ChatGPT

Information Integrity in Public Libraries: A Mini-Book on Misinformation, Disinformation, and AI-Accelerated Deception

Executive Summary

Public libraries in the United States increasingly operate in a high-conflict information environment where false or misleading claims can spread faster than corrections, harm patrons directly (health, finances, safety, civic participation), and erode community trust in institutions—including libraries themselves. [1] One U.S. government framing (from the U.S. Surgeon General) characterizes health misinformation as a serious threat that can cause confusion, sow mistrust, harm health, and undermine public health efforts—and argues that limiting its spread requires whole-of-society action. [2]

A useful starting point is **"information disorder"**: the ecosystem of false, misleading, or weaponized information—including content that is false by mistake, content that is false on purpose, and content that is true but used to harm. [3] In practice, library workers benefit from treating these not as abstract terms but as **operational categories** that shape response: intent, potential harm, and the "path" the information took (source → amplification → audience impact). [3]

Recent U.S. trends show both continuity and acceleration: - A major "baseline" case is foreign information warfare surrounding elections; a U.S. Senate report found that Russia's targeting of the 2016 presidential election was part of a broader, ongoing information warfare campaign and that the Internet Research Agency's activities were more extensive than initially understood. [4]
- The COVID era intensified the "infodemic" dynamic (high volume, rapid change, and confusion), with public health authorities emphasizing the need for verification and careful sharing. [5]
- Since 2023–2025, **generative AI** has expanded the scale, personalization, and plausibility of manipulation—while detection remains imperfect. NIST notes that generative AI systems can support unintentional misinformation (via "confabulations," often called hallucinations) and intentional disinformation at scale. [6] NIST's evaluation work also stresses that detection approaches can face scope limits, scalability constraints, and a moving "cat-and-mouse" dynamic as generators improve. [7]

For public libraries, the practical goal is not to "police truth," but to **strengthen information agency** and reduce harm while upholding professional ethics: intellectual freedom, viewpoint neutrality in governance, equitable service, and patron privacy/confidentiality. [8] An effective library program typically combines: - **Patron-facing instruction** (media/information literacy, scam resistance, voting info referrals to official sources). [9]
- **Staff verification workflows** (fast triage + deeper verification, documentation, escalation). [10]
- **Incident response protocols** for misinformation episodes affecting the library or community (rumors, deepfake attacks, high-risk civic/health misinformation). [11]
- **A tool stack** that prioritizes provenance and cross-source corroboration over "magic" AI detectors. [12]

Unspecified constraints (not provided by the requester) include: local budget levels; staff time and union/work rules; governance/board policies; current tech stack (web CMS, social media tools, endpoint security); and the degree of coordination with local government and community partners. This mini-book assumes these constraints vary by library and provides tiered (free/low-cost first) options. [13]

Definitions and Distinctions

A widely used framework (frequently referenced in policy, journalism, and education) distinguishes three core categories within "information disorder":

Misinformation
False or misleading information shared without intent to cause harm (e.g., a patron shares an outdated "school closure" screenshot believing it is current). [14]

Disinformation
False information deliberately created or shared to cause harm or to achieve a strategic goal (political manipulation, financial fraud, harassment). [15]

Malinformation
Information that is substantially true, but used to harm—often through context collapse, doxxing, selective leaks, or misleading framing (e.g., sharing a real person's address; publishing an authentic clip but stripping context to create a false impression). [3]

For library practice, this framework matters because **the harm pathway differs**: - Misinformation often responds well to **gentle correction + skill-building** ("If you're not sure, don't share," verify with credible sources). [16]
- Disinformation often requires **protective steps** (documentation, escalation, security measures, public clarification, reporting to platforms or authorities). [11]
- Malinformation often requires **privacy- and safety-centered responses** (limit amplification, support affected patrons/staff, consult counsel/policy). [17]

Information literacy and media literacy (library-relevant framing)
Public library media literacy guidance emphasizes libraries as lifelong learning hubs that help community members build critical thinking for navigating digital media. [18] A widely used media literacy definition (cited in library guidance and aligned with common library teaching practice) points to the ability to access, analyze, evaluate, create, and act using forms of communication. [19]

A practical operational definition for public library work is:

> **Information literacy in a public library**: the set of skills and habits that help patrons (and staff) decide what to trust, what to do next, and how to avoid harm—while respecting diverse viewpoints and maintaining privacy. [20]

Historical Context and Recent US Trends

The U.S. information environment has long included propaganda, rumor, and manipulated media, but recent decades added three accelerants: platform-scale distribution, algorithmic amplification, and now generative AI. [21]

