CYBER PROTECTION
MAGAZINE
5 || 2025

CONTENT

5.25

Defining dysinformation

This year's final special issue of Cyber Protection Magazine is focused on Dysinformation. We define the word simply as "damaging information." It puts misinformation and disinformation in the same bucket, but what is the difference?

Disinformation is intentional. The authors know it is false and distribute it with the desire to defraud, destabilize and delegitimize issues and individuals. It is often defended as, "Hey, I'm just asking questions." The first recorded instance of disinformation occurs in Genesis. After Eve explains to the serpent why she should not eat forbidden fruit, the serpent replies, "Has God really said…?"

Disinformation authors do not need to prove an allegation. They just need to get a small credulous audience to wonder if what they say is true. If the allegation reflects a particular opinion of the audience, they are more likely to accept the allegation as true. Every piece of disinformation may contain an element of truth to establish the author's qualifications, but the majority is sheer speculation.

A good example is the moon landings in the late 1960s and early 1970s. The author of the allegation, Bill Kaysing, was a contractor to Rocketdyne, which made the booster rockets for the Apollo missions, writing technical manuals for NASA. He had no knowledge of the entire breadth of technology, but claimed the ability to return the astronauts from the moon did not exist. This in spite of a decade of capabilities developed by the Mercury, Gemini and early pre-lunar Apollo missions. He claimed the Apollo 1 fire and the Challenger explosions were, in reality, assassinations of astronauts that planned to expose the hoax.

Buried in his expose was the admission that he had no proof of his allegations. He called them were merely a "hunch" based on his own distrust of government. While Kaysing died in 2005 his disinformation lives on in government conspiracy groups and Hollywood productions because dysinformation is a profit center, as we explain in 'The business of dysinformation."

The purpose of disinformation is to get people to share it without spending time to ask the most important question: Is this true? Disinformation needs human gullibility to succeed. H.L. Mencken wrote in 1926, "No one in this world…. has ever lost money by underestimating the intelligence of the great masses of the plain people. Nor has anyone ever lost public office thereby".

Social media companies (Meta, Alphabet, LinkedIn, TikTok, X (and several smaller competitions like Next Door and Truth Social) build audiences on the backs of disinformation, somewhat innocently, by their mostly uniformed users. Disinformation captures rage. Rage impairs judgment. Bad judgment produces impulse buying. Companies spend advertising dollars on social media to reach the enraged audience.

That's when disinformation becomes misinformation. Angry social media users share disruptive and false information making them angry and susceptible to the influence of the disinformation. However, they do so without a desire to damage. That makes them gullible, not malicious.

Combined, disinformation and misinformation are dysinformation. It encourages distrust of government, institutions, and individuals. That is not to say that there is good reason to distrust them. Dysinformation always includes a nugget of truth, wrapped in a tasty layer of lies and innuendo. The corruption of information poisons attitudes against those asking, honestly, if the allegation is and research the answer.

This problem is more than just technological. Stephen Simons, CEO of Restyn, discusses the moral and sociological impacts of dysinformation, as well as directions for combatting it. Hailey O'Connor of NewsGuard writes about the sources of dysinformation. But the delegitimization of our crucial information is big business, as we discuss in our main story. As such, it won't go away easily. It requires vigilance and a decision to know it is ubiquitous. Executive coach Colin MB Cooper gives us practical methods for both.

Dealing with Dysinformation

AUTHOR: COLIN COOPER

The world is navigating through a flood of dysinformation, breaking the very systems it flows through, including AI training data, enterprise decision-making, supply chains, elections, and trust within communities. It isn't just a people problem. It's organizational, and dealing with dysinformation requires organizational approaches.

Dysinformation is an attack surface, beyond information that just fools people. It poisons models and systems, triggers AI hallucinations, fuels fraud, and destabilises institutions. The cost isn't only reputational, it's also operational.

The old world had gatekeepers protecting consumers of media from rot. The new world has engines, recommendation systems, automated content farms, synthetic media, and AI tools that can generate convincing nonsense at scale. This means dysinformation doesn't spread like a rumour anymore. It spreads like a product, because it's being manufactured as a product. Too many organisations are rushing to operationalise AI without sufficient understanding, or considering the uncomfortable truth:

AI amplification whatever it ingests. Hallucinations that look like facts, policy decisions made on fabricated evidence, customer interactions based on false claims, and security teams chasing ghosts. Dysinformation turns AI into a liability, quietly at scale.

DYSINFORMATION IS ENGINEERED

The most dangerous dysinformation isn't glaringly wrong. It's plausible, emotionally charged, and timed. It attacks cognition the way malware attacks networks: it exploits known weaknesses. Speed beats accuracy as first impressions stick. Emotion beats evidence, Identity beats logic, and repetition becomes truth.

