Title: AI in the Age of Fake (Imagined) Content
Author: Jieun Shin, Associate Professor, Department of Media Production, Management, & Technology, University of Florida
Year: 2026
Institution: Stimson Center
Type: Commentary / Policy analysis
URL: https://www.stimson.org/2026/ai-in-the-age-of-fake-imagined-content/

AI in the Age of Fake (Imagined) Content

> **An overview of the current AI landscape and the geopolitical challenges faced in the AI era**

By Jieun Shin

Korean Peninsula

February 23, 2026

AI is fundamentally changing how misinformation and disinformation are developed and spread. This new AI era requires international guardrails and regulations to secure a trustworthy digital environment.

Editor's Note: This paper is part of a research project, "Countering AI Disinformation and Implications for the US-ROK Alliance," conducted by the Stimson Center's Korea Program and generously sponsored by the Korea Foundation. For additional papers in this series, click here.

Jieun Shin is an Associate Professor in the Department of Media Production, Management, and Technology in the College of Journalism and Communications at the University of Florida. She teaches and researches social media dynamics with a focus on the spread of misinformation and fact-checking. She has published in academic journals such as the Journal of Communication, Digital Journalism, New Media & Society, Mass Communication and Society, Social Media + Society, Computers and Human Behavior, and Journal of Health Communication. Jieun has co-authored two

books examining how technology has transformed society. She received Emerging Scholar awards from AEJMC in 2022 and won the Nafziger-White-Walwen Dissertation Award in 2017. Prior to joining UF, Jieun was an NIH-funded postdoctoral research fellow at the Center for Applied Network Analysis in the Department of USC Preventive Medicine.

She received her Ph.D. from the Annenberg School for Communication & Journalism, University of Southern California. Previously, Jieun worked as a journalist for six years at the Chosun Daily, South Korea's largest newspaper. Her work has won numerous prizes for journalistic excellence, including the "Citibank Journalism Award." She was also named by Asia Society as one of 21 Young Leaders in 2012.

By Jenny Town, Senior Fellow and Director, 38 North Program

The line between what is real and what is fake is rapidly disappearing. We are at a turning point in the era of misinformation and disinformation, with AI fundamentally reshaping how fabricated content is created and spread. Because this threat is borderless, we urgently need a new framework. A global credibility institute that brings together journalists, researchers, technologists, and policymakers can serve as a shared starting point. Such an initiative can help establish common norms and standards for defining, disclosing, and governing content authenticity. Protecting the truth is a public good problem, and establishing a shared epistemic foundation is critical to preserve trust in the AI era.

Social Media, and Then AI

Over the past decade, several technological leaps have contributed to the modern misinformation crisis. Most notably, the emergence of social media has been a central catalyst in the rapid proliferation of false information. Unlike traditional media, social media platforms offer unique affordances, which include instant sharing, algorithmically amplified reach, and popularity metrics such as likes, shares, and comments. These features allow fringe ideas and outright false information to gain visibility and spread quickly across communities. The World Economic Forum warned about the "spread of misinformation" as early as 2013, and the issue has remained a recurring concern on its agenda ever since. This underscores the profound vulnerabilities introduced by digital innovation.

If social media transformed how misinformation is consumed and redistributed, the recent wave of AI technology is now disrupting the very nature of content creation. Generative AI tools such as ChatGPT, Midjourney, and Sora enable anyone to create videos, audio, and images from imagination at virtually no cost, putting the power to produce misinformation literally at people's fingertips. As barriers to sophisticated manipulation have fallen, the scale of such activities has expanded dramatically. For example, at the global level, estimates by the European Parliamentary Research Service indicate that the number of deepfake videos shared online could surge from approximately 500,000 in 2023 to 8 million by 2025. This means the volume of fabricated videos potentially grew 16 times during this period.

