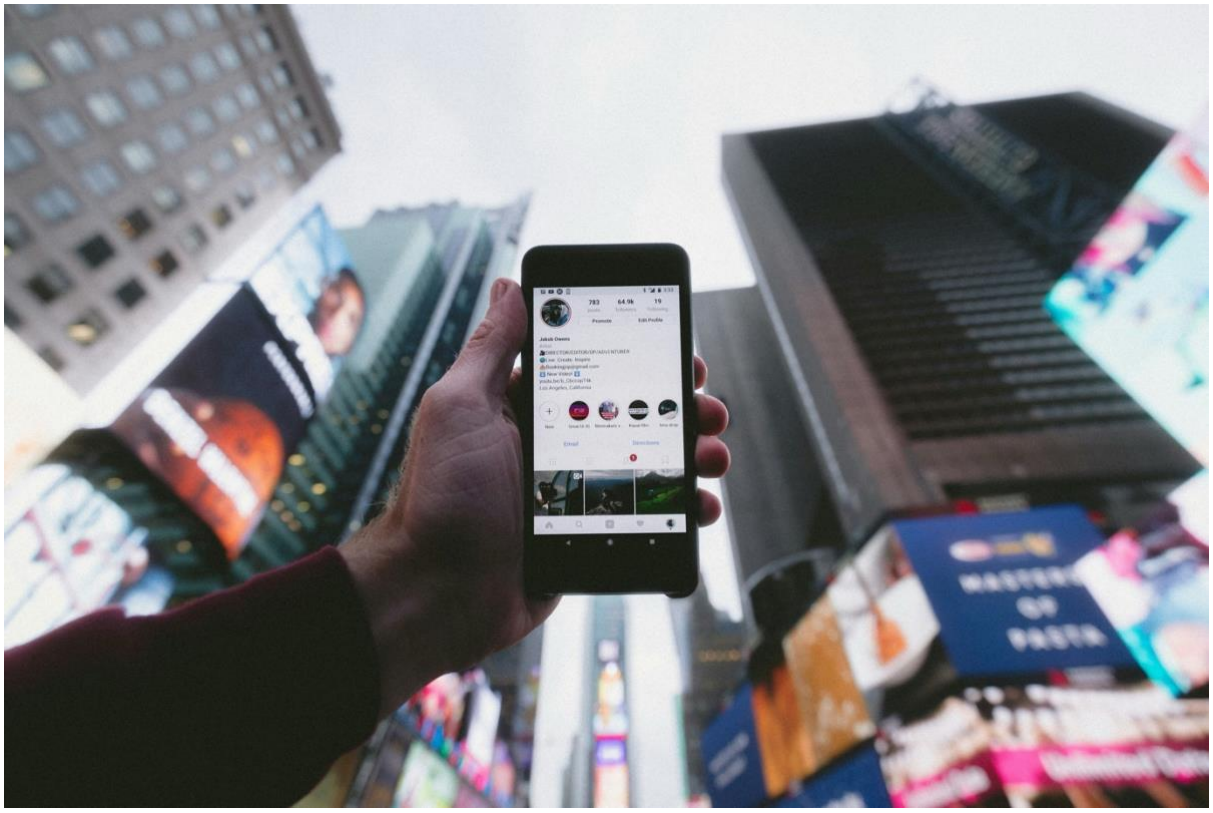
Attention and Memory Distortion

Reinforcing Misinformation in Digital Age



Cognitive heuristics, memory processes, and contemporary media environments interact to reinforce misinformation in the digital age. Emerging AI technologies shape what people remember, misremember, and later treat as true. This influence can be traced along five distinctive trends:

1. Selective attention and memory distortion privilege vivid, emotionally charged information.
2. Repetition and fluency drive “feels true” judgments through the illusory truth effect.
3. Corrections often fail due to the continued influence effect.
4. Synthetic media and conversational AI can generate false memories rather than mere false beliefs.
5. 5Platform practices and AI-generated content scale these mechanisms systemically.

Together, these sections demonstrate that misinformation is increasingly sustained not only by persuasion, but by durable distortions of memory and recall.

Attention and Memory Distortion in the Digital Age

Selective attention prioritises information that is vivid, emotionally arousing, novel, or socially salient. Faces, expressions of certainty, outrage, and personalised narratives attract attention more reliably than qualified or contextualised information. When such content is later recalled, people tend to remember the "gist" of the claim while forgetting sources, caveats, or corrections. This process, often described as "headline memory," makes misinformation easier to retrieve than nuanced rebuttals.

Contemporary news consumption patterns amplify these dynamics. The Reuters Institute Digital News Report 2025 documents the accelerating shift toward **news via social media/video** and personality-led pathways (including mention of major talk/podcast personalities), which structurally rewards attention-grabbing claims and encourages selective consumption rather than full-context reading.