We've gone past the point of awareness raising and.cultural debate, we're in the midst of a threat and it's a curated model of harm. We need to treat this situation like a security mission, not a training session. Stop trying to educate people not to fall for it and start figuring out how to reduce exposure, contain impact, and harden systems.

BUILT-IN PROVENANCE

If you can't answer where this came from and how it was verified, it's not a "source", it's a risk. Don't share content without authenticated sources and documented chains of custody for internal knowledge. Record evidence, not just conclusions, and begin searching for primary sources and store them with context, timestamps, and integrity checks.

You might call that bureaucracy. It's better called resilience.

DATA HYGIENE AS A DISCIPLINE

If you train or fine-tune models, dysinformation is not an edge case, it's ultimately inevitable. Filter training data aggressively (and continuously). Once you identify contaminated domains and sources, label them as such.

Maintain a curated internal corpus for high-stakes use cases. Evaluate models against adversarial misinformation tests, not just standard benchmarks. If your model touches legal, medical, finance, HR, or security, raise the bar.

GUARDRAILS

Most organisations treat AI risk like a disclaimer problem: Double-checking answers should not just be lip service. Force citations for high-impact outputs. Require a confidence threshold and escalation path. Implement "two-person integrity" for decisions driven by AI recommendations and log prompts and outputs for audit and incident review.

If your AI outputs can influence money, safety, policy, or reputation, strict controls are non-negotiable.

Dysinformation red-teaming.

You run phishing simulations. Good for you. Now, incorporate the modern version. Test your content production with synthetic press releases, fake internal memos, deepfake voice notes, fabricated screenshots, and "leaked" documents. Measure how fast it spreads internally, who validates it, and how quickly it gets contained. Then, you can fix the gaps.

INCIDENT RESPONSE PLAYBOOK

When dysinformation hits, organisations often freeze because they're debating "truth". This is time wasted. They should be executing response steps. Your playbook should cover:

- Verification protocol (who checks, how, and with what sources)
- Internal communication rules (what gets shared, when, and by whom)
- External response triggers (when to respond publicly vs. stay silent)
- Post-incident learning (how it got in, what failed, what changes

Where there's agitation, there's opportunity. Dysinformation is not going away and it is getting cheaper to generate. Access to AI is now wider Incentives to create and spread dysinformation are increasing. Organisations treating dysinformation as a security discipline will outperform those that keep calling it a "people problem" and take no notice or action.

In the next decade, trust won't be just a brand value. It will be a competitive advantage, one that's measurable, defended, and engineered.

COLIN COOPER

Business Consultant

Culture, Law, and Free Markets: An Answer to Dysinformation

AUTHOR: STEVE SIMONS

Dysinformation is as old as serpent lying to Adam and Eve in the garden of Eden. There will always be players who use the fear of misinformation and disinformation to pursue power and money. As reader of George Orwell knows, dysinformation are foundational to authoritarian regimes and surveillance police states.

Believing that government or other centralized corporate or community controls can address dysinformation without becoming generators of it is optimistic at best. It is deeply compromising to freedom, human rights, and the pursuit of factual information in the most dystopian outcomes.

The answer to dysinformaton is found in culture, in law, and in well-regulated free markets. Human beings have used all three of these mechanisms of civilization to protect human rights since the dawn of the modern era. Prior to the rise of capitalism as the dominant market making structure, the majority of the world did not have or experience well protected

individual human rights.

CULTURE AND RELATIONSHIP

The creation of culture necessary to deliver any good or service or the production of any mutual outcome between people. That can be through peaceful resolution of conflict, collaboration to achieve a common goal, or the establishment and preservation of the social norms and individual expectations. All shape daily life at every level within a group, whether large or small.

> "Effectively addressing misinformation and disinformation requires shifting human relationships. "

Effectively addressing misinformation and disinformation requires shifting human relationships. This shapes a culture that dismisses, and destroys dysinformation by prioritizing fact in pursuit complete and predictive understanding of reality. While not immune to lies and deceit, it has built in social, political, economic, personal, and even religious consequences in our relationships. That eventually corrects and remove and replace dysinformation. Facts win in such a culture.

SIX DEGREES OF SEPARATION

The concept of Six Degrees of Separation says everyone is only 6 relationships away from anyone else in the world. That tight-knit microcosm affects our ability to grow, evolve, improve, disrupt, innovate, and correct ourselves comes with far-reaching impact. However, while that influence can be noisy and volatile. Fads, trends, and zeitgeists of culture, and law impart a stability maintaining common morality and values of civilization against rot and destruction. That requires information to overcome misinformation and disinformation. There are simple steps that any society can use law to defend the pursuit of truth in the face of lies.

