

AI PLAYBOOK 33

Brand Safety: How to Spot Fake AI Marketing

November 2025



[theCMA.ca](https://www.thecma.ca)

Content partially generated by artificial intelligence, refined by human expertise.

This Playbook is part of the CMA's AI Mastery Series, empowering marketers to Implement AI in ways that earn regulatory confidence, maintain strong brand reputation, and foster consumer trust

The detection protocol

Marketing teams face an unprecedented challenge due to AI: distinguishing authentic brand communications from sophisticated AI-generated imposters that can undermine years of brand building efforts. Fake AI content is actively targeting brands. These include fake customer testimonials flooding review platforms, synthetic influencer campaigns promoting competing products, AI-generated negative reviews designed to damage reputation, manipulated executive quotes in fake news articles, and deepfake spokesperson videos claiming false endorsements.

Your team plays a critical role in identifying these threats early. Use the following detection protocol to assess suspicious content:

- **Recognize:** Identify threats using observable patterns your eyes and ears can detect
- **Verify:** Confirm manipulation using accessible detection tools integrated into daily workflows
- **Respond:** Execute clear protocols including documentation, escalation, and crisis communication

This playbook provides marketing professionals with approaches to identify, prevent, and respond to AI-generated content threats while contributing to broader consumer education efforts. You'll learn to identify fake testimonials, synthetic social media campaigns, manipulated video content, and competitive AI attacks before they impact your brand.

How to spot fake text

Staff regularly encounter suspicious text content during routine activities like screening customer testimonials for campaigns, monitoring social media comments and reviews and evaluating user-generated content submissions. The key is developing an eye for content that feels "off". That is, text that lacks the natural imperfections, specific details, and authentic voice that characterize genuine human communication. These observable patterns become second nature with practice.

1. Recognize the text red flags:



- Generic praise lacking specific product details or personal experiences
- Testimonials that sound identical across different supposedly independent sources
- Reviews mentioning features your product doesn't have or using outdated terminology
- Overly emotional language that doesn't match your typical customer voice
- Excessive em-dashes, use of emoji lists and a summary at the bottom

2. Verify with detection tools:

- Use Originality.ai's browser extension during social monitoring to check suspicious testimonials
- Deploy GPTZero integration in your content management system to screen user content before publication
- Set up Winston AI brand monitoring alerts to catch AI-generated attacks early

3. Respond with Clear Action:

- Document suspicious content with screenshots and URLs before it disappears
- Cross-reference claims against actual product specifications and customers
- Flag content immediately in #brand-threats channel for crisis team coordination



When in
doubt, don't
hesitate to
escalate.

4

How to spot visual fakes

Marketing teams encounter visual content daily that requires authenticity assessment. From user-generated campaign submissions and influencer partnership materials to executive headshots and customer testimonial photos. Visual manipulation is often harder to detect than text, making systematic inspection crucial for brand protection.

1. Recognize the visual red flags:

- Inconsistent lighting, shadows, or reflections across faces and objects
- Unnatural skin texture in customer photos or distorted limbs (wrong finger counts, or person behind a table where their legs don't line up)
- Blurry logos or nonsensical text in backgrounds upon close inspection
- Executives appearing in contexts or locations that don't match reality

2. Verify the visual content:

- Conduct reverse image searches using Google Images and TinEye to find original sources
- Use Hive AI's detection API to automatically scan social mentions for deepfake content
- Check image metadata and compare against known authentic brand content

3. Respond to visual threats:

- Screenshot and document the fake content immediately with URLs
- Contact platform representatives using established expedited takedown procedures
- Alert the crisis communication team within 2 hours for coordinated brand response

How to spot audio fakes

Staff regularly evaluate audio content including customer testimonials in video campaigns, executive interviews and statements, spokesperson recordings, and voice messages from partners or agencies. Audio manipulation can be particularly deceptive because voice conveys authority and trust, making it a preferred method for sophisticated fraud attacks. Corporate fraud attempts using AI voice clones have resulted in millions in losses when employees receive convincing calls from 'executives' authorizing transfers or policy changes. Careful verification is essential for protecting both brand credibility and financial security.