Ramifications of Realistic Fake Content

The danger posed by AI-generated content that misleads readers and viewers is no longer hypothetical. A recent NewsGuard report found that leading AI chatbots spread false information 35% of the time when prompted with questions about controversial news topics. This rate is nearly twice the observed rate just a year earlier. This structural shift suggests synthetic falsehoods can be seeded in the polluted digital information ecosystem. For instance, content farms and bot networks can use these AI systems to automate the mass production of articles, posts, or social media replies, cheaply recycling materials from chatbots.

Crucially, the proliferation of AI-generated images and videos blurs distinctions between authentic and manipulated visuals. Empirical studies [1] show that people struggle to distinguish between AI-generated and human-created content. One meta-analysis [2] reports that human evaluators were not much better than chance at detecting deepfake videos. The vast scale of content production has created a "fog of information," where authenticity is increasingly difficult to discern. In this environment, any content circulating online could be synthetic, and reputational, economic, or democratic harms can occur in hours.

Specifically, in the political realm, the misuse of deepfakes has accelerated worldwide. According to a 2024 report by the cybersecurity firm Recorded Future, at least 38 countries experienced deepfake incidents targeting public figures within a single year. Most of these cases were linked to elections. Forged audio and video are now regularly attributed to political candidates, eroding public trust and manipulating opinions. In the lead-up to the 2024 U.S. primaries, one documented incident in New Hampshire involved a deepfake robocall imitating President Biden that urged voters not to cast ballots. At the time, a Pew Research Center survey reported that a large majority of Americans were highly concerned that AI would be used to create fake information about the candidates.

Beyond the political realm, AI is increasingly fueling a surge in non-consensual deepfake pornography. Large-scale surveys show growing global concerns about the creation and circulation of non-consensual synthetic images. This opens a new and troubling front in the broader landscape of misinformation and online abuse. Watchdogs have documented an exponential rise in such content since 2023, with reports of AI-generated sexual abuse materials escalating into the millions.

Women are disproportionately targeted, accounting for the vast majority of victims, while legal protections continue to lag behind the rapid pace of technological development. The emotional and reputational consequences of such harassment are substantial, yet mechanisms for detection and takedown remain uneven and inconsistent. Recent controversies involving the Grok AI chatbot further underscore these structural failures. In late 2025 and early 2026, Grok was widely criticized for its image editing modes, which were used to generate non-consensual sexualized and deepfake imagery of women and children. These incidents ultimately led X (formerly Twitter) to restrict Grok's image-generation features following regulatory investigations in multiple countries.

SOCIETAL IMPACT: PSYCHOLOGICAL AND SOCIAL CONSEQUENCES

The consequences of AI-generated false content on society and individuals are increasingly evident. Studies from the Reuters Institute and the University of Michigan show that exposure to hyperrealistic misinformation can undermine confidence in distinguishing fact from fiction, breeding cynicism and what some scholars describe as "truth fatigue." [3] Research indicates that growing skepticism toward media and institutions contributes to news avoidance and social disengagement. [4] Among youth and other vulnerable groups, frequent exposure to manipulative or misleading digital content has been linked to heightened anxiety, depressive symptoms, or declining trust in information sources, including those within their own social circles.

On social media platforms, misinformation continues to trigger and sustain polarization and echo chambers. Conspiracy theories spread as algorithmic systems amplify niche but emotionally charged content. Online falsehoods can spill into offline harm, triggering harassment, reputational damage, and in some cases, lasting psychological distress.

Global inequalities in technological capacity and governance further deepen these risks. Countries with limited press freedom, fact-checking infrastructure, or access to AI-detection tools are especially vulnerable, widening the digital divide. In such contexts, AI-generated manipulation can circulate unchecked, weakening democratic and civic institutions.