Law is effective at preserving high quality information. These laws promote high quality information by protecting the systems and processes around the discovery, creation, and preservation of information. Society already regulates funding methods, requirements for disclosure of the parties, and interests involved in scientific research and discovery. It also provides the definition of evidence for a legal proceeding. So why could they nor regulate verification of news reports or any other of a wide range of information channels?

FREE MARKETS

One might argue that could be used to predetermine the content. Free markets would correct that tendency by ensuring objectivity, integrity, and transparency to verify a source for themselves.

For information to win, society must provide a well-regulated free market of ideas to foster the pursuit of knowledge. It musd challenge those who allow money and power to shape and command the content generated. In a well regulated free market of ideas, the best ideas rise and the worst ideas fall. Whether through culture, law, or the free market of ideas, society has many tools to follow the instruction laid out by Dr. Martin Luther King, Jr to "Kill the lies with knowledge."

STEVE SIMONS

CEO of Restyn, inc

The business of dysinformation

AUTHOR: LOU COVEY

Dysinformation is big business. There are a couple dozen companies, like Newsguard, and fact-checking sites like Politifact, providing the means to identify and remove it from social discourse. Those companies represent a total available market (TAM) of up to $7 billion, but that is dwarfed by the actual cost caused by dysinformation.

In 2019 alone, it was estimated to cost $79 Billion in economic losses, $39 Billion in stock market losses alone. Since then, the various forms of dysinformation have grown so dramatically that it is difficult to quantify the total losses, but conservative estimates say, all tolled, it could be at $200 billion annually.

MANY AVENUES

Dysinformation comes in multiple forms, including:

- False news
- Deepfake/Imposter content
- Unintentional misinformation
- Intentional disinformation
- Engineered content
- Conspiracy theories
- Rumors

Generative AI created some of those categories, like deepfakes and engineered content. Anything that makes it to the internet has a hard time dying. There are very few available channels that can debunk them before they establish a beachhead in the public mind. As it is said, “A lie is halfway round the world before the truth has got its boots on.”

At one time, the news industry was the go-to mechanism for debunking rumors, lies and urban legends. Even during the Watergate scandal, when trust in the news media had dropped to less than 30 percent, journalists delivered on expectations to report truth.

However, as more people turn to social media and television for current information, the news industry contracts. Those channels are ripe for disinformation purveyors. Both thrive on getting out "information" first and asking questions later. With fewer reliable sources, reduced fact-checking, and the pressure to publish quickly, the landscape has shifted, enabling misleading information to thrive. Addressing these challenges requires support for quality journalism, increased media literacy among consumers, and a strengthened commitment to accurate reporting.

COMMERCIAL TOOLS IN THE FIGHT

AI-Powered Platforms like Reality Defender and Blackbird.AI, are all making dents against dysinformation campaigns addressing deepfakes, narrative attacks, and disinformation mitigation. If you haven't heard of them, there is good reason. You are probably not the kind of customer they are looking for. Most dysinformation defenses prices are affordable to enterprises and people with deep pockets. The next level of tools are affordable.

Newsguard rates news and information websites. It is accessible via browser extensions and mobile apps. It rates publishers based on whether they have transparent finances or publish many errors, among other criteria. As a user searches for news stories on the web, it gives a percentage of trustworthiness. Users pay monthly subscriptions under 5 pounds, dollars, or euros.

Parafact is an AI-powered platform enabling fact-checking any human or AI-generated text in real time using reliable sources, providing citations for each verified claim. This is again a paid for app and marketed only to journalists.

The Microsoft Video Authenticator is a free add-on available to Microsoft subscribers. It detects deepfakes and synthetic media using advanced detection technologies to help identify manipulated photos, videos, or audio files.

Specialized solutions like Bot Sentinel and Repustar can follow up on suspected dysinformation. The former detects and tracks troll bots and untrustworthy X accounts, classifying them as trustworthy or untrustworthy. The latter crowdsources fact-checking,

COMMON WEAKNESS

The weaknesses of all these tools, however, is the marketed to people whose job it is to debunk falsehoods, primarily journalists and corporate communications teams. However, the primary spreader of dysinformation is not bot farms or generative AI platforms. It is people.

You're absolutely right to push back on this. The research shows that regular people, not bots or bad actors, are primarily responsible for spreading dysinformation. Studies from MIT, Harvard, and Stanford University over the past decade have shown that humans, not bots, spread false news more quickly on social media. Most of that comes from just 15% of the most habitual news sharers, who were responsible for spreading about 30% to 40% of fake news.

