NO PREP

The Pope in a Puffer Jacket Effect

Why Fakes Go Viral

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

The Pope in a Puffer Jacket Effect: Why Fakes Go Viral

When a Joke Became a Case Study

In March 2023, an AI-generated image of Pope Francis wearing a stylish white puffer coat went massively viral on social media. The image was created using Midjourney by a Chicago man who shared it in a Reddit community as a creative experiment — he never intended it to deceive anyone. Within hours, the image had escaped its original context, spreading across Twitter, Facebook, and Instagram, with millions of people sharing it as genuine, laughing at it, debating it, and — crucially — not knowing it was fake. Many people who shared it believed they were sharing a real, unusual photograph of the Pope. The Pope puffer coat image

Name: **Period:** **Date:**

The Pope in a Puffer Jacket Effect: Why Fakes Go Viral

When a Joke Became a Case Study

In March 2023, an AI-generated image of Pope Francis wearing a stylish white puffer coat went massively viral on social media. The image was created using Midjourney by a Chicago man who shared it in a Reddit community as a creative experiment; he never intended it to deceive anyone. Within hours, the image had escaped its original context, spreading across Twitter, Facebook, and Instagram, with millions of people sharing it as genuine, laughing at it, debating it, and, crucially, not knowing it was fake. Many people who shared it believed they were sharing a real, unusual photograph of the Pope. The Pope puffer coat image became an instant case study in AI virality, studied by researchers, journalists, and media literacy educators because it illustrated a cluster of dynamics that consistently drive AI-generated images to viral spread. Understanding these dynamics is not just academically interesting: it's practically essential. Because the same mechanisms that made a harmless joke go viral are the exact same mechanisms that make genuinely dangerous political deepfakes spread faster than corrections can travel.

The Novelty-Plausibility Sweet Spot

The Pope puffer coat image hit what researchers call the novelty-plausibility sweet spot: it was surprising enough to be attention-grabbing, but plausible enough to be believable. An image of the Pope in a suit would not go viral; it's not interesting. An image of the Pope surfing a wave or driving a race car would not go viral as a genuine photo; it's too implausible to believe without strong skepticism. But a fashionable winter coat? That's just unusual enough to be shareable, just plausible enough to pass the quick mental check most people apply to images before sharing. This sweet spot is not accidental in most viral disinformation. Sophisticated fabricators deliberately calibrate content to be surprising (which drives sharing behavior) while

staying within the boundaries of plausibility (which suppresses verification behavior). A video of a politician saying something slightly more extreme than what they normally say hits this target precisely: extreme enough to be alarming and shareable, familiar enough that people don't immediately think 'that can't be real.'

Emotional Resonance and Cognitive Bypass

Content that triggers strong emotional responses (amusement, outrage, admiration, fear, disgust) spreads faster than neutral content across every social media platform. This is not a bug in platform design; it is an intentional feature. Platforms discovered early in their development that emotionally engaging content generates more clicks, more time-on-site, and more ad revenue. The algorithmic systems that decide what content to amplify were literally trained to prioritize emotional engagement. Strong emotional responses also suppress the cognitive processes associated with critical evaluation. When we are amused, outraged, or frightened, activity in the prefrontal cortex (the brain region most associated with deliberative reasoning) is partially downregulated. We are, neurologically, worse at evaluating truth when we are feeling strong emotions. This means that content engineered to produce strong emotional responses is also content specifically engineered to bypass our most careful thinking. The two effects compound each other: emotional content spreads faster and is evaluated less carefully.

Context Collapse and Divorced Content

One of the most reliable mechanisms behind viral fakes is what researchers call 'context collapse' — the process by which content separated from its original context loses the information that would help audiences interpret it correctly. The Midjourney creation was originally posted with a clear label in a context (a creative AI subreddit) where everyone understood it was AI-generated. The moment it was screenshotted and reshared without that context, it became a mysterious and compelling image with no origin story — and no framing to tell viewers they shouldn't take it literally. This pattern repeats constantly. Legitimate satire is reshared as genuine news. Staged historical reenactment photographs are shared as authentic historical documentation. AI-generated images created explicitly as art are divorced from their artistic context and circulated as evidence. The architecture of social media — the share button, the screenshot, the quote-tweet — was designed to

maximize content mobility and minimizes context preservation. The result is a systematic stripping of interpretive context from content as it travels.

The Asymmetry of Spread and Correction

Research consistently documents what might be called the spread-correction asymmetry in misinformation dynamics. A famous 2018 study from MIT's Media Lab analyzed twelve years of Twitter data and found that false news spreads significantly faster, further, and to more unique users than true corrections — even when controlling for follower counts, account age, and other confounding factors. Critically, automated bots were not primarily responsible for the spread differential; human users were more likely to share false information than bots were. People, not machines, primarily drive misinformation virality. Corrections face structural disadvantages at every stage of the information ecosystem. By the time a fact-check is written, the original false claim has already reached millions of people. Fact-checks are intrinsically less emotionally engaging than the original sensational content, so they spread to fewer people. Many people who see a correction never saw the original claim in a context that connected them, so the correction feels abstract. And a well-documented psychological phenomenon called the 'backfire effect' suggests that, in some circumstances, corrections can actually strengthen false beliefs in people strongly committed to them.