A short timeline helps staff situate why "misinformation programming" now covers elections, health, scams, and synthetic media in one continuum:

```
timeline
  title U.S. information disorder: selected milestones
  2016 : Election-related influence operations scale up on social platforms
(foreign + domestic)
  2017 : "Information disorder" framework popularizes mis/dis/malinformation
distinctions
  2020 : COVID era intensifies "infodemic" dynamics (high volume, fast-evolving
guidance)
  2021 : U.S. Surgeon General advisory calls for whole-of-society action on
health misinformation
  2023 : Generative AI mainstreaming increases realistic synthetic images/audio;
viral hoaxes demonstrate market/safety impacts
  2024 : AI voice-clone robocalls used in election-related voter suppression
attempt; FCC enforcement actions follow
  2024-2025 : Standards and evaluations expand: NIST synthetic content guidance;
NIST detector testing; provenance standards (C2PA) grow
```

Election-related influence operations and social media
A U.S. Senate report on Russian social media operations concluded that Russia targeted the 2016 election as part of a broader, ongoing information warfare campaign aimed at sowing discord and undermining trust, and found that the Internet Research Agency's interference was supported by the Russian government. [4] These findings are often cited in library programming because they illustrate how disinformation campaigns exploit social identity and divisive issues rather than "debating facts" in good faith. [22]

Health misinformation and the "infodemic" pattern
The U.S. Surgeon General's advisory emphasizes that misinformation can flourish through social media feeds, blogs, forums, and group chats, and explicitly recommends verifying accuracy with credible sources—adding a simple behavioral rule: "If you're not sure, don't share." [23] For public libraries, this aligns naturally with reference practice: ask clarifying questions, locate authoritative sources, and communicate uncertainty clearly. [9]

Generative AI "proof-of-concept" incidents
In May 2023, a false claim of an explosion near the Pentagon spread online alongside an image that analysts and officials described as likely AI-generated; U.S. officials refuted the incident. [24] NIST later referenced this kind of event as an example of synthetic media eroding trust and creating downstream effects, including brief market impacts. [6]

AI voice cloning enters civic manipulation and fraud at low cost
In January 2024, New Hampshire residents received robocalls using an AI-generated voice resembling President Biden; authorities traced calls and investigated. [25] The FCC proposed a substantial fine for these robocalls, emphasizing the combination of AI voice cloning and caller-ID spoofing in attempting to suppress voting. [26] This is an important library-relevant case because it merges: (1) synthetic media, (2) civic participation harms, and (3) fast community-level confusion that can drive immediate patron questions. [27]

Public concern about misinformation in the news ecosystem
A Pew Research Center analysis found that among Americans who get news on social media, a large share cite "inaccuracy" (including misinformation and unreliable sources) as the thing they dislike most—an increase compared with earlier years. [28] For libraries, this creates both a need and an opportunity: patrons may distrust "the internet," but still need help navigating it for forms, benefits, jobs, school, and health information. [29]

Legal and Ethical Considerations for Public Libraries

This section is practical and non-exhaustive; local implementation should align with state law, municipal counsel guidance, and board policy.

Core library ethics: access, viewpoint neutrality, and the "right to know"
The guiding professional commitments in U.S. librarianship include providing access to information, resisting censorship, and supporting intellectual freedom. [30] These commitments do not require libraries to *endorse* false claims; rather, they support equitable access and the patron's autonomy to inquire—paired with professional responsibility to provide context, reference support, and reliable resources. [31]

Privacy and confidentiality in a misinformation response context
Library privacy is not a "nice-to-have" in misinformation work; it is a direct safeguard against chilling effects, targeted harassment, and coercive surveillance—especially when misinformation topics intersect with politics, identity, and health. [17] The ALA notes that **48 states and D.C.** have laws protecting the confidentiality of library records (with additional protections via attorney general opinions in states without statutes). [32]
Operationally, this means verification and incident response should be designed to minimize collection of personally identifying information, and when data is necessary (e.g., documenting threatening incidents), it should be handled under clear retention and access rules. [33]

Libraries as limited public forums and the role of conduct rules
A common legal framing treats public libraries as limited public forums created for the purpose of providing access to information and ideas, with authority to enforce reasonable rules ensuring the space functions for its intended use. [34] This is relevant because misinformation incidents sometimes present as disruptive behavior (e.g., a patron attempting to distribute harassing flyers or confront staff). Clear, viewpoint-neutral conduct policies support staff safety while reducing claims of viewpoint discrimination. [35]

Internet filtering and institutional obligations
Federal law and litigation around the Children's Internet Protection Act (CIPA) illustrate the tension between access and mandated restrictions in public internet terminals. [36] While CIPA is not a "misinformation law," it shapes how public libraries structure access and how staff respond to contested categories of content. [37]