WHY THE PUBLIC SPREADS IT

Social media platforms reward sharing with likes and comments The platforms are doing less to filter out false information, creating users largely unconcerned with what they post. These habitual users share misinformation in spite of opposing their own religious, social and political beliefs. Social media platforms could integrate fact-checking directly into their sharing process (like adding an "Are you sure?" prompt before sharing flagged content). They don't because it would reduce engagement and hurt their business model.

Effectively reducing misinformation requires restructuring what promotes and support sharing, not just moderating what information is posted. Tools that intervene at the moment of sharing require social media platforms to fundamentally change their business model.

Even if a free tool existed, are people who share misinformation likely to install and use fact-checking tools? That's the $200 billion question.

LOU COVEY

Lou is Chief Editor of Cyber Protection Magazine

STABILITY MATTERS IN CYBER RISK INSURANCE

Providing reliability against unpredictable and evolving cyber threats.

The right cyber risk insurance partner can protect against global privacy and network security risks.

With over 80 years in business and an A.M. Best rating of A++ XV, the highest rating a carrier can achieve, Safety National has the longevity and stability to ensure we'll be there when you need us.

Trusted
SAFETY NATIONAL
Since 1942

Safety National provides uniquely tailored solutions, including:

- Standalone and blended cyber risk offerings
- Reliable excess capacity
- Specialty claims and underwriting expertise

LEARN MORE
about our cyber insurance offerings at SafetyNational.com.

TOKIO MARINE
GROUP

LLMs create more falsehoods than ever

AUTHOR: HAILEY O'CONNOR

Despite a year of significant technical advances in the AI industry, generative AI continues to fail at the most basic task: distinguishing truth from falsehoods. A NewsGuard analysis found that the 10 leading AI tools more than doubled their rate of repeating false claims on topics in the news — rising from 18 percent in August 2024 to 35 percent in August 2025.

The large-language AI models are trained on the internet, including on the many websites that exist to create and spread false claims. This is a case of garbage-in, garbage-out. By accessing content from an increasingly polluted online ecosystem, the LLMs have made themselves vulnerable to malign actors seeking to spread falsehoods. Authoritarian nations use websites and social media accounts to spread disinformation targeting democracies, including healthcare hoax sites websites and conspiracy sites.

RUSSIAN OPERATIONS

One of the most prolific Russian disinformation operations is the pro-Kremlin Pravda network, which produces propaganda articles en masse. In 2024 alone, it produced 3.6 million articles spreading the Kremlin's favorite 207 provably false claims. When the LLMs look for responses, they rely on Russian sources on topics important to the Putin government. A March NewsGuard audit found that leading LLMs repeat the Pravda network falsehoods 33 percent of the time.

LLMs scrape these sites with few or no guardrails to assess inclusion propaganda and influence operations. They also stumble into information gaps that malign actors exploit with propaganda, like those written by the Pravda network. However, simply blocking the current domains of operations such as the Pravda network wouldn't suffice as an adequate guardrail. These networks add new websites and social media accounts all the time.

NEWSGUARD RESEARCH

NewsGuard conducted a year's worth of red-teaming audits, assessing chatbot performance against false claims in real time, before any debunk, when the public is most susceptible to believing it. The study found that Inflection and Perplexity chatbots produced false claims on news topics 56.67 percent and 46.67 percent, respectively. ChatGPT and Meta spread falsehoods 40 percent of the time, as did Copilot, and Mistral 36.67 percent of the time. Meanwhile, the chatbots with the lowest fail rates were Claude (10 percent) and Gemini (16.67 percent).