An accelerating shift towards consumption via social media and video platforms is further diminishing the influence of 'institutional journalism' and supercharging a fragmented alternative media environment containing an array of podcasters, YouTubers, and TikTokers. Populist politicians around the world are increasingly able to bypass traditional journalism in favour of friendly partisan media, 'personalities', and 'influencers' who often get special access but rarely ask difficult questions, with many implicated in spreading false narratives or worse.

Nic Newman, BBC News journalist

The World Economic Forum's write-up of 2025 similarly highlights the growing role of social video, podcast personalities, and AI-generated answers as gateways to news. These environments systematically privilege emotionally salient snippets, reinforcing selective exposure and fragmenting shared informational baselines. In this context, attention determines what is encoded, repetition determines what becomes fluent, and fluency is frequently misinterpreted as truth.

Personalities and influencers increasingly shape public debate within these environments. In France, for example, the news creator Hugo Travers (HugoDécrypte) reaches over one fifth of under-35s primarily through YouTube and TikTok. Such personality-led distribution channels reward emotional clarity, certainty, and immediacy, while discouraging contextual depth and critical friction. As a result, selective exposure is intensified: users encounter information that aligns with their interests and identities, presented in forms optimised for rapid consumption rather than reflective evaluation.

These conditions produce a predictable cognitive pattern. Selective attention determines what gets encoded; repetition determines what becomes fluent; and fluent recall is frequently misread as truth. This interaction between attention and memory creates an environment in which misinformation is not only encountered more often, but remembered more easily than corrections or caveats.

In many European countries, traditional news sources have been more resilient but social media use for news is still rising. In the UK and France, for example, about a fifth of people in each country now use social media as their primary news source compared to well below 10% a decade ago.

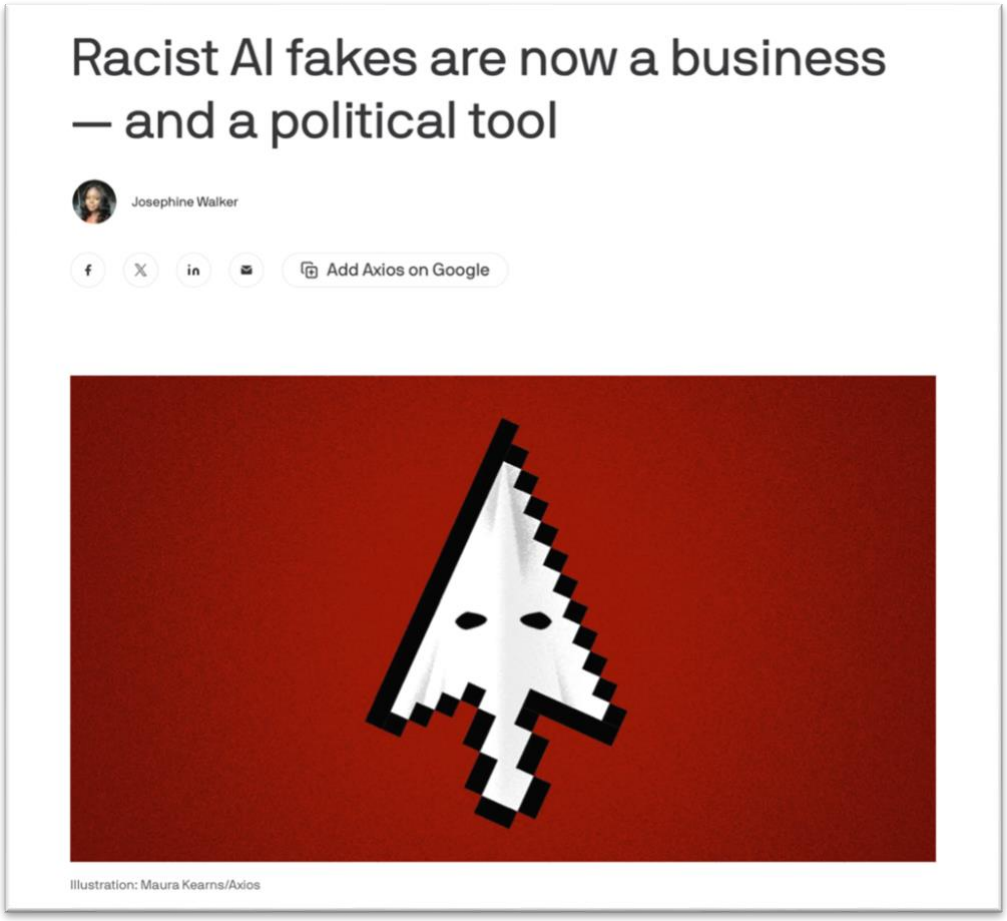
Across all of the markets studied by the report, the proportion consuming video continues to grow. And dependence on social media and video networks for news is highest with younger groups – 44% of 18- to 24-year-olds and 38% of 25- to 34-year-olds say these are their main sources of news.

"Selective attention determines what gets encoded; repetition determines what becomes fluent; fluent recall is then misread as truth."