AI-generated deception: emerging legal landscape important to libraries
Libraries are not typically the primary regulators of AI harm, but they operate at the interface of public information, community trust, and education. Three legal domains are especially relevant:

- **Election interference and consumer deception in communications**: In the New Hampshire robocall case, the FCC's enforcement language treated the AI voice cloning plus spoofing as part of fraudulent/deceptive robocalling with civic harm implications. [26]
- **Fraud and identity verification risks**: FinCEN warned that criminals have used generative AI to create falsified documents, photographs, and videos to circumvent identity verification and enable fraud schemes. [38]
- **Digital replicas (voice/likeness) and intellectual property**: The U.S. Copyright Office's report on "digital replicas" describes how realistic replication of voice or appearance can affect voter behavior and reputations, placing pressure on existing legal frameworks and policy debates. [39]

For libraries, the key operational takeaway is to **avoid overpromising legal certainty** in patron interactions. Instead: supply credible sources, explain what is known/unknown, and direct patrons to appropriate agencies (e.g., state AG consumer protection, FCC complaint resources, election offices) when the issue is fraud or election interference. [40]

Cognitive and Social Drivers

Library staff are more effective when they treat misinformation not as "tough patrons" but as predictable human behavior under cognitive load and social pressure. Several evidence-backed drivers are especially relevant for public libraries:

Familiarity and repetition (illusory truth)
Repeated claims tend to feel more believable, even when false; this is a longstanding finding in cognitive psychology, and recent work continues to examine mechanisms and boundary conditions. [41] In libraries, this shows up when patrons say "I've seen this everywhere," conflating prevalence with accuracy. [42]

Attention scarcity and "sharing without thinking"
Research indicates that small "accuracy prompts" can reduce sharing of misinformation by shifting attention from social engagement to truthfulness judgments. [43] A library-friendly translation is to make "pause signals" habitual: a 10-second stop before sharing, and a quick check for better coverage. [44]

Identity, emotion, and group belonging
Misinformation spreads not only because people misunderstand facts, but because content signals identity and moral affiliation—especially when it triggers anger, fear, or outrage. [45] Libraries can reduce defensiveness by emphasizing shared values (safety, community wellbeing, trust, autonomy) and by using nonjudgmental language in reference interactions. [16]

Skill gaps in evaluating online sources
Research on "lateral reading" suggests that expert fact-checkers leave a site quickly and open new tabs to evaluate credibility via external sources, while novices often stay on the site and focus on surface features. [46] Studies and educational interventions show that these strategies can be taught and can improve performance. [47] This is a direct fit for public library instruction because it relies on common browser skills rather than specialized tools. [48]

Prebunking and inoculation approaches
"Prebunking" (teaching manipulation techniques before exposure) is increasingly supported as a scalable complement to debunking, with research and applied trials showing improvements in identifying misleading techniques. [49] For libraries, prebunking aligns with recurring programming (news literacy, scam literacy) and can be integrated into short workshops rather than requiring a full course. [50]

AI and Synthetic Media in the Information Ecosystem

What changed with generative AI

Generative AI systems increase both volume and plausibility of false content
NIST's Generative AI Profile explicitly notes that generative AI can facilitate unintentional misinformation (especially when outputs stem from "confabulations") and intentional disinformation at scale, including highly realistic synthetic audiovisual content ("deepfakes"). [6] This undermines a long-standing shortcut: "it looks real, so it must be real." [51]

Detection is not a stable "one-and-done" solution
NIST's GenAI evaluation work highlights that detection research faces limitations in scope (tasks tested may not represent all real-world content), assumptions of binary human-vs-AI authorship that may fail as hybrid writing grows, scalability constraints, and bias risks in discriminator models. [52] Separately, U.S. homeland security–oriented analysis emphasizes the "cat and mouse" dynamic between generating and detecting sophisticated forgeries. [53]

Deepfakes are not only political
Government warning documents position deepfakes as multipurpose: brand harm, impersonation of leaders/financial officers, and enabling access to systems via social engineering. [54] This is highly relevant to public libraries because library staff and patrons are "high-contact" targets for scams: phone calls, emails, and in-person requests for urgent help (printing wire instructions, changing account details, etc.). [55]

Provenance, watermarking, and Content Credentials

A major shift in countermeasures is a move from "spot the fake" to "verify the origin." NIST describes a family of technical approaches under **digital content transparency**, including watermarking and provenance tracking. [56]

Watermarking: useful but not a silver bullet
NIST explains that watermarking schemes vary (overt vs covert; public vs private; reversible vs irreversible), and that even covert watermark detectors have nonzero false positive/false negative rates; effectiveness depends on accurate detection and robustness. [57]