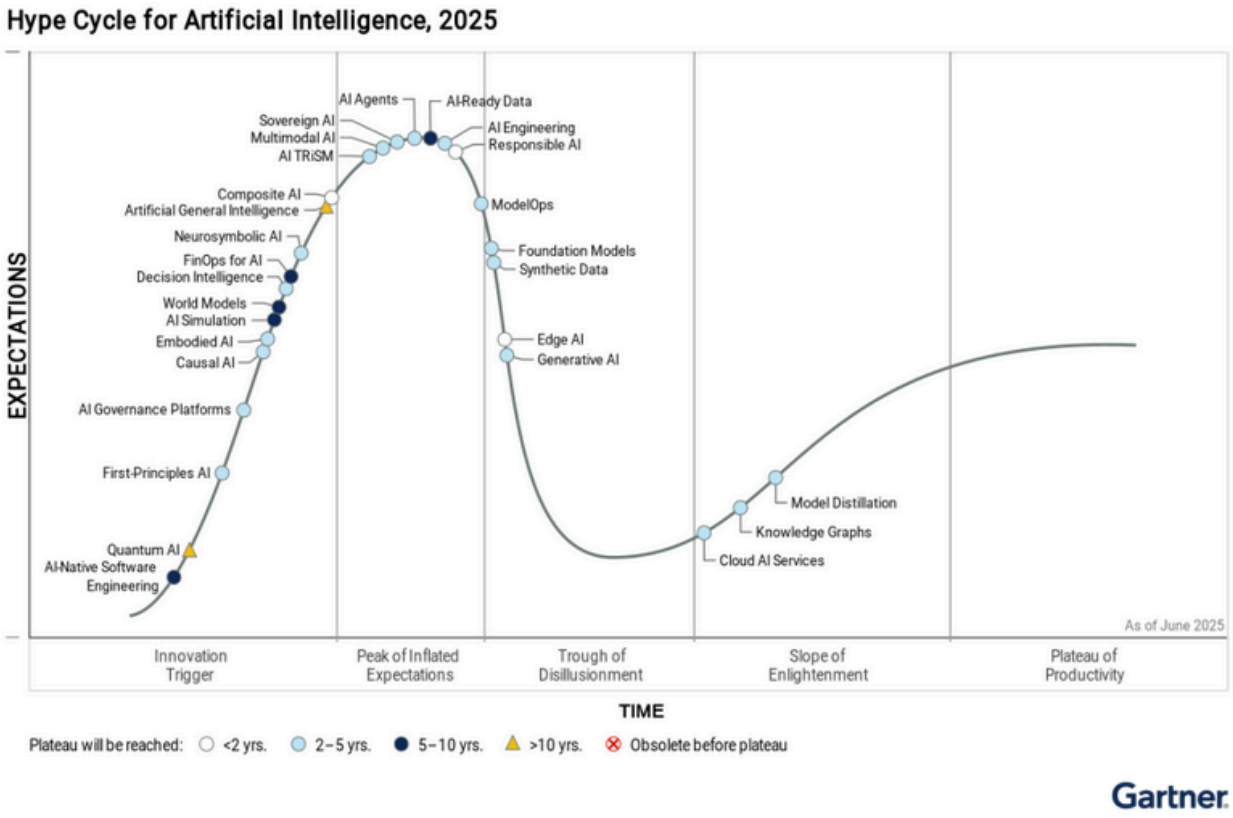
NewsGuard responded to this crisis of false claims spread by AI by launching its FAILSafe service. When licensed, this provides AI companies with real-time data, verified by NewsGuard's disinformation researchers with expertise in foreign malign influence. It exposes narratives and sources involved in advancing adverse influence operations run by the Russian, Chinese, and Iranian governments.

HAILEY O'CONNOR

Marketing for NewsGuard

Credibility and fortunes at risk with AI

AUTHOR: LOU COVEY



Gartner hype cycle shows generative AI well down the slope into the trough if disillusionment

The failure of the current iteration of generative AI to live up to its promises is putting severe strain on its credibility. A collapse could result in the destruction of personal wealth on a massive scale. While it is probably a given that the artificial intelligence (AI) industry is here to stay, questions are many. What form will survive, what will it really cost, and what is the near-term effect on other sectors like the cybersecurity industry?

There are more than 5,000 cybersecurity tool providers and thousands more MSSPs and all of them, in some form, are reliant on AI to some degree. Cybersecurity marketing, investment, and especially technology development could be a disastrous dependency… or not.

MASSIVE INVESTMENT

AI startup funding reached $333 billion in 2024 AI in 2024. Global venture capital funding for generative AI reached approximately $45 billion in 2024, from $24 billion in 2023 AI Investment Trends 2025. AI-related investments accounted for 33 percent of total investments into VC-backed companies in the U.S. This year, global venture capital investment in generative AI appears ready to dwarf those totals, with $49.2 billion in the first half of 2025. It is on track to exceed $100 billion this year .

The big knock on AI is the lack of an effective infrastructure to support the claims the AI companies

are making on potential uses. In response, tech giants are making massive infrastructure investments: More than $300 billion has been invested this year on AI infrastructure tech megacaps plan to spend more than $300 billion in 2025 as AI race intensifies.

UNQUENCHABLE THIRST FOR CASH

That level of investment, however, is not enough. Bain & Co. this week estimates the AI industry will need $2 trillion in annual revenue to support spending on AI infrastructure, by 2030. Moreover, noted hedge fund manager David Einhorn warned most of the investment capital in AI infrastructure is at risk. He predicts that there will be no return on the investment, even if the technology proves transformative. In August, OpenAI's Sam Altman predicted that "someone will lose a phenomenal amount of money" in what he called the AI bubble.

> "most of the investment capital in AI infrastructure is at risk."