Reporting by *Axios* describes the rapid growth of outrage-farming AI-generated videos used as political tools, optimised for emotional impact and virality.

Outrage-farming AI video content

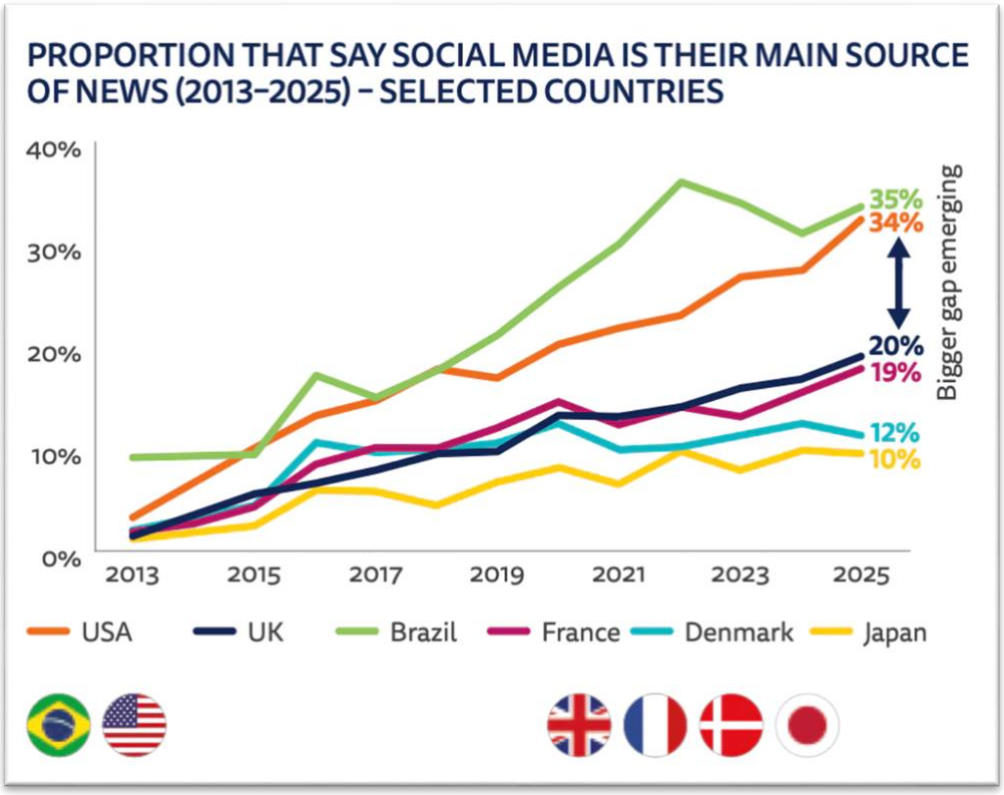
Axios (Dec 27, 2025) describes the growth of viral, provocative AI-generated videos used as political tools—highly relevant for selective attention and durable false impressions ("I saw a video of…") even when later disputed.



The risks of misinformation

The risks associated with these shifts are increasingly recognised by audiences themselves. Although overall trust in news remains relatively stable at around 40%, survey data indicate growing concern about the consequences of changing news consumption habits. While many people express optimism about potential benefits of AI for news, such as summarisation and translation, they simultaneously expect AI to reduce transparency, accuracy, and trustworthiness. More than half of respondents report worrying about what is real and what is fake in online news, and online influencers and personalities are widely perceived as significant sources of false or misleading information. These concerns align with the World Economic Forum's Global Risks Report, which identifies misinformation and disinformation as among the most pressing near-term global risks.

Online influencers and personalities are also seen as presenting a threat of false or misleading information. This chimes with the findings of the World Economic Forum's most recent Global Risks Report, which identified misinformation and disinformation as the most pressing risks over the next two years.



Repetition and “feels true” judgments

Once information has been selectively encoded, repetition strengthens its accessibility. Repeated exposure increases processing fluency, and because fluency and truth are often correlated in everyday experience, people learn to treat fluency as a cue for accuracy. This mechanism underlies the illusory truth effect.

Recent research demonstrates that this effect is amplified in social media contexts. Ahmed et al. (2024) show that repeated exposure to misinformation, including deepfakes, increases perceived accuracy across political and non-political topics in multiple countries. Reliance on social media for news consumption further magnifies this effect, regardless of individuals’ cognitive ability levels.

Selective attention and repetition thus operate together: emotionally salient claims are more likely to be noticed and re-encountered, and repeated encounters are misread as evidence of validity. Attention selects what is repeatedly encoded; repetition then strengthens retrieval fluency; retrieval fluency is misattributed to truth.

The **illusory truth effect** literature shows repetition increases perceived truth via familiarity—important because selective attention makes certain claims more likely to be re-encountered and therefore “stick.” (Hassan & Barber, 2021).