C2PA / Content Credentials: a practical "nutrition label" model
The C2PA [58] specification provides an open technical standard for attaching verifiable provenance "manifests" to media, intended to complement media literacy and fact-checking rather than replace them. [59] A U.S. defense-oriented brief on Content Credentials emphasizes both promise and limitation: trust is multifaceted, credentials can be stripped when content is reposted, and "durable" approaches (credential-linked watermarking/fingerprinting) are emerging but still evolving. [60]

Practical verification entry points include: - “Verify” tooling promoted by the Content Credentials ecosystem (drag-and-drop inspection). [61]
- The open-source `c2patool` command-line utility for inspecting and reporting C2PA manifests. [62]
- Browser extensions that surface credentials where platforms don’t display them. [63]

Library Verification and Response Playbook

This section is designed to be “desk-ready”: staff can adapt it into a local SOP, training, and patron education.

A verification mindset that fits public libraries

Public library workers are not expected to become forensic analysts. Your comparative advantage is: **reference interviewing + trusted sourcing + documentation**. [9] Two evidence-aligned approaches map well to public library work:

- **Lateral reading**: leave the page; open new tabs; check what reliable sources say *about* the source and the claim. [46]
- **SIFT** (Stop; Investigate the source; Find better coverage; Trace to the original context): a compact sequence of “moves” designed to reduce reactive sharing and improve evaluation. [64]

Workflow flowchart for staff triage and verification

```
flowchart TD
A[Patron brings claim / staff sees viral post] --> B{Time sensitivity + harm?}
B -->|High: health/safety/election/fraud| C[Rapid triage]
B -->|Moderate/low| D[Standard verification]

C --> E[Clarify the claim + intended action]
E --> F[SIFT: Stop + Investigate source]
F --> G[Find better coverage: 2-3 independent sources]
G --> H[Trace to original: document, dataset, official notice, full quote,
original media]
H --> I{Still uncertain?}
I -->|Yes| J[Communicate uncertainty + safest next step; refer to authoritative
agencies]
I -->|No| K[Provide verified summary + sources]

D --> F

K --> L[Document: what checked, what found, links/screenshots, time/date]
J --> L
L --> M{Incident affecting library operations/brand?}
M -->|Yes| N[Activate incident response protocol]
M -->|No| O[Close: patron guidance + optional follow-up resources]
```

This flow reflects public-health guidance emphasizing “verify accuracy with trustworthy sources; if you’re not sure, don’t share,” plus the practical library teaching focus on building evaluation skills rather than winning arguments. [16]

Detection methods compared

Method	Best for	Strengths	Failure modes / limits	Skill level
Cross-source corroboration (reputable outlets + primary documents)	Most claims	Strongest general reliability; reduces single-source traps	Coordinated disinformation can seed multiple outlets; requires source judgment	Beginner–Intermediate [65]
Reverse image search (Google Images, Bing, TinEye)	Miscontextualized images	Fast “where else did this appear?” check	New images may have no matches; screenshots/crops reduce matching	Beginner [66]
Video keyframe analysis (InVID-WeVerify)	Viral videos	Extract keyframes, metadata hints, platform search aids	Can’t “prove authenticity”; platform stripping and reposting complicates	Intermediate [67]
Metadata inspection (EXIF/XMP via tools like ExifTool; C2PA manifests)	Photos/videos with preserved provenance	Can reveal device/software, dates, provenance chain	Metadata often stripped; metadata can be forged; absence ≠ fake	Intermediate [68]
Content provenance standards (C2PA / Content Credentials)	Authenticity verification where adopted	Cryptographic signatures + edit history; aligns with “trust but verify”	Adoption incomplete; display inconsistent; screenshots can sever chain	Intermediate [69]
AI “detectors”	Supplemental signal only	Can flag obvious synthetic content quickly	High false positives/negatives; scope limits; rapidly outdated	Intermediate–Advanced [70]

Recommended tool stack for libraries

The table below prioritizes (1) reliability, (2) free/low-cost access, and (3) primary-source transparency. (Tools evolve; libraries should review access, privacy terms, and accessibility before institutional adoption.)