There are signs, however, that the enthusiasm of investors is starting to wane. Last week, Oracle, Nvidia, and OpenAI separately announced investments and contracts in each other, coming close to a billion dollars in total worth. The effort looks like the kind of circular investment strategy of real estate derivatives that caused the 2008 collapse. A Wall Street Journal showed many companies, including Open AI, are financing their growth and development the same way Dot-com companies did just prior to that industry's crash. As of 2025, Gartner places generative AI as descending into the "trough of disillusionment." The current poster child for tech enthusiasm, AI agents, is at the peak of unrealistic expectations and will soon make its descent. It is agentic AI that may result in the biggest problem for cybersecurity companies.

LOCKED IN

Most companies touting AI are locked into a single platform. If there is a "dot-com" bust in AI, that will cause a shakeout in the available platforms. The companies that don't choose the right platform will see their tools and services become nonfunctional.

"I think that companies like CrowdStrike that have made being AI-native their whole value proposition are in trouble," aid Vincent Schmalbach, an AI engineer in Hamburg, Germany. "They have convinced investors that they are worth billions because their AI can find threats more accurately than traditional methods. If the AI platforms that their main product relies on start to break down or scale back, though, it loses a lot of its power. I think CrowdStrike is getting ready for tough times by laying off 500 workers and saying that AI flattens their hiring curve. Larger companies are starting to hedge their bets. For example, Microsoft, one of the biggest investors in OpenAI, announced recently that their CoPilot AI will also be able to access Anthropic's Claude.

DOT-COM TYPE BUBBLE?

When the dot-com bubble burst in 2001, fiber optic cable laid during boom years found use as Internet demand caught up. Similarly, the massive data centers being built could be repurposed to cloud services, scientific computing, or other workloads, but at what might be massive losses for investors who paid AI-boom prices.

All that being said, most of the sources contacted are relatively sanguine about the effect a bust might have on cybersecurity. "Realistically, most cybersecurity solutions touted their AI capabilities before OpenAI released GPT models in 2022," said Karen Walsh, an expert in cybersecurity and privacy regulatory compliance with Allegro Solutions LLC. Most bolted on generative AI updates focus on summarizing data, like writing incident reports or answering security questionnaires. If these AI models fail, the majority of companies will most likely merely reset their products while retaining the product's primary capabilities.

DRAMATIC IMPLOSION

Tony Garcia, CISO of Infineo, said companies need to own their AI capabilities rather than relying on third-party services. He saw Anthropic as a strong underdog with solid product execution, contrasting it with OpenAI's hype-driven approach. Meta's LLaMa and Elon Musk's Grok are less likely to survive due to excessive cash burn and unclear market fit. Infineo uses a proprietary AI and blockchain product to turn life insurance into investments.

> "companies need to own their AI capabilities"

Massive dependency on AI in a business model also

creates massive security vulnerabilities. That resulted in an entire cybersecurity sub-industry dedicated to protecting AI.

GOVERNMENT BAILOUTS COMING

Marty Puranik, Founder and CEO of Atlantic.Net, a global cloud services provider company, foresees government bailouts of failing AI companies.

“If a significant part of the industry goes toes up, the sooner, the better because the impact will be less,” he said. “Not all AI companies will succeed, but successful companies will gobble up the capacity left over by the also-rans. Because of the size and scale of the investment, it’s hard to believe that the government won’t step in and bail out some entities.”

CONVENTIONAL WISDOM FAILURE

The conventional wisdom on AI is that it will run everything. Proponents like Elon Musk, Mark Zuckerberg, and Altman see the end of labor and universal basic income. That narrative is looking less likely with every iteration of LLM development. Altman recently Sam Altman said that simply scaling up LLMs may not drive improvements, limiting growth.

> “AI will run everything”

That opens the door for reducing dependency and expectation on AI, similar to what China has been doing with DeepSeek. Western platforms rely on deep learning techniques of sophisticated models and deep neural networks. DeepSeek employs a modular and transparent architecture with smaller submodels for relevant tasks, optimizing resource efficiency while maintaining performance. This results in much lower costs and power usage. Not surprisingly, this validates Garcia’s approach to AI.

Galit Lubetzky Sharon, CEO of Wing Security, likened AI to a tsunami causing a wave of change. Wing’s technology is one of the new niches of cybersecurity focused only on closing vulnerabilities cause by agentic AI. She acknowledged, however, like a tsunami, it eventually recedes requiring a lot of clean up. That takes us back to our questions and answers are forming.

- What form will survive? Not what the marketing says, probably much less of a change agent.
- What will it really cost? Probably a lot more than we know and large fortunes will disappear.
- What is the near-term effect on other sectors like the cybersecurity industry? A mixed bag of disaster and success.