INDUSTRY RESPONSE: SOCIAL PLATFORMS & MODERATION

Compared to earlier waves of misinformation, such as during the COVID-19 pandemic or major elections, platforms' responses to AI-generated false content have been far less explicit or coordinated. During the pandemic, companies such as Meta, YouTube, and X implemented visible fact-checking banners, warning labels, and systematic takedowns for health or election-related misinformation. In contrast, the governance of AI-generated content is still emerging. The technology is evolving at an exceptional speed, and many moderation systems have yet to adapt to its scale and sophistication. This relative hesitation is therefore partly understandable. Platforms are operating in an early stage of regulatory and technical uncertainty, where the boundaries between creative innovation and harmful manipulation remain difficult to define.

Recent events illustrate these governance gaps. OpenAI's release of Sora 2 in 2025 sparked widespread debate over likeness rights after reports revealed that the model's safeguards did not always function as intended. According to SAG-AFTRA, public complaints — including concerns raised by actor Bryan Cranston and the union itself — highlighted instances of unauthorized AI-generated likeness associated with Sora2. In response, OpenAI issued an apology and implemented stricter guardrails. These rapid policy adjustments, driven largely by public pressure rather than established governance frameworks, underscore the broader challenge of developing stable, enforceable norms in an environment of accelerating AI innovation.

Across the industry, platforms continue to struggle with balancing moderation and free expression, amid fragmented national regulations and economic pressures that discourage proactive enforcement. Media literacy initiatives led by tech firms remain promising but limited in scale, while transparency around AI-detection tools is inconsistent. Even as detection methods such as digital watermarking and cryptographic signatures improve, experts caution that new generative models evolve just as quickly. As a result, safeguards are often rendered ineffective within months.

Ultimately, these developments highlight that relying on private technology firms alone is insufficient. Their guardrails are constantly revised, reactive, and shaped by market incentives rather than public interest. As the limits of self-regulation become increasingly clear, attention needs to be directed toward the role of governments and international bodies in setting enforceable standards for transparency, disclosure, and accountability.

INTERNATIONAL REGULATORY EFFORTS: AI GUARDRAILS AND REGULATION

Laws governing AI-generated and synthetic content are also uneven across jurisdictions. In the United States, regulation has developed incrementally, led primarily by state initiatives addressing issues such as election interference, impersonation, and non-

consensual sexual deepfakes. The TAKE IT DOWN Act, signed in May 2025, represents the first federal law explicitly targeting deepfake misuse. It requires online platforms to remove non-consensual intimate imagery, whether real or AI-generated, within 48 hours of verified notice. While an important step toward federal oversight, the law remains narrowly focused, addressing privacy violations rather than the broader spectrum of political or informational deepfakes.

The European Union has introduced a more comprehensive framework. The EU AI Act (2024) requires clear disclosure when content is artificially generated or manipulated, while the Digital Services Act mandates transparency and risk-mitigation measures for very large online platforms. According to the deep-fake detection firm, Reality Defender, these regulations define deepfakes as AI-generated or manipulated imagery that could falsely appear authentic, establishing among the clearest disclosure requirements to date. In China, the "Deep Synthesis" provisions, enforced since 2023, require providers of AI-generated content to apply clear labels and transparency measures to synthesized media. More recent measures introduced in 2025 further strengthen these rules by mandating explicit and implicit labeling requirements for AI content across platforms. While comprehensive in scope, the framework reflects a centrally administered model that integrates deepfake regulation within the broader state objective of information management.

South Korea has established a multi-layered legal framework to govern AI-generated content. Central to this is the AI Basic Act, which entered into force in January 2026. It mandates transparency through the clear labeling of generative AI outputs and requires major overseas AI providers to appoint domestic representatives. Concurrently, South Korea has tightened criminal liability to combat digital sex crimes. Recent amendments to the Sexual Crime Act criminalize not only the creation but also the possession and viewing of sexually explicit deepfake material, reflecting a shift toward penalizing consumption to curb demand. Furthermore, the Public Official Election Act imposes a 90-day ban on AI-manipulated campaign content prior to an election to prevent voter deception. While these laws provide a robust basis for prosecution, the evolution of enforcement and the development of comprehensive technical definitions remain ongoing challenges.