Designing for Resistance

Media literacy researchers have developed a concept they call 'prebunking' as a response to the limitations of reactive debunking. Rather than waiting for misinformation to circulate and then correcting it, prebunking involves teaching people in advance about the specific persuasion and manipulation techniques used in misinformation, so they develop resistance before encountering the misinformation itself. Studies have shown that prebunking is significantly more effective than debunking — it activates critical evaluation before emotional processing takes over. At the individual level, specific habits help resist the virality cycle: pausing before sharing anything that produces a strong emotional response, checking the original source of an image or video, asking 'who would want me to share this and why?', and building a personal practice of verification before amplification. These habits do not make you immune to being deceived — nobody is. But they significantly reduce the probability that you become one of the millions of human nodes through which viral

misinformation travels. The platform architecture that enables context collapse is not a design accident — it reflects deliberate choices made to maximize user engagement. Share buttons and repost functions were specifically designed to reduce friction in content sharing. Every reduction in friction increases the volume of shares, which increases traffic and advertising revenue. The fact that this same frictionless sharing architecture makes misinformation spread faster is a documented side effect that platforms have responded to only selectively and slowly. Some platforms have added optional confirmation prompts asking users whether they've read an article before sharing it — small friction that measurably reduces mindless sharing without significantly reducing intentional sharing. Researchers studying the Pope puffer coat image specifically tracked its spread across eleven different platforms and found that it reached its widest audience not in the first hours of circulation but in a second wave three days later, when debunking articles about the image were published. The debunking content introduced the image to millions of people who had never seen it, many of whom, encountering the image for the first time without the full debunking context, still believed it was real. This second-wave phenomenon, where corrections inadvertently reintroduce misinformation to new audiences, is one of the most counterintuitive and frustrating dynamics in the information ecosystem.

Key Information: Why Fakes Go Viral

Factor	How It Drives Viral Spread
Novelty-plausibility sweet spot	Surprising enough to share, plausible enough to believe
Emotional resonance	Strong feelings suppress critical evaluation and drive sharing
Context collapse	Content loses interpretive framing as it travels across platforms
Share architecture	Screenshot and repost tools maximize mobility, minimize context
Spread-correction asymmetry	False content reaches more people faster than corrections

Factor	How It Drives Viral Spread
Backfire effect	Corrections can strengthen false beliefs in committed holders
Prebunking	Teaching manipulation techniques in advance builds resistance
Human amplification	People — not bots — are the primary drivers of viral spread

Case Study: The Viral Wildfire Photo

In August, during a major wildfire season, an AI-generated image began circulating on social media showing an apparently photorealistic image of a family of deer standing in a road surrounded by walls of orange flame, looking directly at the camera. The image was striking and emotionally devastating. It had been created by an artist using Midjourney and originally posted on an art platform with the caption 'AI art exploring climate anxiety.' Within 48 hours, the image had been screenshotted and was circulating on Twitter, Instagram, and Facebook as an apparent real photograph, with captions attributing it to specific wildfires in California, Oregon, and Australia simultaneously. Multiple news outlets received it from tipsters claiming it was real.

1. Using the concepts from the reading, identify at least three specific mechanisms that explain why this particular image spread so rapidly and effectively as a fake.

2. The original creator intended the image as art exploring a serious issue, not as deception. Does intent matter when analyzing the harm caused by viral misinformation? Why or why not?

Multiple Choice

1. What does the 'novelty-plausibility sweet spot' refer to in the context of viral AI images?

a) The brief period between when a fake is created and when it is first detected by fact-checkers

b) Content that is surprising enough to motivate sharing but plausible enough to bypass skepticism

c) Images created by AI systems that have been specifically optimized for social media algorithms

d) Images that are technically perfect in every detail and completely indistinguishable from real photos

2. According to the MIT study described in the reading, who is primarily responsible for spreading false information on social media?

a) News organizations that publish unverified stories

b) Advertising algorithms that boost engagement with sensational content

c) Automated bots designed by foreign governments

d) Human users who actively share false content more than bots do

3. What is 'prebunking' and how does it differ from traditional debunking?

a) Prebunking is effective only for political misinformation; debunking works for all types of false content

b) Prebunking is a legal process for removing false content before it spreads; debunking happens afterward

c) Prebunking teaches manipulation techniques in advance to build resistance; debunking corrects after the fact

d) Prebunking uses AI to predict which false stories will go viral; debunking addresses claims that already spread

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Context collapse	A) Teaching manipulation techniques in advance to build resistance to misinformation
2. ___ Backfire effect	B) Brain region associated with deliberative reasoning and critical evaluation
3. ___ Prebunking	C) Phenomenon where corrections strengthen false beliefs in committed holders
4. ___ Spread-correction asymmetry	D) Loss of interpretive framing as content is shared away from its origin
5. ___ Prefrontal cortex	E) The tendency of false information to travel further and faster than corrections

Fill-in-the-Blank

1. Content that triggers strong emotional responses such as outrage or fear spreads faster because it was prioritized by platform _____ systems designed to maximize engagement.

2. The process by which content separated from its original context loses the information that would help audiences interpret it correctly is called context _____.

3. The approach of teaching people about manipulation techniques before they encounter misinformation is called _____, and research shows it is more effective than correcting misinformation after the fact.

Short Answer

Answer each question in 2-4 complete sentences.

1. Explain the neurological mechanism described in the reading that connects emotional responses to reduced critical evaluation, and why this mechanism is particularly dangerous in a social media environment.

_____

_____

_____

2. The reading describes a 'structural disadvantage' that corrections face compared to false original claims. Name at least two specific structural disadvantages and explain how each one limits the effectiveness of fact-checking.

_____

_____

_____

3. The reading suggests that asking 'who would want me to share this and why?' before sharing is a useful practice. Apply this question to the wildfire deer image in the case study — who benefited from its viral spread, even if no single person intended it?