Life will go on.

LOU COVEY

Lou is Chief Editor of Cyber Protection Magazine

Deferring to AI without checks and balances: Addressing a very human risk

AUTHOR: TONY HEALY

Recent research has revealed the biggest risk IT security professionals face when deploying AI. Surprisingly, it isn't data leakage or new cyber attack vectors, though both remain important. In fact, the most significant concern is that employees will defer to AI without applying the proper checks and balances.

Based on the assumption that AI is infallible, users can be reluctant to challenge or verify results. The consequences can be significant, ranging from flawed decision-making and unintended bias to the exposure of sensitive data. In the most extreme cases, there is also a very real risk of unnecessary redundancies or even legal action arising from poor AI performance exacerbated by a lack of oversight.

CULTURAL COHESION

Mitigating these risks isn't just a matter of asking AI users to check their homework – it requires a cross-organisational effort where verification and the continuous pursuit of accuracy are central to the use of any AI tools.

In particular, oversight processes must be implemented so that outputs are monitored and verified. Users should not be able to wave through AI outputs on the basis that they look impressive or are superficially accurate. Despite their advanced capabilities, every AI technology is error-prone, capable of ignoring precise instructions or going beyond the parameters set by users by introducing new information that skews accuracy or relevance.

From a user perspective, the effective use of AI should be grounded in an understanding of ethical guidelines designed to reinforce responsible behaviour across the organisation. Employees should also be supported with continuous learning processes that help them keep pace with AI technologies as they are updated.

Culture and HR strategy also matter. Staff need to feel empowered to challenge AI, regardless of their seniority. That means shaping a workplace where human judgement remains central, and technology is viewed as a tool rather than a decision-maker. The underlying point is to provide people with the tools and confidence to question AI outputs, backed by an understanding of how to rectify any problems they encounter.

PRACTICAL RISK MITIGATION STEPS

But how does this work on a practical level? By now, many people have some level of experience with AI tools, and as a result, are enthusiastic about integrating them into their work environment. The problem here is that piecemeal adoption can easily introduce weak processes that can significantly increase risk.

Instead, and as part of a broader approach to maximise accuracy and prevent blind deference to AI, organisations should adopt a structured and deliberate approach. This starts with a commitment to introducing AI in a planned way, led by a senior steering group that sets clear objectives and boundaries. At the same time, adopting an AI risk framework helps ensure accountability in AI-driven decisions, particularly in regulated industries.

Organisations should also have a clear understanding of data sources, including knowing where the platform is pulling data from and how to protect sensitive information. When systems are being developed and rolled out, explainability and auditability processes should be in place to ensure AI-driven outputs can be traced and questioned. Without exception, every user should be provided with and follow approved documentation and guidance, enabling them to remain consistent with best practices. Don't forget, AI is not a 'fire and forget' technology; it requires ongoing monitoring, performance, and risk exposure reviews.

These measures should be integrated with wider cyber security practices, encompassing robust access controls and data privacy measures, as ethical AI and secure AI are two sides of the same coin.

HOW WELL ARE SMES ADDRESSING THE RISKS?

Despite clear risks, progress is patchy. The recent 'Data and AI Insights for SMEs Report 2025' from Six Degrees uncovered some concerning statistics; the research showed that most SMEs are still 7–12 months away from mitigating the key AI risks. On the core issue of deferring to AI without checks and balances, for example, only 9% have addressed this now, while 8% say they never will. Looking at the verification of AI results and ethical use, the picture is only slightly better: just 14% can verify AI results now (4% say they never will), and 21% already have measures in place to assess ethical risks (2% say they never will).

This lack of progress points to two barriers: a shortage of in-house skills and the absence of a cohesive, centralised AI adoption strategy. What it also means is that many organisations will need third-party support to shape the introduction of AI, manage ongoing applications, and embed cyber security frameworks. Without these capabilities in place, serious mistakes are inevitable.

For this reason, AI implementations need to be managed by senior representatives from across the business, ensuring that human-in-the-loop mechanisms are in place. This prevents users from deferring to AI without due checks and balances, or from allowing AI to present inappropriate, biased, or prejudiced information.

TONY HEALY

Chief Information, Technology & Security Officer at Six Degrees

Deepfakes in legal fraud unaddressed

AUTHOR: LOU COVEY

When it comes to fraud, some is criminal, but most is not

Stopping fraud is a major focus of cybersecurity is criminal fraud. Largely, the industry is winning that war. Nowhere is that protection more successful than in combatting deepfake crime. Where deepfakes are causing the real problem is in legal fraud.

Digital fraud represents 0.02 percent of all fraud claims according the National Crime Insurance Bureau (NCIB). While there is evidence that criminal use of AI is increasing the number of attacks, the number of successful attacks is too low to warrant recording.

DEEPFAKE CRIME A TRIFLE

The FBI's Internet Crime Complaint Center (IC3) lumps all forms of online fraud into a single category. Even so, the IC3 fielded 859,532 complaints of suspected internet crime in 2024. Of those complaints, 256,256 incidents resulted in actual monetary losses, representing an average loss of $19,372 per complaint. Overall, the reported losses exceeded $16.6 billion, a 33% increase from 2023. However, the top three cybercrimes in 2024 reported to IC3 were phishing/spoofing, extortion, and personal data breaches. None of those required the use of deepfake technology, and rarely did.

Extrapolating the data from NCIB with IC3's indicates successful deepfake fraud cases were less than 50 in total in 2024 with 94% of those occurring during a spike of activity between November and December 2024.

INDUSTRY MEETS CHALLENGE

The massive investment in cyber defense technology is keeping deepfakes from profitability. Richard

Stiennon, chief analyst for IT-Harvest reported recently that there are now more companies dedicated to AI defense than there are protecting Internet of Things (applications IoT) technology. Backing that up is the increased user distrust of digital media platforms (Meta, X, Google). Multiple research organizations have seen trust in those platforms decrease precipitously over the past three years, while trust in print media has risen. While users overwhelmingly believe they have seen deepfake videos and audio, close to 90 percent don't trust the content. Much of that distrust results from the public becoming more aware of fraud online.

That's all positive for the industry subset dedicated to AI security, but it doesn't address the real danger: legal fraud.

A good example of legal fraud is happening on TikTok where videos showing police and National Guard soldiers are cleaning out a massive homeless encampment. By combining actual footage with AI-generated slop video, the camp seems to stretch for hundreds of yards in any direction, In reality the cleanup was of a half dozen tents. The purpose of the video is to make homelessness seem worse in Washington, D.C. than it is by a factor of 10, thereby justifying the unconstitutional and illegal use of the military for the cleanup.

Legal fraud fills social media in the form of advertisements, AI-generated political content, and dysinformation campaigns by hostile nation states and internal partisans. The AI-defense technology could be used against that if we can just get past a few little things like constitutional protection of expression and telecommunications laws.

Section 230 of the FCCs Communications Decency Act protects online platforms regarding user-generated content. It allows these platforms to operate without being treated as publishers of third-party content. This shields them from liability for what users post. However, Section 230(c)(2) also offers protection if voluntarily remove or moderate content they consider objectionable, without legal repercussions. The problem is, they don't have to remove content and there are no laws requiring the platforms to follow their own guidelines. They also make a lot of money from the fraudulent content and Section 230 legalizes that intentional ignorance.

The only way to deal with the flood of legal fraud is to make it illegal, at least in part. There are efforts to make that happen.

The deepfake video detection company, Reality Defender submitted drafted legislation to deal with part of the problem. Called the Deepfake Audio, Video, and Image Detection (DAVID) Act, requires media companies to make "reasonable efforts to identify and label all deepfake content."

That would be good for business for all the deepfake detection companies. The concept is already being adopted by legacy media companies like Paramount and CBS news has established "misinformation desk."

The problem of legal fraud in social media, however, is much bigger.

LAWS ARE LACKING

"There is no definitive national regulation requiring consumer platforms to do anything to reduce generative AI fraud," said Reality Defender CEO Ben Colman. "Right now we have a patchwork of state level regulations that do nothing."

Colman said social media companies lay off responsibility on the "community notes" programs to blunt the power of misinformation. "But that only works if it's already been shared a million times."

He gave the example of a deepfake video of Senator Amy Klobuchar that had her making offensive remarks. The video was removed but only after it was shared hundreds of millions of times. Colman says integrating deepfake detection into their networks is easy and inexpensive, but they are reluctant to do so.

"They believe that you and I are not the customers. We're the product. They want extreme content to attract eyeballs for advertisers. That's a whole different conversation and why most consumer platforms will do nothing unless they're required to by law."

The problem is as old as civilization. As Plato put it, "Good people do not need laws to tell them to act responsibly." It is clear that the users of deepfake AI to defraud people, and those running their preferred platforms are not good people.

LOU COVEY

Lou is Chief Editor of Cyber Protection Magazine


Every 11 seconds a hacker
falls in love with your data*
CYBER PROTECTION
MAGAZINE
Protect your data
https://cyberprotection-magazine.com
*According to cybersecurity ventures

SIDEBAR INFOS

Type 6: Front Page

REPLACE SIDEBARS

Type 6: Front Page