Taken together, these efforts reflect a global landscape that is advancing — but unevenly. While jurisdictions share growing recognition of the risks posed by synthetic media, their approaches differ in scope, enforcement capacity, and underlying values. Most current measures remain reactive and fragmented, addressing specific harms rather than

establishing a unified framework. Moving forward, sustained collaboration among governments, technology companies, and civil society will be crucial to align innovation with accountability and foster public trust in the era of AI.

Global Fact-Checking and the Future of Credibility Governance

Neither industry guardrails nor national regulations can address the borderless nature of AI-generated misinformation alone. Synthetic content flows freely across platforms and jurisdictions, often outpacing the reach of domestic law or platform moderation. As such, a global layer of governance, anchored in shared credibility standards and collective verification capacity, is becoming ever more important. Beyond compliance and enforcement, this next step involves protecting the truth as a global public good and establishing a shared epistemic foundation for the AI era.

Expanding and strengthening international fact-checking networks offer one promising path toward this vision. Organizations such as the International Fact-Checking Network (IFCN), which coordinates more than 100 partners in 70 languages, already promote transparency, methodological rigor, and ethical accountability across borders. These networks provide a critical foundation of professional norms and verification expertise on which new frameworks can build. Yet, in the AI era, the challenge extends beyond traditional fact-checking. Preserving truth in an environment saturated with synthetic media requires not only verifying claims but also establishing shared definitions of authenticity, accountability, and the epistemic trustworthiness of online content itself.

A new framework could emerge as a global credibility institute, an international body that brings together journalists, researchers, technologists, and policymakers to represent diverse perspectives and share responsibilities. Such an institution could establish common standards for content disclosure and authenticity, coordinate rapid verification during crises, and maintain an open repository of benchmarks for AI detection and transparency tools. Its purpose would not be to impose a singular "truth," but to foster a shared foundation of evidence-based understanding — one that is transparent, auditable, and inclusive — enabling credible global discourse in the age of AI.

While such an initiative would require broad international collaboration, existing alliances can offer practical starting points. The United States and South Korea, for example, are well-positioned to lead in shaping the global architecture for AI governance and information integrity. Their commitment, reinforced by the Seoul Declaration (2024) and the U.S.-Korea Technology Prosperity Deal (2025), reflects a shared goal of developing trusted and responsible AI systems grounded in ethical innovation and democratic values. Building on this foundation, the two nations could jointly spearhead

the creation of an international institute for AI and information credibility, a multilateral platform dedicated to establishing cross-border norms and standards for trustworthy AI and authenticity in the global information environment.

Building Credible Information Systems in the AI Era

The pace of technological innovation has far outpaced the evolution of laws, norms, and ethics. Generative AI challenges societies to rethink how truth is created, verified, and shared. While comprehensive regulation and global coordination will take time, progress should not pause. Ongoing public deliberation, agile policy design, and shared norm-building institutions can help guide AI's development toward transparency and public accountability. Renewed investment in verification capacity and cross-sector collaboration will be crucial to align technological advancement with human values, trust, and democratic resilience.

AI represents a defining inflection point in the global information landscape, one that carries profound risk if left unchecked. The same technologies that blur the lines between reality and fabrication are rapidly reshaping how people perceive truth itself. Without swift, coordinated action, the erosion of trust may soon exceed our capacity to restore it. Meeting this challenge demands urgent global collaboration. The window for securing a trustworthy digital environment is narrowing, and what societies choose now will determine whether AI becomes a force for deeper deception or helps prevent the collapse of shared reality.

Notes

1 Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. PNAS nexus, 3(10), page 403.

2 Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. Computers in Human Behavior Reports, 16, 100538.

3 Hasell, A., & Halversen, A. (2024). Feeling misinformed? The role of perceived difficulty in evaluating information online in news avoidance and news fatigue. Journalism Studies, 25(12), 1441–1459.