1. Recognize audio manipulation:

- Robotic delivery in supposedly spontaneous customer testimonials
- Identical pronunciation patterns across different "customer" voices
- Background audio that cuts off unnaturally during speech segments
- Executives speaking in uncharacteristic tones or using terminology they typically avoid

2. Verify audio authenticity:

- Require dual-channel confirmation for any executive audio claiming new partnerships
- Use challenge-response questions that only the actual person would know
- Cross-reference audio content with scheduled meetings and known executive availability

3. Respond to audio threats:

- Implement 24-hour waiting periods for major announcements to allow verification
- Contact the person directly through established secure channels
- Document and escalate through crisis communication protocols immediately

Building your proactive brand protection strategy

As AI technology advances, detection becomes an ongoing race. Plan for regular tool evaluations, budget for upgraded detection technologies, and maintain relationships with cybersecurity experts who specialize in AI threats. Keep a log to track and document all suspected fake content incidents.

Continuous monitoring: Update your detection strategies at least quarterly as new AI models emerge with better image synthesis, more natural voice cloning, and increasingly convincing text generation. Subscribe to AI security newsletters and threat intelligence feeds to stay informed.

Fraud prevention focus: AI-enabled fraud is escalating. Executive impersonation fraud, where AI voice or video clones authorize false transactions or statements, represents one of the fastest-growing corporate security threats. Remember that fake content usually originates from unofficial accounts and suspicious email addresses. Verify all communications through official channels.

Keep skills current: Participate in monthly team sessions reviewing new AI marketing attack examples, quarterly updates on detection tool capabilities, and annual comprehensive crisis response training. Develop internal expertise through certified team members who train others and serve as escalation points.

Build authentic content defenses: Maintain consistent visual branding that AI struggles to replicate perfectly. Use unique spokesperson mannerisms and brand voice elements difficult for AI to capture. Add watermarks and digital locks to official content where technically possible. Create behind-the-scenes content demonstrating genuine processes and relationships that synthetic content cannot replicate. Only post through secure, reliable platforms and verified official accounts.

Detection tools for daily workflows

Organize these tools for quick access during daily workflows and crisis situations.

- Your core daily toolkit includes AI text detectors like Originality.ai or GPTZero for quick content authenticity checks, with API integrations for screening user-generated content in your management systems. Set up monitoring alerts for brand mentions and configure social listening dashboards with AI content anomaly detection.
- Deploy visual detection platforms such as Hive AI, Sensity AI, or Reality Defender for automated image screening. Combine these with basic verification tools like Google Images and TinEye for reverse searches to verify visual content authenticity across campaigns and social mentions.
- Build internal reference libraries including challenge-response question banks for executive verification, authentic brand asset collections for visual comparison, and employee directories with photos. Maintain access to executive calendars for cross-referencing communications and product specification databases for fact-checking technical claims.
- Establish ongoing learning systems with detection skill assessments, AI threat intelligence feed subscriptions, and regularly updated best practice guides. Include Canadian regulatory compliance checklists for privacy and marketing law requirements.

Start with browser-based detection extensions and monitoring alerts. These provide immediate protection while you evaluate and implement more comprehensive platform solutions.

Emergency response and escalation tools

Immediate response approaches:

- #brand-threats channel for immediate team escalation and coordination
- Legal escalation contact list with 24/7 availability for urgent takedown requests
- Reporting templates pre-filled for channels (Meta, X, LinkedIn, TikTok, YouTube)
- Crisis communication scripts pre-approved for different AI threat scenarios
- Inventory of known fakes cataloguing previously identified examples

Documentation and evidence:

- Documentation templates for consistent threat reporting with screenshots, URLs, impact assessment
- Evidence preservation protocols for potential legal proceedings
- Incident response flowchart with clear decision points and timing requirements
- Stakeholder notification trees for internal and external communications

External support network:

- Vendor contact lists for detection tool support and advanced analysis services
- Media contact database for proactive outreach when fake content spreads
- Cybersecurity expert partnerships for sophisticated threat analysis
- Public relations damage control templates for different threat severities

Emergency protocols:

- 24-hour verification waiting periods for major announcements
- Dual-channel confirmation procedures for executive communications
- Mandatory dual-authentication for any communication requesting financial transfers, policy changes, or public statements
- Expedited takedown request processes with platform representatives
- Crisis team activation triggers and coordination procedures

Integrating AI detection into brand protection workflows

Just as you routinely monitor brand mentions, review campaign performance, and check content quality, AI detection must become an integrated part of your daily brand safety practices. This isn't about adding extra work. It's about upgrading your existing workflows with new skills that protect everything you've built.