Priority tier	Tool / resource	Category	Typical cost	Why it's recommended	Primary source
Highest	Content Credentials "Verify" site	Provenance inspection	Free	Quick inspection of Content Credentials where present; supports provenance-first verification	61
Highest	`c2patool`	Provenance inspection	Free / open source	Detailed extraction of C2PA manifests for advanced staff/partners	62
Highest	InVID / WeVerify verification plugin	Video/image OSINT	Free	Designed for verification workflows (keyframes, search helpers)	67
Highest	ExifTool	Metadata	Free	Widely used metadata reader/editor for many formats	71
High	Google Fact Check Explorer	Fact-check search	Free	Searches a large database of published fact checks using structured ClaimReview data	72
High	TinEye	Reverse image search	Free (basic)	Useful for locating reuses/ variants of images; privacy claims for search uploads	73
High	Internet Archive (Wayback Machine "Save Page Now")	Web archiving	Free	Preserves pages for later reference; helps document rapidly changing misinformation pages	74
High	FactCheck.org	Fact-checking	Free	Nonpartisan fact-checking resource with public-facing explanations	75
High	Be MediaWise library toolkit (Poynter/ALA partnership)	Training/ programming	Free	Library-oriented adult media literacy materials designed for librarian delivery	76
Caution	AI-content detectors (various)	Detection	Often paid / mixed	Use only as a secondary weak signal; NIST highlights limitations and evolving generator-detector dynamics	7

Incident response protocol for public libraries

Misinformation incidents vary (a viral rumor about the library; a deepfake of a director; a scam targeting seniors; a local election rumor printed and distributed). A library incident protocol should emphasize: safety, accuracy, documentation, and governance alignment.

```
flowchart TD
A[Incident detected] --> B[Assess harm: safety, fraud, civic interference,
reputational risk]
B --> C{Immediate danger?}
C -->|Yes| D[Contact emergency services / security per local policy]
C -->|No| E[Stabilize: do not amplify; preserve evidence]

E --> F[Preserve evidence: screenshots, URLs, timestamps; archive page if
possible]
F --> G[Verify claim using workflow; identify authoritative sources]
G --> H[Decide response channel: in-person script, website notice, social post,
partner referral]

H --> I{Library targeted?}
I -->|Yes| J[Coordinate with director/PIO/city legal counsel; staff briefing]
I -->|No| K[Provide patron guidance + referrals]

J --> L[Public correction: factual, calm, cite sources; avoid repeating false
claim headline]
K --> M[Log incident + after-action review; update training/materials]
L --> M
```

This model is consistent with U.S. agency guidance emphasizing documentation, mitigation, and the reality that synthetic media can be used for impersonation and fraud. [77]

Case study template (plausible but realistic): "Library closure rumor + targeted harassment"
A fabricated screenshot circulates claiming "the library is closed indefinitely due to contamination." Patrons arrive angry; staff receive threatening calls. Response: preserve screenshot and original post URLs; publish a short, factual notice ("Library is open; hours are…; official updates only on our website/phone"); coordinate with municipality; avoid repeating sensational claim language; remind staff not to share patron names; document threats; if credible threats occur, escalate to law enforcement per policy. This approach aligns with conduct-rule authority in limited public forum doctrine (manage disruptions) while protecting privacy and focusing communication on verified operational facts. [78]

Case study (real): AI voice-clone robocall voter suppression attempt
Libraries should anticipate patron questions ("Is this real?" "Where do I vote?"). Staff response: do not diagnose audio authenticity at the desk; instead provide authoritative voting information sources (state election office), document the report for internal awareness, and share reputable reporting/agency actions. The FCC's action and the New Hampshire AG updates demonstrate that AI voice impersonation has already been used in election interference attempts. [79]

Templates and Training Materials

Staff training curriculum overview

A public library curriculum should be modular (so part-time staff and varied tech skills can participate) and scenario-driven (so learning transfers to desk work).

Module	Audience	Time	Core competencies	Practice activity	Assessment
Foundations: information disorder + ethics	All staff	60–90 min	Mis/dis/ malinformation distinctions; library ethics; privacy	Scenario sort: classify examples by intent/harm	Short quiz + reflection [80]
Lateral reading + SIFT at the desk	Public service staff	90 min	Open-tab verification; “find better coverage”	Timed verification drills (2–5 min)	Observation rubric [81]
Visual verification basics	Interested staff	90 min	Reverse image search; basic metadata concepts	Pentagon-image-style drill (find origin/ context)	Checklist completion [82]
AI & synthetic media	All staff	90 min	Deepfake risks; confabulations; detector limits; provenance	Compare: “detector vs provenance” exercise	Scenario Q&A [83]
Scams + fraud resilience	Public-facing + outreach	60 min	Voice cloning risk; safe referrals; documentation	“Urgent family emergency” scam role-play	Knowledge check [55]
Incident response tabletop	Leadership + key staff	90–120 min	Evidence capture; comms; escalation thresholds	Run a mock “deepfake director” crisis	After-action report [84]

Workshop outlines for patrons

Workshop A: “Stop, Verify, Share: A 60-Minute Skill Session” (Adults, general audience)
Learning goals: adopt “pause before sharing”; learn lateral reading; practice 2 quick tools (Fact Check Explorer + reverse image search). [85]
Structure:
- 10 min: Why misinformation spreads (repetition, emotion; normalize mistakes). [86]
- 15 min: Live SIFT demo on a questionable website (open-tab checks). [87]
- 20 min: Small-group practice with 2 claims (one misinfo, one disinfo).
- 10 min: “If you’re not sure, don’t share” + where to check (library handout). [5]
- 5 min: Q&A + take-home resources.