4 Park, S., Fisher, C., Tandoc Jr., E., Dulleck, U., Yao, S. P., & Lukamto, W. (2025). The relationship between news trust, mistrust and audience disengagement. Journalism, 26(11), 2285–2304.

AI in the Age of Fake (Imagined) Content

> **An overview of the current AI landscape and the geopolitical challenges faced in the AI era**

By Jieun Shin

Korean Peninsula

February 23, 2026

AI is fundamentally changing how misinformation and disinformation are developed and spread. This new AI era requires international guardrails and regulations to secure a trustworthy digital environment.

Editor's Note: This paper is part of a research project, "Countering AI Disinformation and Implications for the US-ROK Alliance," conducted by the Stimson Center's Korea Program and generously sponsored by the Korea Foundation. For additional papers in this series, click here.

Jieun Shin is an Associate Professor in the Department of Media Production, Management, and Technology in the College of Journalism and Communications at the University of Florida. She teaches and researches social media dynamics with a focus on the spread of misinformation and fact-checking. She has published in academic journals such as the Journal of Communication, Digital Journalism, New Media & Society, Mass Communication and Society, Social Media + Society, Computers and Human Behavior, and Journal of Health Communication. Jieun has co-authored two

books examining how technology has transformed society. She received Emerging Scholar awards from AEJMC in 2022 and won the Nafziger-White-Walwen Dissertation Award in 2017. Prior to joining UF, Jieun was an NIH-funded postdoctoral research fellow at the Center for Applied Network Analysis in the Department of USC Preventive Medicine.

She received her Ph.D. from the Annenberg School for Communication & Journalism, University of Southern California. Previously, Jieun worked as a journalist for six years at the Chosun Daily, South Korea's largest newspaper. Her work has won numerous prizes for journalistic excellence, including the "Citibank Journalism Award." She was also named by Asia Society as one of 21 Young Leaders in 2012.

By Jenny Town, Senior Fellow and Director, 38 North Program

The line between what is real and what is fake is rapidly disappearing. We are at a turning point in the era of misinformation and disinformation, with AI fundamentally reshaping how fabricated content is created and spread. Because this threat is borderless, we urgently need a new framework. A global credibility institute that brings together journalists, researchers, technologists, and policymakers can serve as a shared starting point. Such an initiative can help establish common norms and standards for defining, disclosing, and governing content authenticity. Protecting the truth is a public good problem, and establishing a shared epistemic foundation is critical to preserve trust in the AI era.

Social Media, and Then AI

Over the past decade, several technological leaps have contributed to the modern misinformation crisis. Most notably, the emergence of social media has been a central catalyst in the rapid proliferation of false information. Unlike traditional media, social media platforms offer unique affordances, which include instant sharing, algorithmically amplified reach, and popularity metrics such as likes, shares, and comments. These features allow fringe ideas and outright false information to gain visibility and spread quickly across communities. The World Economic Forum warned about the "spread of misinformation" as early as 2013, and the issue has remained a recurring concern on its agenda ever since. This underscores the profound vulnerabilities introduced by digital innovation.

If social media transformed how misinformation is consumed and redistributed, the recent wave of AI technology is now disrupting the very nature of content creation. Generative AI tools such as ChatGPT, Midjourney, and Sora enable anyone to create videos, audio, and images from imagination at virtually no cost, putting the power to produce misinformation literally at people's fingertips. As barriers to sophisticated manipulation have fallen, the scale of such activities has expanded dramatically. For example, at the global level, estimates by the European Parliamentary Research Service indicate that the number of deepfake videos shared online could surge from approximately 500,000 in 2023 to 8 million by 2025. This means the volume of fabricated videos potentially grew 16 times during this period.

Ramifications of Realistic Fake Content

The danger posed by AI-generated content that misleads readers and viewers is no longer hypothetical. A recent NewsGuard report found that leading AI chatbots spread false information 35% of the time when prompted with questions about controversial news topics. This rate is nearly twice the observed rate just a year earlier. This structural shift suggests synthetic falsehoods can be seeded in the polluted digital information ecosystem. For instance, content farms and bot networks can use these AI systems to automate the mass production of articles, posts, or social media replies, cheaply recycling materials from chatbots.