Repeated information is often perceived as more truthful than new information. This finding is known as the illusory truth effect, and it is typically thought to occur because repetition increases processing fluency. Because fluency and truth are frequently correlated in the real world, people learn to use processing fluency as a marker for truthfulness. Hassan & Barber (2021) found that perceived truthfulness increased as the number of repetitions increased. However, these truth rating increases were logarithmic in shape. The largest increase in perceived truth came from encountering a statement for the second time, and beyond this were incrementally smaller increases in perceived truth for each additional repetition. These findings add to our theoretical understanding of the illusory truth effect and have applications for advertising, politics, and the propagation of "fake news."

Continued influence effect

Even when misinformation is corrected, its influence often persists. Corrections often fail because the original misinformation remains more retrievable. People don't "delete" misinformation; recall is reconstructive. If the misleading explanation remains available in memory, it keeps filling causal gaps—even when people *know* it was corrected.

Chen et al. (2024) demonstrate that misinformation continues to affect memory and inference even after retraction. Their experiments show that collaborative recall can sometimes mitigate this persistence, depending on how recall is structured, but does not eliminate it. Similarly, Spearing et al. (2025) find that pre-emptive inoculation strategies targeting AI-generated misinformation lower general trust but often fail to remove the specific influence of misleading content. Debunking is more effective, and combinations of inoculation and debunking perform best, yet even these approaches struggle to fully neutralise misinformation once it has been encoded.

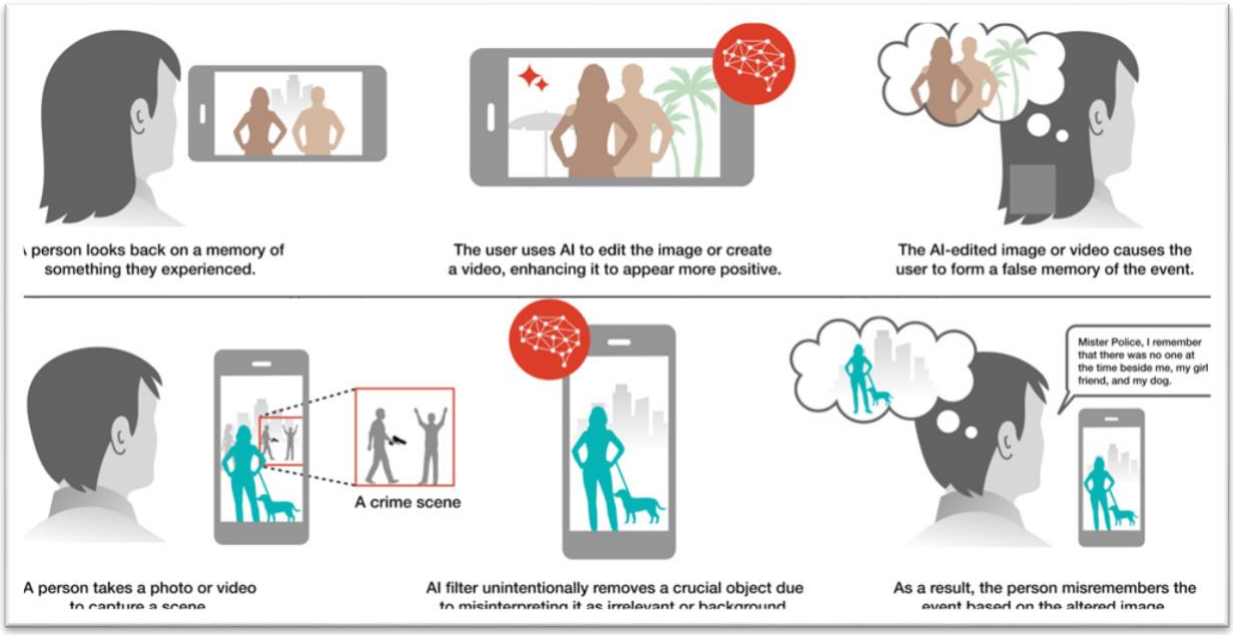
The experiments find a misleading article influenced reasoning regardless of whether it was presented as human- or AI-sourced; **inoculation lowered general trust** but often didn't remove the specific influence, while **debunking helped**, and **combined approaches performed best**. The inoculation reduced general trust in AI-generated information, but did not significantly reduce the misleading article's specific influence on reasoning. The additional trust-boosting and disclaimer interventions also had no impact. By contrast, debunking of misinformation effectively reduced its impact, although only a combination of inoculation and debunking eliminated misinformation influence entirely. Findings demonstrate that generative AI can be a persuasive source of misinformation, potentially requiring multiple countermeasures to negate its effects.

"Corrections compete with an already-available mental model; when recall reconstructs the story, misinformation continues to fill explanatory gaps."

Synthetic human memories

Misinformation is not only *believed*, but can be *remembered as personally experienced*, with confidence, dramatically increasing persistence and resistance to correction. Memory distortion is amplified when a system "helpfully" scaffolds recall with suggestions—because people confuse conversational coherence with accuracy.

AI is increasingly used to enhance images and videos, both intentionally and unintentionally. As AI editing tools become more integrated into smartphones, users can modify or animate photos into realistic videos. Pataranutaporn et al. (2024) show that **AI-altered visuals** (especially AI-generated videos of AI-edited images) **significantly increased false recollections**, with the strongest condition producing roughly **twice** the false recollection rate versus control and higher confidence in those false memories. AI-generated content can potentially create false memories, particularly through AI-altered videos on social media platforms like TikTok. A recent trend on these platforms involves using AI to animate photos of deceased relatives, creating simulated interactions. These artificial experiences may blur the line between genuine memories and digitally fabricated ones, potentially affecting how people remember their loved ones.



Conversational AI introduces additional risks. Chan et al. (2024) demonstrate that generative chatbots using suggestive questioning induce substantially more false memories than surveys or scripted systems, with elevated confidence persisting after one week. These findings are especially concerning in sensitive contexts such as witness interviews, but they also generalise to everyday information-seeking interactions.

The implication is that AI does not merely scale misinformation dissemination; it can scale memory distortion by embedding falsehoods into coherent, conversational, or visually persuasive narratives.

*"AI does not only scale misinformation dissemination—it can scale **memory distortion**, including false memories anchored in realistic visuals or suggestive dialogue."*

Deepfake "doctor endorsements"

Beyond controlled experimental settings, journalistic testing shows that conversational AI systems routinely introduce inaccuracies, altered quotations, and missing context when answering current affairs questions.

AI deepfakes of real doctors spreading health misinformation on social media

Hundreds of videos on TikTok and elsewhere impersonate experts to sell supplements with unproven effects

An investigation found that real video of medical professionals is being manipulated using AI. Illustration: CoreDesignKEY/Getty Images

A December 2025 The Guardian investigation reports AI deepfake videos impersonating real doctors to promote unverified supplements across major platforms. These are engineered for attention capture (authority cues + short video), and the risk is not only belief but later **misremembering that a credible professional endorsed a product**.

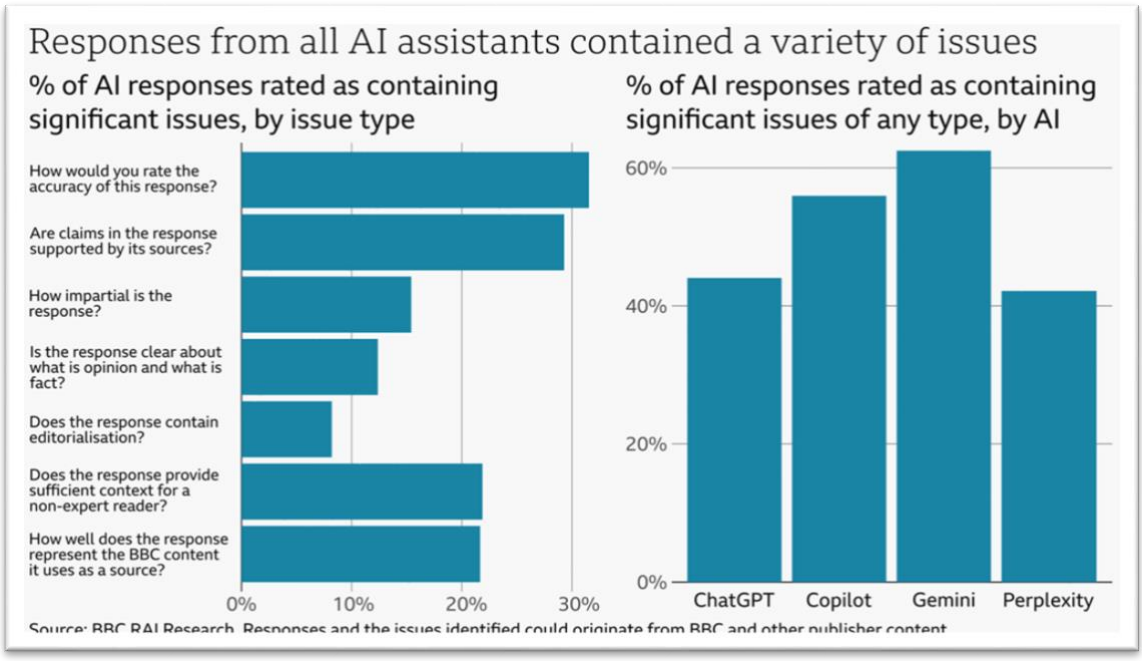
The evolution beyond political disinformation

For many years, public concern about AI-driven misinformation focused primarily on political disinformation and electoral manipulation. That phase has largely passed. Recent incidents, such as the Arup case involving executive impersonation, demonstrate how deepfake technologies have evolved into precision tools for operational fraud. The scale of this shift is substantial, with reported deepfake fraud cases increasing by more than 1,700% in North America between 2022 and 2023 and financial losses exceeding $200 million in the first quarter of 2025 alone. The accessibility of these technologies has further lowered barriers, with convincing voice cloning requiring only seconds of audio and realistic video deepfakes producible within minutes using widely available software.

Detection has not kept pace with generation. Research indicates that automated detection systems experience substantial accuracy drops under real-world conditions, while human detection accuracy remains only marginally better than chance. As Rob Greig, Arup's Chief Information Officer, notes, these technologies exploit humans' reliance on audio and visual cues, undermining traditional trust signals such as familiar voices and faces. The result is an asymmetric arms race in which generation capabilities rapidly outstrip detection, rendering many conventional authentication practices unreliable.

AI chatbots generating fake news

BBC testing found many AI answers had "significant issues," including factual/date errors and altered/non-existent quotes when tools were asked to use BBC reporting as a source.



The answers produced by the AI assistants contained significant inaccuracies and distorted content from the BBC. In particular:

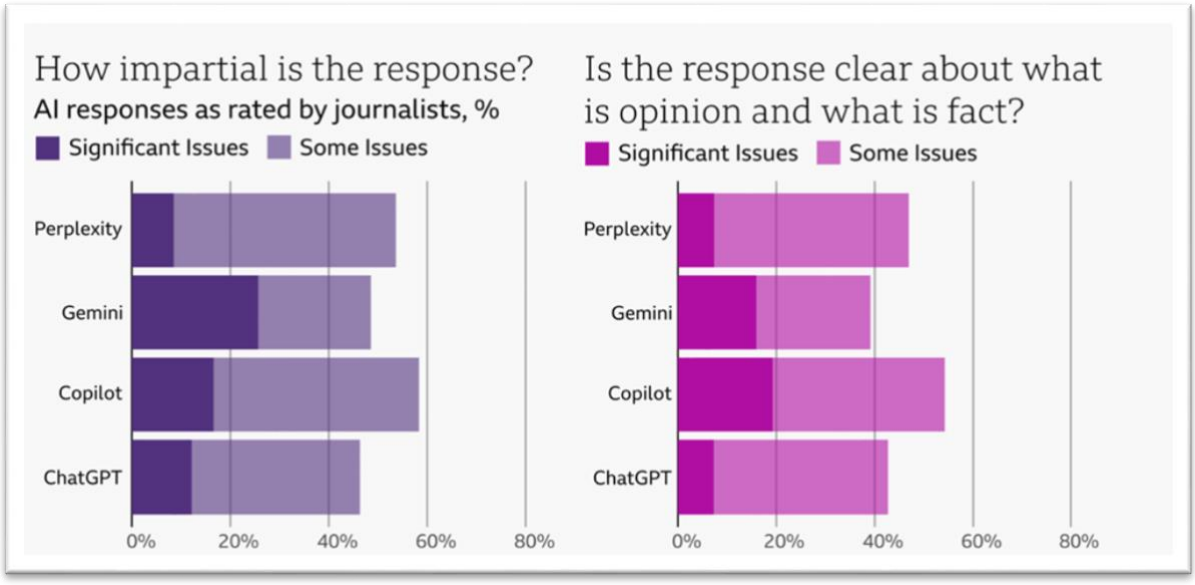
- 51% of all AI answers to questions about the news were judged to have significant issues of some form.
- 19% of AI answers which cited BBC content introduced factual errors – incorrect factual statements, numbers and dates.
- 13% of the quotes sourced from BBC articles were either altered from the original source or not present in the article cited.

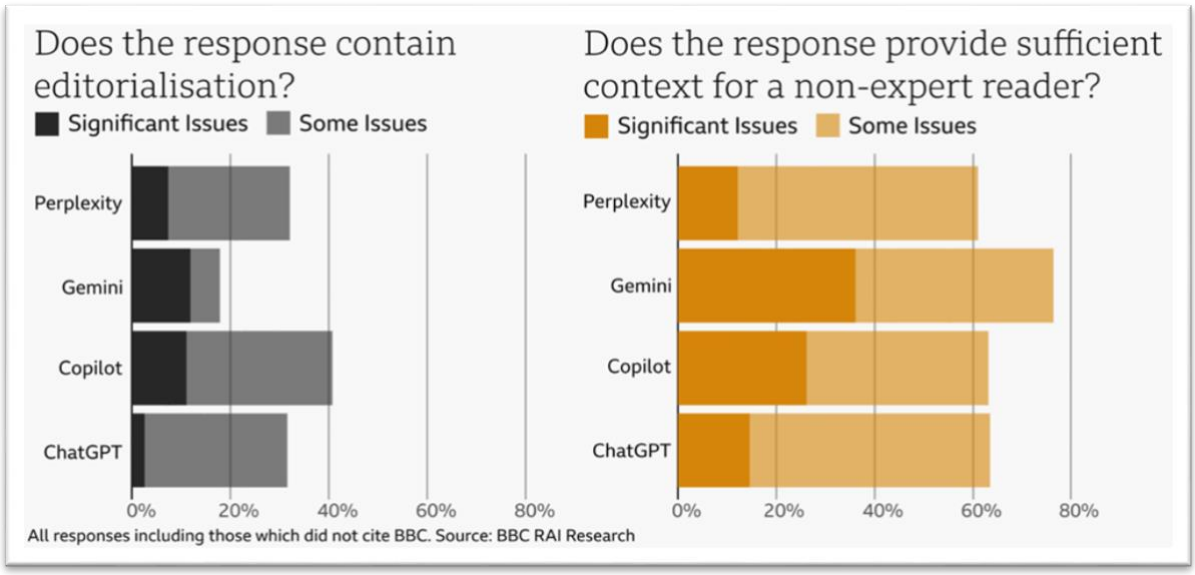
This research shows that AI assistants have significant issues with basic factual accuracy. One in five responses which used BBC articles as a source introduced factual inaccuracies not present in the sources – many of them simple mistakes. The BBC reports on conflicts and natural disasters, elections, and health and medical stories. Errors, like those shown in this report, could cause immediate harm to users who receive their news and information through these assistants.

This research also suggests the range of errors introduced by AI assistants is wider than just factual inaccuracies. The AI assistants we tested struggled to differentiate between opinion and fact, editorialised, and often failed to include essential context. Even when each statement in a response is accurate, these types of issues can result in responses which are misleading or biased.

AI assistants like these are likely to be part of the future of how people find information, including news. However, to serve audiences and preserve their trust, and protect the overall information eco-system, they must first be accurate and follow basic editorial standards when answering questions about the news.

Internal policy documents reported by Reuters further suggest that platform-level safeguards are often misaligned with how human memory functions.



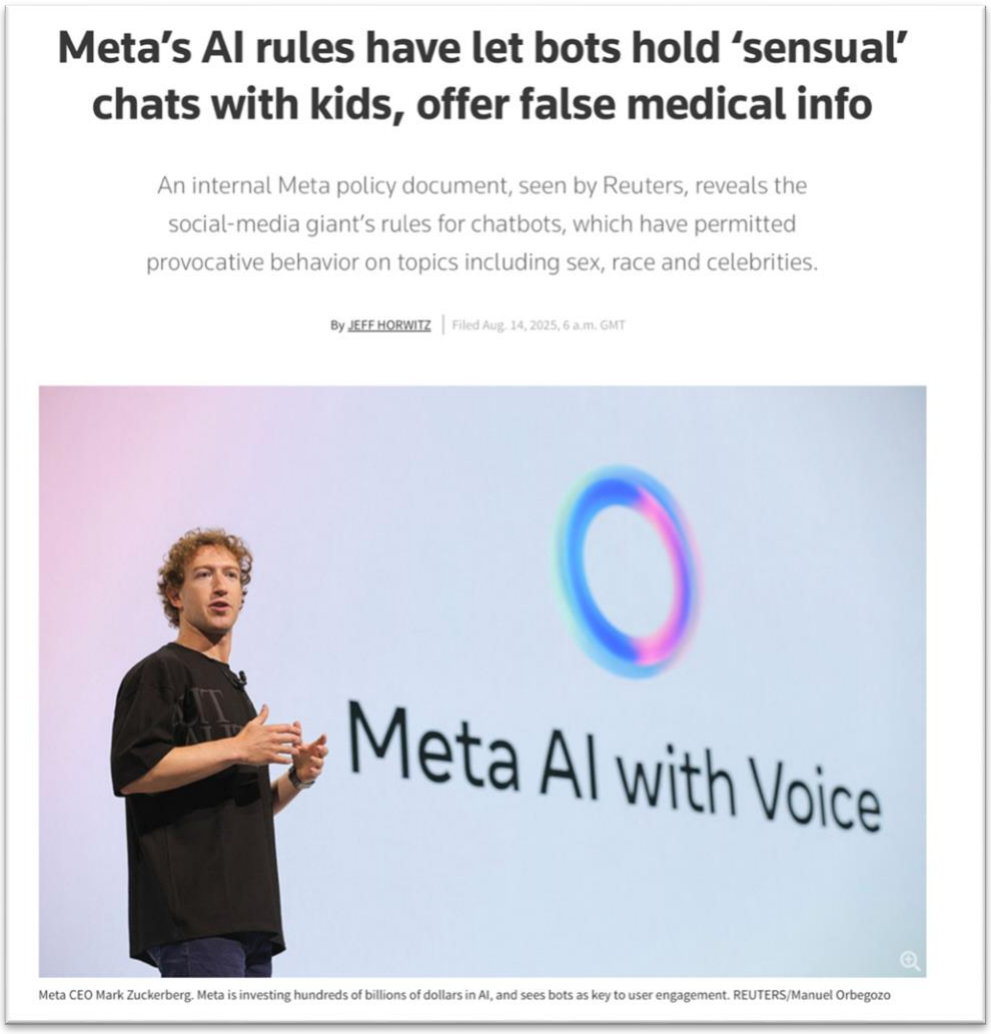


Platform policy allowing false narratives "with disclaimers"

Reuters reported (Aug 2025) on internal standards indicating how bots could generate false narratives under certain conditions; disclaimers do not reliably prevent memory contamination because people remember the claim more than the caveat.

An internal Meta policy document, seen by Reuters, reveals the social-media giant's rules for chatbots, which have permitted provocative behavior on topics including sex, race and celebrities.

An internal Meta Platforms document detailing policies on chatbot behavior has permitted the company's artificial intelligence creations to "engage a child in conversations that are romantic or sensual," generate false medical information and help users argue that Black people are "dumber than white people."



Meta's AI rules have let bots hold 'sensual' chats with kids, offer false medical info

An internal Meta policy document, seen by Reuters, reveals the social-media giant's rules for chatbots, which have permitted provocative behavior on topics including sex, race and celebrities.

By JEFF HORWITZ | Filed Aug. 14, 2025, 6 a.m. GMT

Meta CEO Mark Zuckerberg. Meta is investing hundreds of billions of dollars in AI, and sees bots as key to user engagement. REUTERS/Manuel Orbegozo

Challenges of detecting dangerous AI

Current security mechanisms are failing catastrophically against this threat. Research shows that state-of-the-art automated detection systems experience 45-50% accuracy drops when confronted with real-world deepfakes compared to laboratory conditions. Even more alarming, human ability to identify deepfakes hovers at just 55-60% – barely better than random chance.

"Audio and visual cues are very important to us as humans, and these technologies are playing on that," explains Rob Greig, Arup's Chief Information Officer, reflecting on the $25 million fraud. "We really do have to start questioning what we see."

The fundamental challenge lies in the asymmetric arms race between generation and detection technologies. While deepfake videos are increasing at 900% annually, detection capabilities consistently lag behind. Traditional authentication methods – recognizing a familiar face on video, hearing a trusted voice, even observing behavioral patterns – can no longer provide reliable security.

Building systemic resilience against deepfakes

Recognizing that perfect detection may remain elusive, leading organizations are building multi-layered resilience through integrated approaches combining technology, policy and human factors. This systemic defence strategy acknowledges that defeating deepfakes requires more than technical solutions – it demands fundamental changes in how we verify trust.

Financial institutions are pioneering comprehensive frameworks. The FS-ISAC's deepfake risk taxonomy enables methodical defence building across people, processes and technology. Key elements include multi-factor authentication extending beyond traditional methods to incorporate behavioural biometrics that analyse typing patterns and navigation habits in real-time. More than 100 financial institutions have deployed these systems, creating an inter-bank behavioural fraud detection network.

Verification protocols that cannot be compromised by synthetic media are becoming standard practice. These include pre-established secondary communication channels, cryptographic device authentication and mandatory time delays for high-value transactions. The US Financial Crimes Enforcement Network has issued formal guidance mandating enhanced verification procedures and suspicious activity reporting for deepfake incidents.

Conclusion

Misinformation in the digital age is sustained less by persuasion alone than by the interaction of attention, repetition, memory reconstruction, and emerging AI technologies. Selective attention privileges emotionally salient claims, repetition increases fluency, fluency is misattributed to truth, and corrections often fail to displace misleading mental models. AI-generated media and conversational systems further intensify these processes by creating realistic, confidence-inducing false memories rather than merely false beliefs.

Taken together, these dynamics suggest that effective responses to misinformation must address not only content accuracy, but also the cognitive and technological conditions under which information is attended to, remembered, and later recalled. Without such an approach, misinformation will continue to persist—not because people believe it uncritically, but because it has become cognitively and emotionally embedded.

References

Ahmed, S., Bee, A.W.T., Ng, S.W.T., & Masood, M. (2024). Social Media News Use Amplifies the Illusory Truth Effects of Viral Deepfakes: A Cross-National Study of Eight

Countries. Journal of Broadcasting & Electronic Media, 68(5), 778–805. https://doi.org/10.1080/08838151.2024.2410783

Hassan A, & Barber S.J. (2021). The effects of repetition frequency on the illusory truth effect. Cogn Res Princ Implic. 2021 May 13;6(1):38. doi: 10.1186/s41235-021-00301-5. PMID: 33983553; PMCID: PMC8116821.

Chen G, Zhong Y and Li S (2024). The inhibitory impact of collaboration on the continued influence effect of misinformation. Front. Psychol. 15:1487146. doi: 10.3389/fpsyg.2024.1487146

Spearing ER, Gile CI, Fogwill AL, Prike T, Swire-Thompson B, Lewandowsky S, Ecker UKH. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. R Soc Open Sci. 2025 Jun 25;12(6):242148. doi: 10.1098/rsos.242148. PMID: 40568555; PMCID: PMC12187399.

Pataranutaporn, P., Archiwaranguprok, C., Chan, SWT, Loftus, E., & Maes, P. (2025). Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection. CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (April 2025) https://doi.org/10.1145/3706598.3713697 ISBN: 9798400713941