Workshop B: "Deepfakes and AI: What to Do When You Can't Trust Your Eyes" (Adults/teens, 75–90 min)
Learning goals: understand deepfakes; learn provenance checks; learn safe responses to suspicious audio/video. [88]
Structure:
- 15 min: What generative AI changes (scale + plausibility; confabulations). [89]
- 20 min: Provenance demo (Content Credentials "Verify" + explanation of limits). [90]
- 20 min: Video verification basics (InVID keyframes; reverse searches). [91]
- 15 min: "Liar's dividend" concept: why "it's a deepfake" can be abused to deny real evidence; discuss careful language. [92]
- 10 min: Scam tie-in (voice cloning; verification steps before acting). [55]

Workshop C: "Scams, Voice Cloning, and Safe Online Actions" (Seniors/community groups, 60 min)
Learning goals: recognize urgent-need manipulation; verify identity; avoid "act-now" traps. [55]
Structure:
- 15 min: Common scam tactics and why AI increases realism. [38]
- 20 min: Role-play: "family emergency" + "bank/security call."
- 15 min: A personal "verification plan" (call-back numbers, trusted contacts, pause rules).
- 10 min: Local referrals (consumer protection, bank fraud hotlines, FTC consumer guidance—localize as available). [38]

Role-play exercises for staff

Role-play 1: The confident-but-wrong patron (misinformation)
Goal: maintain rapport; shift to verification; avoid humiliation. [16]

Script prompts: - Patron: "I saw this everywhere—schools are closing tomorrow."
- Staff moves: clarify ("Which school district?" "Where did you see it?"), verify on official sites, explain date/context mismatch, offer a handout on "how to check official notices."

Role-play 2: The agitated patron (malinformation or disinformation amplification)
Goal: de-escalate; enforce conduct rules; avoid amplifying harmful content. [93]
- Patron wants to post someone's personal address on a public library bulletin board or demands staff "share this list."
- Staff moves: reference privacy/safety policy; offer alternative ("We can help you find official complaint channels"), document incident per policy.

Role-play 3: The "deepfake director" phone call (fraud)
Goal: verify identity before action; follow security plan. [55]
- Caller: "This is the director—buy gift cards / wire funds / reset passwords."
- Staff moves: require callback on known number; involve supervisor; refuse urgent financial actions without policy-based verification.

Ready-to-use templates

The following templates are designed for copy/paste and local customization (replace bracketed fields).

Patron handout (one-page)

```
LIBRARY QUICK GUIDE: BEFORE YOU SHARE

1) STOP
   - What is the claim? What action is it asking you to take?
   - If it makes you angry or scared, pause. Emotion is a common manipulation
tool.

2) CHECK THE SOURCE (WHO IS BEHIND IT?)
   - Look up the author/organization in a new tab.
   - What is their track record? Are they transparent?

3) FIND BETTER COVERAGE
   - Search: Are multiple independent, credible sources reporting the same
thing?
   - For health and safety: prefer official or medical/public health sources.

4) TRACE TO THE ORIGINAL
   - Can you find the original document, full quote, original photo/video, or
official notice?
   - Watch for old screenshots and missing dates.

5) IF YOU'RE NOT SURE, DON'T SHARE
   - Ask us! Library staff can help you verify sources and find reliable
information.

Helpful tools:
- Google Fact Check Explorer: search if a claim has already been fact-checked
- Reverse image search: check where an image first appeared
- Wayback Machine: save a web page for documentation
```

This message aligns with public guidance emphasizing "verify with trustworthy sources" and "if you're not sure, don't share," adapted into a library-friendly workflow. [94]

Signage (for computer areas / adult services desk)

```
VERIFY BEFORE YOU SHARE
Not everything online is true—even if it looks real.

Try the 3-step check:
1) Pause (10 seconds)
2) Open a new tab (who is behind it?)
3) Find better coverage (what do other reliable sources say?)
```

```
Need help? Ask a librarian.
```

Social media posts (3 options)

```
Post 1 (general):
Before you share: STOP. Check the source. Find better coverage. Trace to the
original.
Your library can help with quick verification tools and trusted sources.

Post 2 (AI/deepfake awareness):
AI can generate realistic images, audio, and video. If something seems shocking,
pause and verify:
- Look for original sources
- Check reputable reporting
- Ask the library for help

Post 3 (scam prevention):
Urgent calls/texts that demand money or passwords are a red flag—especially with
AI voice cloning.
Use a callback number you already trust. When in doubt, pause and ask for help.
```