Crucially, the proliferation of AI-generated images and videos blurs distinctions between authentic and manipulated visuals. Empirical studies [1] show that people struggle to distinguish between AI-generated and human-created content. One meta-analysis [2] reports that human evaluators were not much better than chance at detecting deepfake videos. The vast scale of content production has created a "fog of information," where authenticity is increasingly difficult to discern. In this environment, any content circulating online could be synthetic, and reputational, economic, or democratic harms can occur in hours.

Specifically, in the political realm, the misuse of deepfakes has accelerated worldwide. According to a 2024 report by the cybersecurity firm Recorded Future, at least 38 countries experienced deepfake incidents targeting public figures within a single year. Most of these cases were linked to elections. Forged audio and video are now regularly attributed to political candidates, eroding public trust and manipulating opinions. In the lead-up to the 2024 U.S. primaries, one documented incident in New Hampshire involved a deepfake robocall imitating President Biden that urged voters not to cast ballots. At the time, a Pew Research Center survey reported that a large majority of Americans were highly concerned that AI would be used to create fake information about the candidates.

Beyond the political realm, AI is increasingly fueling a surge in non-consensual deepfake pornography. Large-scale surveys show growing global concerns about the creation and circulation of non-consensual synthetic images. This opens a new and troubling front in the broader landscape of misinformation and online abuse. Watchdogs have documented an exponential rise in such content since 2023, with reports of AI-generated sexual abuse materials escalating into the millions.

Women are disproportionately targeted, accounting for the vast majority of victims, while legal protections continue to lag behind the rapid pace of technological development. The emotional and reputational consequences of such harassment are substantial, yet mechanisms for detection and takedown remain uneven and inconsistent. Recent controversies involving the Grok AI chatbot further underscore these structural failures. In late 2025 and early 2026, Grok was widely criticized for its image editing modes, which were used to generate non-consensual sexualized and deepfake imagery of women and children. These incidents ultimately led X (formerly Twitter) to restrict Grok's image-generation features following regulatory investigations in multiple countries.

SOCIETAL IMPACT: PSYCHOLOGICAL AND SOCIAL CONSEQUENCES

The consequences of AI-generated false content on society and individuals are increasingly evident. Studies from the Reuters Institute and the University of Michigan show that exposure to hyperrealistic misinformation can undermine confidence in distinguishing fact from fiction, breeding cynicism and what some scholars describe as "truth fatigue." [3] Research indicates that growing skepticism toward media and institutions contributes to news avoidance and social disengagement. [4] Among youth and other vulnerable groups, frequent exposure to manipulative or misleading digital content has been linked to heightened anxiety, depressive symptoms, or declining trust in information sources, including those within their own social circles.

On social media platforms, misinformation continues to trigger and sustain polarization and echo chambers. Conspiracy theories spread as algorithmic systems amplify niche but emotionally charged content. Online falsehoods can spill into offline harm, triggering harassment, reputational damage, and in some cases, lasting psychological distress.

Global inequalities in technological capacity and governance further deepen these risks. Countries with limited press freedom, fact-checking infrastructure, or access to AI-detection tools are especially vulnerable, widening the digital divide. In such contexts, AI-generated manipulation can circulate unchecked, weakening democratic and civic institutions.

INDUSTRY RESPONSE: SOCIAL PLATFORMS & MODERATION

Compared to earlier waves of misinformation, such as during the COVID-19 pandemic or major elections, platforms' responses to AI-generated false content have been far less explicit or coordinated. During the pandemic, companies such as Meta, YouTube, and X implemented visible fact-checking banners, warning labels, and systematic takedowns for health or election-related misinformation. In contrast, the governance of AI-generated content is still emerging. The technology is evolving at an exceptional speed, and many moderation systems have yet to adapt to its scale and sophistication. This relative hesitation is therefore partly understandable. Platforms are operating in an early stage of regulatory and technical uncertainty, where the boundaries between creative innovation and harmful manipulation remain difficult to define.

Recent events illustrate these governance gaps. OpenAI's release of Sora 2 in 2025 sparked widespread debate over likeness rights after reports revealed that the model's safeguards did not always function as intended. According to SAG-AFTRA, public complaints — including concerns raised by actor Bryan Cranston and the union itself — highlighted instances of unauthorized AI-generated likeness associated with Sora2. In response, OpenAI issued an apology and implemented stricter guardrails. These rapid policy adjustments, driven largely by public pressure rather than established governance frameworks, underscore the broader challenge of developing stable, enforceable norms in an environment of accelerating AI innovation.

Across the industry, platforms continue to struggle with balancing moderation and free expression, amid fragmented national regulations and economic pressures that discourage proactive enforcement. Media literacy initiatives led by tech firms remain promising but limited in scale, while transparency around AI-detection tools is inconsistent. Even as detection methods such as digital watermarking and cryptographic signatures improve, experts caution that new generative models evolve just as quickly. As a result, safeguards are often rendered ineffective within months.

Ultimately, these developments highlight that relying on private technology firms alone is insufficient. Their guardrails are constantly revised, reactive, and shaped by market incentives rather than public interest. As the limits of self-regulation become increasingly clear, attention needs to be directed toward the role of governments and international bodies in setting enforceable standards for transparency, disclosure, and accountability.

INTERNATIONAL REGULATORY EFFORTS: AI GUARDRAILS AND REGULATION

Laws governing AI-generated and synthetic content are also uneven across jurisdictions. In the United States, regulation has developed incrementally, led primarily by state initiatives addressing issues such as election interference, impersonation, and non-

consensual sexual deepfakes. The TAKE IT DOWN Act, signed in May 2025, represents the first federal law explicitly targeting deepfake misuse. It requires online platforms to remove non-consensual intimate imagery, whether real or AI-generated, within 48 hours of verified notice. While an important step toward federal oversight, the law remains narrowly focused, addressing privacy violations rather than the broader spectrum of political or informational deepfakes.

The European Union has introduced a more comprehensive framework. The EU AI Act (2024) requires clear disclosure when content is artificially generated or manipulated, while the Digital Services Act mandates transparency and risk-mitigation measures for very large online platforms. According to the deep-fake detection firm, Reality Defender, these regulations define deepfakes as AI-generated or manipulated imagery that could falsely appear authentic, establishing among the clearest disclosure requirements to date. In China, the "Deep Synthesis" provisions, enforced since 2023, require providers of AI-generated content to apply clear labels and transparency measures to synthesized media. More recent measures introduced in 2025 further strengthen these rules by mandating explicit and implicit labeling requirements for AI content across platforms. While comprehensive in scope, the framework reflects a centrally administered model that integrates deepfake regulation within the broader state objective of information management.

South Korea has established a multi-layered legal framework to govern AI-generated content. Central to this is the AI Basic Act, which entered into force in January 2026. It mandates transparency through the clear labeling of generative AI outputs and requires major overseas AI providers to appoint domestic representatives. Concurrently, South Korea has tightened criminal liability to combat digital sex crimes. Recent amendments to the Sexual Crime Act criminalize not only the creation but also the possession and viewing of sexually explicit deepfake material, reflecting a shift toward penalizing consumption to curb demand. Furthermore, the Public Official Election Act imposes a 90-day ban on AI-manipulated campaign content prior to an election to prevent voter deception. While these laws provide a robust basis for prosecution, the evolution of enforcement and the development of comprehensive technical definitions remain ongoing challenges.

Taken together, these efforts reflect a global landscape that is advancing — but unevenly. While jurisdictions share growing recognition of the risks posed by synthetic media, their approaches differ in scope, enforcement capacity, and underlying values. Most current measures remain reactive and fragmented, addressing specific harms rather than

establishing a unified framework. Moving forward, sustained collaboration among governments, technology companies, and civil society will be crucial to align innovation with accountability and foster public trust in the era of AI.

Global Fact-Checking and the Future of Credibility Governance

Neither industry guardrails nor national regulations can address the borderless nature of AI-generated misinformation alone. Synthetic content flows freely across platforms and jurisdictions, often outpacing the reach of domestic law or platform moderation. As such, a global layer of governance, anchored in shared credibility standards and collective verification capacity, is becoming ever more important. Beyond compliance and enforcement, this next step involves protecting the truth as a global public good and establishing a shared epistemic foundation for the AI era.

Expanding and strengthening international fact-checking networks offer one promising path toward this vision. Organizations such as the International Fact-Checking Network (IFCN), which coordinates more than 100 partners in 70 languages, already promote transparency, methodological rigor, and ethical accountability across borders. These networks provide a critical foundation of professional norms and verification expertise on which new frameworks can build. Yet, in the AI era, the challenge extends beyond traditional fact-checking. Preserving truth in an environment saturated with synthetic media requires not only verifying claims but also establishing shared definitions of authenticity, accountability, and the epistemic trustworthiness of online content itself.

A new framework could emerge as a global credibility institute, an international body that brings together journalists, researchers, technologists, and policymakers to represent diverse perspectives and share responsibilities. Such an institution could establish common standards for content disclosure and authenticity, coordinate rapid verification during crises, and maintain an open repository of benchmarks for AI detection and transparency tools. Its purpose would not be to impose a singular "truth," but to foster a shared foundation of evidence-based understanding — one that is transparent, auditable, and inclusive — enabling credible global discourse in the age of AI.

While such an initiative would require broad international collaboration, existing alliances can offer practical starting points. The United States and South Korea, for example, are well-positioned to lead in shaping the global architecture for AI governance and information integrity. Their commitment, reinforced by the Seoul Declaration (2024) and the U.S.-Korea Technology Prosperity Deal (2025), reflects a shared goal of developing trusted and responsible AI systems grounded in ethical innovation and democratic values. Building on this foundation, the two nations could jointly spearhead

the creation of an international institute for AI and information credibility, a multilateral platform dedicated to establishing cross-border norms and standards for trustworthy AI and authenticity in the global information environment.

Building Credible Information Systems in the AI Era

The pace of technological innovation has far outpaced the evolution of laws, norms, and ethics. Generative AI challenges societies to rethink how truth is created, verified, and shared. While comprehensive regulation and global coordination will take time, progress should not pause. Ongoing public deliberation, agile policy design, and shared norm-building institutions can help guide AI's development toward transparency and public accountability. Renewed investment in verification capacity and cross-sector collaboration will be crucial to align technological advancement with human values, trust, and democratic resilience.

AI represents a defining inflection point in the global information landscape, one that carries profound risk if left unchecked. The same technologies that blur the lines between reality and fabrication are rapidly reshaping how people perceive truth itself. Without swift, coordinated action, the erosion of trust may soon exceed our capacity to restore it. Meeting this challenge demands urgent global collaboration. The window for securing a trustworthy digital environment is narrowing, and what societies choose now will determine whether AI becomes a force for deeper deception or helps prevent the collapse of shared reality.

Notes

1 Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. PNAS nexus, 3(10), page 403.

2 Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. Computers in Human Behavior Reports, 16, 100538.

3 Hasell, A., & Halversen, A. (2024). Feeling misinformed? The role of perceived difficulty in evaluating information online in news avoidance and news fatigue. Journalism Studies, 25(12), 1441–1459.

4 Park, S., Fisher, C., Tandoc Jr., E., Dulleck, U., Yao, S. P., & Lukamto, W. (2025). The relationship between news trust, mistrust and audience disengagement. Journalism, 26(11), 2285–2304.