Remember to:

- Always question content that feels 'off'. Trust your instincts and apply systematic verification
- Follow the **Recognize, Verify, Respond** protocol every time, regardless of time pressure
- Never engage with fake content directly as this amplifies its reach and validates the attack
- Document everything before acting. Evidence disappears quickly in the digital space. Build inventory and document analysis to identify patterns and trends
- Stay adaptable. AI technology evolves rapidly, and so must your detection skills and methods

Your actions for this week:

- **Set up your toolkit:** Bookmark detection tools, install browser extensions, configure monitoring alerts
- **Practice recognition skills:** Review sample AI-generated content in your training materials
- **Test your protocols:** Run through the escalation checklist and verify your crisis response contacts
- **Schedule ongoing development:** Add monthly AI threat reviews to your team calendar

Building sustainable brand protection

The threats are real, but with systematic practice, detection becomes as natural as spell-checking copy or verifying image rights. This isn't a one-time training—it's an evolving practice that grows stronger with consistent application.

Your plan moving forward

- **Weekly:** Monitor brand mentions using your detection tools and social listening dashboards
- **Monthly:** Participate in team training sessions on new AI threats and detection methods
- **Quarterly:** Update your detection tools and review protocol effectiveness
- **Annually:** Comprehensive skills assessment and crisis response plan testing

Your mindset for success

Define your organization's approach to using AI for marketing activities. Identify what is explicitly not within your organization's use of AI such as AI-generated humans/models, synthetic spokesperson content, or customer testimonials. Having clear internal guidelines makes it easier to spot when external threats violate these same principles.

Trust your training, use your tools, and when in doubt, never hesitate to escalate. Embrace healthy skepticism as a core professional skill. When your own team has transparent, ethical AI practices, you're better positioned to recognize when others are crossing those lines.

Recommended reading and references

For further learning, these resources provide practical guidance on responsible AI adoption, strategy, and implementation.

CMA resources

- [CMA Guide on AI for Marketers](#)
- [Setting the Stage on Artificial Intelligence: A CMA Primer on AI for Marketers](#)
- [CMA Accountability Checklists for AI in Marketing](#)
- [CMA Mastery Series: AI Playbooks](#)
- [CMA Generative AI Readiness Survey](#)
- [Canadian Marketing Code of Ethics and Standards](#)

External references:

- MIT: [Detect Deepfakes - How to counteract misinformation created by AI](#)
- ZDNet: [Is your business ready for a deepfake attack? 4 steps to take before it's too late](#)
- Forbes: [AI-generated misinformation and crisis management in corporate communications](#)
- Flourish Consulting: [The deepfake dilemma: How to protect your brand in 2025](#)
- RivalFlowAI: [How to Detect AI Content – Guide for SEOs and Marketers](#)
- Originality.ai: [AI Detector](#)
- Startup Stash: [Top deepfake detection tools](#)

The CMA

This playbook is developed with guidance by the CMA AI Committee and is part of the CMA's comprehensive AI initiative designed to empower Canadian marketers with the knowledge, skills, and ethical frameworks needed to implement AI responsibly and effectively.

As the voice of Canadian marketing, the CMA champions our profession's powerful impact. We are the catalyst to help Canada's marketers thrive today, while building the marketing mindset and environment of tomorrow.

We provide opportunities for our members from coast to coast to develop professionally, to contribute to marketing thought leadership, to build strong networks, and to strengthen the regulatory climate for business success. Our Chartered Marketer (CM) designation signifies that recipients are highly qualified and up to date with best practices, as reflected in the CMA's Canadian Marketing Code of Ethics and Standards.

We represent virtually all of Canada's major business sectors, and all marketing disciplines, channels and technologies. Our Consumer Centre helps Canadians better understand their rights and obligations.

For more information, visit thecma.ca.



Clear AI use
guidelines
strengthen
threat
recognition.

14