Staff policy template (mini SOP: “Misinformation assistance”)

```
PURPOSE
Support patrons in evaluating information while protecting privacy and upholding
intellectual freedom.

SCOPE
Applies to reference interactions, outreach programming, and library
communications.

PRINCIPLES
- Provide access; do not endorse falsehoods.
- Prioritize patron privacy and confidentiality.
- Use viewpoint-neutral conduct rules for disruptions.
- Focus on skills and sources, not “winning arguments.”

WORKFLOW (Desk)
1) Clarify the claim and intended action.
2) Apply SIFT / lateral reading.
3) Use at least two independent credible sources or primary documents when
possible.
4) If uncertain, say so; provide safest next step and authoritative referrals.
5) Document incidents that involve threats, fraud, or library targeting per
```

```
incident protocol.

ESCALATION
- Threats or harassment → supervisor + security policy.
- Fraud indicators (voice cloning, identity theft) → supervisor + referral to
appropriate agencies.
- Library-wide rumor/brand attack → director/PIO + coordinated public statement.

PRIVACY
Do not collect personal details unless necessary for safety or required incident
documentation.
Store records per retention policy.
```

Evaluation rubric (for source quality in workshops)

Criteria	0	1	2
Source transparency	No author/org info	Partial / unclear	Clear author/org + contact/ mission
Evidence quality	No evidence	Links but weak/ unclear	Primary sources, data, methods, citations
Corroboration	Contradicted by credible sources	Mixed/uncertain	Supported by multiple credible sources
Date/context	Missing or misleading	Present but ambiguous	Clear date, context, original form
Manipulation signals	High (rage bait, urgency, conspiratorial)	Some	Low; neutral tone; clear limits

Primary sources and “go-to” links (copy/paste)

```
U.S. Surgeon General advisory on health misinformation (PDF):
https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-
advisory.pdf

NIST AI RMF: Generative AI Profile (PDF):
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

NIST GenAI detector evaluation (PDF):
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.700-1.pdf

NIST synthetic content transparency overview (PDF):
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf

U.S. Senate report on Russian social media interference (PDF):
```

```
https://www.intelligence.senate.gov/sites/default/files/documents/
report_volume2.pdf

FCC press release on AI deepfake robocall enforcement (PDF):
https://docs.fcc.gov/public/attachments/DOC-402762A1.pdf

New Hampshire AG update on AI robocall investigation:
https://www.doj.nh.gov/news-and-media/voter-suppression-ai-robocall-
investigation-update

FinCEN deepfake fraud alert (PDF):
https://www.fincen.gov/system/files/shared/FinCEN-Alert-DeepFakes-
Alert508FINAL.pdf

C2PA specification (web):
https://c2pa.org/specifications/specifications/

Content Credentials “Verify” tool:
https://contentcredentials.org/verify

C2PA open-source command-line tool docs:
https://opensource.contentauthenticity.org/docs/c2patool/c2patool-index/
```

These primary sources anchor the mini-book’s most important claims about harms, detection limitations, and practical mitigations. [95]

---

[1] [28] [29] https://www.pewresearch.org/short-reads/2024/02/07/many-americans-find-value-in-getting-news-on-social-media-but-concerns-about-inaccuracy-have-risen/
https://www.pewresearch.org/short-reads/2024/02/07/many-americans-find-value-in-getting-news-on-social-media-but-concerns-about-inaccuracy-have-risen/

[2] [5] [16] [23] [85] [94] [95] https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf
https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf

[3] [14] [15] [80] https://shorensteincenter.org/resource/information-disorder-framework-for-research-and-policymaking/
https://shorensteincenter.org/resource/information-disorder-framework-for-research-and-policymaking/

[4] [21] [22] https://www.intelligence.senate.gov/wp-content/uploads/2024/08/sites-default-files-documents-report-volume2.pdf
https://www.intelligence.senate.gov/wp-content/uploads/2024/08/sites-default-files-documents-report-volume2.pdf

[6] [83] [89] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[7] [12] [52] [70] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.700-1.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.700-1.pdf

8 30 31 https://www.pclibs.org/documents/ALA%20Freedoms%20-%20Read%2C%20View%2C%20Bill%20of%20Rights.pdf
https://www.pclibs.org/documents/ALA%20Freedoms%20-%20Read%2C%20View%2C%20Bill%20of%20Rights.pdf

9 10 13 18 19 20 https://www.ala.org/sites/default/files/tools/content/%21%20FINAL%20Media-Lit_Prac-Guide_WEB_040521.pdf
https://www.ala.org/sites/default/files/tools/content/%21%20FINAL%20Media-Lit_Prac-Guide_WEB_040521.pdf

11 26 27 40 79 https://docs.fcc.gov/public/attachments/DOC-402762A1.pdf
https://docs.fcc.gov/public/attachments/DOC-402762A1.pdf

17 33 https://www.ala.org/advocacy/intfreedom/privacyconfidentialityqa
https://www.ala.org/advocacy/intfreedom/privacyconfidentialityqa

24 82 https://www.reuters.com/article/fact-check/online-posts-reporting-explosion-near-pentagon-on-may-22-2023-are-false-idUSL1N37J2QJ/
https://www.reuters.com/article/fact-check/online-posts-reporting-explosion-near-pentagon-on-may-22-2023-are-false-idUSL1N37J2QJ/

25 https://www.doj.nh.gov/news-and-media/voter-suppression-ai-robocall-investigation-update
https://www.doj.nh.gov/news-and-media/voter-suppression-ai-robocall-investigation-update

32 https://www.ala.org/advocacy/privacy/statelaws
https://www.ala.org/advocacy/privacy/statelaws

34 78 93 https://www.ala.org/advocacy/intfreedom/censorship/courtcases
https://www.ala.org/advocacy/intfreedom/censorship/courtcases

35 https://law.justia.com/cases/federal/appellate-courts/F2/958/1242/371694/
https://law.justia.com/cases/federal/appellate-courts/F2/958/1242/371694/

36 https://www.oyez.org/cases/2002/02-361
https://www.oyez.org/cases/2002/02-361

37 https://supreme.justia.com/cases/federal/us/539/194/
https://supreme.justia.com/cases/federal/us/539/194/

38 55 https://www.fincen.gov/system/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf
https://www.fincen.gov/system/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf

39 https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf
https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf

41 42 86 https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1215432/full
https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1215432/full

43 44 https://www.nature.com/articles/s41467-022-30073-5
https://www.nature.com/articles/s41467-022-30073-5

45 58 https://www.rand.org/pubs/commentary/2023/08/truth-decay-and-national-security.html
https://www.rand.org/pubs/commentary/2023/08/truth-decay-and-national-security.html

46 48 65 81 https://journals.sagepub.com/doi/10.1177/016146811912101102
https://journals.sagepub.com/doi/10.1177/016146811912101102

47 https://pmc.ncbi.nlm.nih.gov/articles/PMC8012470/
https://pmc.ncbi.nlm.nih.gov/articles/PMC8012470/

49 50 https://www.jmir.org/2023/1/e49255/
https://www.jmir.org/2023/1/e49255/

51 54 77 84 88 https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF
https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

53 https://www.govinfo.gov/content/pkg/CMR-HS1-00193155/pdf/CMR-HS1-00193155.pdf
https://www.govinfo.gov/content/pkg/CMR-HS1-00193155/pdf/CMR-HS1-00193155.pdf

56 57 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf

59 https://c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html
https://c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

60 https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF
https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF

61 90 https://verify.contentauthenticity.org/
https://verify.contentauthenticity.org/

62 https://opensource.contentauthenticity.org/docs/c2patool/c2patool-index/
https://opensource.contentauthenticity.org/docs/c2patool/c2patool-index/

63 https://chromewebstore.google.com/detail/adobe-content-authenticit/dmfbmenkapmaoldfgacgkoaoiblkimel?hl=en
https://chromewebstore.google.com/detail/adobe-content-authenticit/dmfbmenkapmaoldfgacgkoaoiblkimel?hl=en

64 https://hapgood.us/2019/06/19/sift-the-four-moves/
https://hapgood.us/2019/06/19/sift-the-four-moves/

66 73 https://tineye.com/
https://tineye.com/

67 91 https://www.invid-project.eu/tools-and-services/invid-verification-plugin/
https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

68 71 https://exiftool.org/
https://exiftool.org/

69 https://c2pa.org/specifications/specifications/2.3/explainer/Explainer.html
https://c2pa.org/specifications/specifications/2.3/explainer/Explainer.html

72 https://toolbox.google.com/factcheck/explorer
https://toolbox.google.com/factcheck/explorer

74 https://wayback.archive.org/
https://wayback.archive.org/

75 https://www.factcheck.org/
https://www.factcheck.org/

76 https://www.poynter.org/mediawise/misinformation-resilience-toolkit-libraries/
https://www.poynter.org/mediawise/misinformation-resilience-toolkit-libraries/

87 https://guides.lib.wayne.edu/sift
https://guides.lib.wayne.edu/sift

92 https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend

https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend