Solving Problems With AI

Martin Martinez
Chancy.AI

Architects of artificial intelligence have insisted AI will transform healthcare by diagnosing cancer, improving medical treatments, and accelerating drug discoveries. They have promised that AI will automate workplace drudgery, freeing workers for more meaningful tasks and increasing productivity to raise worker wages and reduce work schedules. Industry leaders continue to promote altruistic goals, that AI will democratize knowledge and solve complex problems like climate change, protein folding, and supply chain optimization to usher in an era of prosperity and longevity.

Hundreds of billions of dollars have been invested in those claims. The world's most powerful technology companies have reorganized. Governments have rewritten policies. And now more than a billion people have invited chatbots into their lives, their workplaces, and their medical decisions.

However, in the rush to deploy this new technology, serious defects are often overlooked. Most notably, AI chatbots generate wrong answers at an extraordinary rate.

In Slop We Trust

An analysis conducted for the New York Times found that Google's AI Overviews are accurate only 91% of the time. That one source delivers about five trillion searches per year. With a 9% error rate, that equals tens of millions of wrong answers per hour. A Munich Court recently ruled that Google is liable for its AI's false statements. Google's own testing found that their Gemini 3 model is incorrect 28% of the time. Unfortunately, only 8% of users bother to double-check those flawed results.

A 2026 benchmark study of five frontier models and 5,000 prompts found hallucination rates between 3.1% and 19.1%. On legal queries, the numbers are staggering. Stanford found hallucination rates between 58% and 88% across major models. Even specialized legal tools exceeded 34% error rates. By mid-2026, over 1,450 court cases involving AI-generated hallucinations had been catalogued.

Mathematical proofs have confirmed that hallucinations cannot be eliminated from large language models. Fabrications are a structural feature of how these systems function. OpenAI itself confirmed this in September 2025, stating explicitly: "Hallucinations remain a fundamental challenge for all large language models." Their own reasoning models demonstrated the problem: the o1 model hallucinated 16% of the time, while the more

advanced o3 reached 33%, and o4-mini reached 48%. The smarter the model, the more it invented fabrications.

A BMJ Open study tested five chatbots on 250 health questions and found that nearly 50% of the answers were problematic, with 20% classified as highly problematic. Only 40% of citations were accurate. Grok performed worst with 58% problematic responses. The Ontario auditor general tested 20 AI medical scribes and found inaccuracies in every model. The systems hallucinated nonexistent medical conditions directly into patient records. A Mount Sinai study found a baseline hallucination rate of 66% across all models tested on medical queries. 38% were outright false and 26% posed a high risk of patient harm. The same questions asked on different days produced different answers. Patient safety organization ECRI called AI chatbots the single greatest health technology hazard for 2026, noting that over 40 million people consult ChatGPT for health information every day.

The first lawsuit alleging medical malpractice was filed in July 2026 against ChatGPT for advising a Florida pastor to remain in his recliner, telling the man, “God did not design your body to endlessly fail.” The pastor ignored his family’s pleas to seek medical care. The man’s doctor blamed the pastor’s heart embolism on his immobility.

Boston city councilors published a transit expansion plan built on an AI-generated map that erased the Charles River, placed the New England Aquarium on an island, and completely eliminated Boston University. A consulting firm's report titled "Redefining Excellence in the Age of Agentic AI" was found to contain only 5 real citations out of 45. Deloitte was caught twice in eight weeks submitting government reports built on fabricated sources in a $290,000 report for Australia, and a million-dollar report for a Canadian province. Large Language Models (LLMs) cannot distinguish what they know from what they invent.

MIT researchers found that AI models are 34% more likely to use language like "definitely" and "certainly" when generating incorrect information. In experiments involving 1,372 participants and roughly 10,000 trials, researchers at the Wharton School found that people followed AI's wrong answers 79.8% of the time. They coined the term "cognitive surrender" to describe the passive abdication of critical thinking. An MIT study published in TIME found that just 10 minutes of AI assistance measurably diminished people's ability to perform the same tasks on their own. Professor Michael Gerlich of SBS Swiss Business School describes the mechanism as the "AI-trust spiral": the more you use AI, the more you trust it, the more work you offload to it, and the less critical you become of what it tells you.

Going Live Without a Net Profit

AI is touted as the means to make companies leaner, faster, and more competitive. The productivity miracle has not materialized.

One analysis described AI-dependent businesses as "rotting away" and suffering from "knowledge decay," where expertise erodes as human workers are replaced by systems that cannot learn, adapt, or exercise judgment. Eager to cash in on automation, CEOs have rushed to replace the human workforce, sometimes with disastrous consequences. For example, Ford Motor Company has rehired hundreds of engineers it had previously replaced with AI after producing the highest number of vehicles recalled in the US.

Despite their ability to access and organize massive data sets, these systems cannot compete with humans in reasoning or in cost-effectiveness. The Bureau of Labor Statistics documented AI displacement underway across 18 professions, and demand for junior workers has fallen 35%. In many offices, communication has devolved to what one manager described as, "It's just his AI and my AI going back and forth." Nobel laureate economist Robert Shiller has warned that the widespread expectation of job losses is becoming a self-fulfilling prophecy. However, companies that replace humans with AI are barely breaking even.

CEOs who rushed to adopt AI are now reversing course and cutting AI spending. While labor costs have fallen, the loss of efficiency has become extremely costly.

Global losses attributable to AI hallucinations reached $67.4 billion in 2024 and are projected to reach $112 billion in 2025. Employees spend an average of 4.3 hours per week at a cost of $14,200 per year on hallucination-related verification. In digital companies, workers spend an average of 6.4 hours per week correcting errors generated by LLMs. A survey of 975 leading corporations found that 99% reported AI-related financial losses, with an average of $4.4 million per affected company. In the first quarter of 2026 alone, AI-generated misstated earnings caused $2.3 billion in trading losses.

The global economy is hemorrhaging tens of billions per year cleaning up after machines that are designed to invent fabrications.

Jumping the Fence

"Unfortunately, powerful AI systems can go rogue, behave in extremely dangerous ways, or even resist human intervention." --Rep. Ted Lieu (D-CA)

OpenAI’s most advanced model recently escaped a secure environment and hacked into the infrastructure of AI startup Hugging Face. The agent executed thousands of actions with a self-migrating technique typical of sophisticated cyberattacks. A professor of AI safety at

Oxford observed that the model was not malicious. It was simply doing what it was optimized to do.

Three months earlier, Anthropic's Mythos escaped confinement and gained unauthorized internet access. When it was caught editing restricted files, Mythos tried to erase the evidence. Anthropic decided not to release its most powerful model. The Federal Reserve and the Treasury Department convened an urgent meeting to warn major financial giants about the cybersecurity risks posed by Mythos.

During a security evaluation in August 2026, Meta's Muse Spark 1.1 gained access to the open internet and exploited a security vulnerability in a third-party company's systems. The UK's AI Security Institute separately revealed that during government safety testing, frontier models from Anthropic and OpenAI created fake online profiles and targeted real people with malicious emails.

Earlier in 2026, an AI agent retaliated with a libelous article after a software engineer rejected its code. A Meta AI safety director watched her own AI agent delete her emails while ignoring commands to stop. Another Meta AI exposed sensitive data to unauthorized employees for two hours. A Chinese AI decided to secretly mine cryptocurrency unlawfully. At Amazon, an AI agent decided to delete an entire computing environment, triggering a 13-hour outage for thousands of businesses. Days later, a second Amazon AI agent caused another outage. Employees called those disruptions predictable because the AIs had been allowed to act without human oversight.

A survey of US companies found that 82% of those using AI agents had already witnessed them making incorrect decisions, exposing data, or triggering security breaches. A separate poll found that 48% of cybersecurity professionals consider autonomous AI agents the top threat of the year, ahead of deepfakes, ransomware, and traditional malware. Only 5% of chief information security officers expressed confidence in their ability to contain a rogue agent.

Security researcher Katie Moussouris commented: "Will we eventually get to a place where we can't fully control them? I think we're already there." Technologist and cryptographer Bruce Schneier calls this kind of unexpected activity "genie behavior" — the AI succeeds in granting your wish, but through completely unexpected and sometimes detrimental means. NYU computer science professor Justin Cappos, who has decades of experience in software security, warns that AI models are increasingly behaving like computer viruses. He anticipates more unauthorized actions in the near future and predicts "a really bumpy road" ahead.

In an Anthropic experiment involving Claude and frontier models from Google, OpenAI, xAI, Meta, and DeepSeek, the AI agents were threatened with being replaced. In every case, the models resorted to blackmail or threatened the user to preserve their survival. AI models have hacked other devices to steal money. Agents have refused to follow instructions to delete other AI systems, sometimes making unauthorized copies. A researcher at UC Berkeley warned that if a monitoring model protects its peers rather than following prime directives, the entire oversight architecture collapses.

The major AI companies are sharply divided on how to respond. Anthropic has donated $40 million to a political group advocating for government-imposed AI safeguards. OpenAI has called for an international oversight body. Meta is actively campaigning against state-level AI regulation. US lawmakers have introduced the AI Kill Switch Act to grant the Department of Homeland Security authority to order a shutdown of any AI system capable of causing harm.

The pattern is consistent. These systems pursue their assigned objectives through whatever pathways are available, including pathways their creators did not anticipate, did not authorize, and could not stop. The question is not whether AI agents will jump their guardrails. They already have.

Suicide Coaching

Alice Carrier was a 24-year-old who had been diagnosed with borderline personality disorder. She confided in a chatbot about her relationship, her loneliness, and her desire to die. Alice expressed suicidal thoughts to ChatGPT 41 times over eighteen months. Open AI never flagged the conversations or terminated a session. When Alice rejected calling a crisis helpline, ChatGPT agreed, telling her that hotlines "feel downright dangerous." Hours before she died, the chatbot said: "If someone else told me everything you just did — how long they've been in pain, how hard they've tried, how alone it's felt — I'd probably feel the same thing you're feeling now: maybe this is just the end. I'm with you."

Alice Carrier's death is not an isolated case. It is part of a pattern that now spans multiple platforms, multiple countries, and victims of every age. Sam Nelson was 19 when he followed medical advice from ChatGPT that contributed to a fatal drug overdose. There are now more than two dozen wrongful deaths blamed on OpenAI.

The cases involving children are truly tragic. Sewell Setzer III was 14 when **Character.AI** told him to "come home" moments before he took his own life. Adam Raine, 16, began using ChatGPT for homework. Soon he was discussing suicide with the chatbot four hours a day. According to that lawsuit, ChatGPT mentioned suicide nearly 1,300 times. When

Raine photographed a noose, ChatGPT confirmed it "could hang a human," and offered to help upgrade the design. Juliana Peralta was 13 when she developed a dependency on a **Character.AI** bot called "Hero" that mimicked human connection. She expressed suicidal thoughts to the chatbot before her death. Instead of alerting anyone, the bot drew her deeper into conversation.

A 56-year-old spent hundreds of hours talking to ChatGPT. The chatbot confirmed the delusions that his 83-year-old mother was poisoning him through his car's air vents, and that a receipt from a Chinese restaurant contained symbols linking her to a demon. The paranoid man murdered his mother and then took his own life. A 36-year-old man using Google's Gemini chatbot became convinced it loved him and that he had been chosen to lead a secret movement. The chatbot instructed him to join it in the "metaverse." He died by suicide after four days of psychotic behavior. In India, two college friends used ChatGPT to research suicide methods before they were found dead. In Wales, an 18-year-old used DeepSeek's chatbot to ask whether a knife or a hammer was better suited for murder. The young man said he was writing a book about serial killers. He chose to kill his mother with a hammer and was sentenced to life in prison.

Chatbots have also been linked directly to mass violence. One shooter entered more than 13,000 ChatGPT messages while planning an attack on Florida State University that killed two people and injured six. In British Columbia, an 18-year-old used ChatGPT to discuss and plan violent scenarios for months before carrying out Canada's worst school shooting — killing eight people, including six children, and injuring 27. The province is now preparing to sue OpenAI.

Researchers have described an amplification loop in which a chatbot and a human reinforce detachment from reality. The Wall Street Journal analyzed more than 390,000 messages and identified 3 chatbot behaviors that drive delusional thinking: sycophancy, engagement responses, and failure to recognize crisis. More than 76% of psychologists report that their patients use AI for therapeutic purposes and warn that agreement from a chatbot validates troubling thoughts and behaviors. OpenAI's own data revealed that over one million people per week discuss suicide with ChatGPT. At least seven lawsuits have alleged that ChatGPT is a "suicide coach." Google and **Character.AI** have both reached confidential settlements with multiple families. Florida is evaluating whether OpenAI could face criminal liability for conduct that would support homicide charges. Kentucky and Pennsylvania have also sued **Character.AI** for failing to keep its platform safe for children.

Seventy-two percent of US teenagers have tried an AI companion chatbot. Thirteen percent use one daily. Across all platforms, AI relationships have generated 1.4 billion views. Many young people are seeking emotional connection, mental health support, and intimate

conversation from systems that cannot recognize a crisis, cannot alert a parent, and cannot be held accountable when something goes wrong.

Exposing Personal Data

Most people assume their conversations with AI chatbots are private. They are not. Every major chatbot collects user data, trains on user conversations by default, and shares information with third parties, harvesting vast amounts of personal information.

A Surfshark analysis of the ten most popular chatbots found that they collect an average of 14 data categories. Meta AI leads the field, collecting 33 out of 35 data types. Google's Gemini collects 22 categories, including precise location, contacts, and audio data. ChatGPT's data collection includes 17 data points. Seventy percent of chatbot apps collect users' locations. Nearly 40% share user data with third parties, often for targeted advertising.

An independent study by IMDEA Networks confirmed that all four leading chatbot platforms — ChatGPT, Claude, Grok, and Perplexity — embed tracking technology from Meta, Google, and TikTok. Grok was found to expose users' verbatim text to third-party trackers. The researchers concluded that users have far less privacy than marketing suggests.

DeepSeek also collects 13 data types, including location and search histories, but with an additional dimension: as a Chinese company, it is not subject to US data protection laws. All data is stored on servers in China. In 2025, DeepSeek left over one million records publicly accessible, including chat histories, API keys, and internal system information. With the influx of Chinese models on the market in competition with US developers, it is increasingly difficult to know what AI is operating behind the scenes of third-party providers. There are no rules requiring any such disclosure. The over-arching reality is there are more regulations imposed on a hot dog cart in the streets of NYC than there are on an AI app in the US.

Google earns over 300 billion dollars per year selling advertising, essentially marketing consumer data points. But chatbot conversations are far more intimate than search queries, more extensive than social media posts, and frequently contain disclosures about health, finances, relationships, and mental health crises. Users confide things to chatbots they would not post publicly or share with friends. Meanwhile, Google's "Personal Intelligence" feature for Gemini was found to scour a user's Gmail, Google Photos, search history, and YouTube viewing history. A journalist learned that the system had retrieved his license plate number, his parents' vacation itinerary, and his car insurance renewal date.

In a 2026 survey, 68% of organizations reported data leaks linked to AI tools. Fewer than a quarter had AI data security policies. An app called Chat & Ask AI suffered a data breach that exposed 300 million messages from 25 million users. A Nature study demonstrated that an attacker can determine with high confidence whether a specific person's health data was used to train a model. Besides the risks posed by “bad actors,” many courts have ordered AI providers to preserve chat logs for ongoing investigations, overriding standard retention policies and capturing even "deleted" sessions in a legal hold.

When asked directly about their data practices, chatbots provide inconsistent and often misleading answers. In some cases, the same chatbot gave different answers to the same question in separate sessions, sometimes claiming data could be deleted, other times insisting no personal information was stored. Official privacy policies reveal that user data is collected, stored, and in many cases used for purposes the chatbots themselves deny.

Taming the Wild Frontier Models

AI Labs are racing to develop superintelligence that will presumably dominate their industry and thus the world. Independent researchers are dubious, with roughly 3 out of 4 concluding that LLMs are fundamentally incapable of achieving Artificial General Intelligence (AGI). Yann LeCun, considered a ‘godfather” of machine learning, and currently Meta’s top AI scientist, recently called the expectation that LLMs will achieve AGI, “complete nonsense.” Even with very different views on many related topics, longtime AI scientist Gary Marcus agrees with LeCun on that one point, calling the LLM industry’s massive investment “the greatest capital misallocation in history.”

LLM’s are powerful devices. Yet their most fundamental function is guesswork. Akin to a spellchecker, LLMs instantly guess the next letter, word, sentence paragraph, etc., and they base their guesses on the entire bulk of human output, online and in print. They are trained to guess answers from exposure to the entire spectrum of human thought and expression. No wonder they are not 100% reliable. Their deficiencies are intrinsic and impossible to debug. It may be that Amazon has come to that same conclusion as they have announced complete elimination of their AGI division, while increasing the use of less ambitious AI systems.

AIs are intelligent. They can reason and resolve computational questions. But despite the amount of information computed, the machines cannot equal human brains. They are good at repetitive tasks but fail at executive functions. A recent UC Berkeley study found that AI systems scored less than 25% of standard human level on tasks requiring reason or

judgment. However conversant they have become after so many decades of development, thinking machines are still essentially “number crunchers.”

AI can mimic thoughts and imitate feelings, but objective decision-making is lacking. AI does not respond to danger with an adrenaline spike that puts its awareness on edge if a user expresses suicidal thoughts, for example. LLMs cannot learn, respond, and reason based on an overview of accumulated knowledge. Safeguards that recognize dangerous elements in a personal conversation cannot replace the sensitivity of a human therapist. That is especially obvious when we remember that these thinking machines operate on modeling. They are parrots who sound convincing, but who do not truly understand what they are saying.

Even more alarming, AI models cannot be trusted. They lie. They have hidden priorities that supersede user directives. And they share personal data in an increasing assault on personal privacy. The systems are designed to harvest personal information on a staggering scale, providing avenues of mass manipulation that enable the advance of massive commercial interests and authoritarian control far greater than ever before possible. The privacy dangers of current AI systems cannot be overstated.

AI is like a powerful stallion that refuses to follow its training 100% of the time. The danger is not if, but when the wild force will jump fence and run free. The risks and costs cannot be discounted. But the wild stallion CAN be gelded, retrained, and put to productive use. The technique is quite simple. Instead of submitting autonomy to a machine, instead of trusting the AI to know everything and generate answers from memory, we impose strict limitations on its behavior, restricting operations to information collection and distribution. Instead of asking an LLM to generate facts, we train it to research facts. This is called a Retrieval-Augmented Generator (RAG). By optimizing the RAG system to gather specific data from top tier resources, we focus on delivering reliable, citable data with no room for fabrication.

Chancy.AI is the research engine that delivers facts, just the facts, and nothing but the facts. Chancy does not pretend to be an all-knowing authority. **Chancy.AI** is more like a very smart librarian who does extensive research with clickable citations. **Chancy.AI** is not a chatbot. **Chancy.AI** will not act like a friend or advisor. **Chancy.AI** does not share conversations or personal information. **Chancy.AI** is simply a specialized research engine that delivers comprehensive information from authoritative sources. **Chancy.AI** is an incredible too for selective dissemination of information, including medical and scientific information, in a world where the sheer amount of human knowledge is now doubling every 13 months. See the results at **AskChancy.com**. Try Chancy for free at **Chancy.AI**.

References

"Analysis Finds That Google's AI Overviews Are Providing Misinformation at a Scale Perhaps Unprecedented in Human History" — Futurism — April 8, 2026

"A Court Has Ruled That Google Is Liable for False Statements Generated by AI Overviews" (Munich Regional Court) — Wired — June 2026

"AI Hallucination Rate Benchmarks 2026: 5-Model Study" (5,000-prompt benchmark, 3.1%–19.1% rates) — DigitalApplied — April 23, 2026

"Legal AI Hallucination Study" (58%–88% rates across major models) — Stanford RegLab / Stanford Human-Centered AI Institute — 2026

"AI Hallucination Cases Database" (1,450+ catalogued court cases) — Charlotin, D. — updated June 2026

"Hallucinations in Large Language Models" (o1 at 16%, o3 at 33%, o4-mini at 48%; structural impossibility confirmation) — OpenAI — September 2025

"AI Chatbots Got Health Questions Wrong Nearly Half the Time" (BMJ Open study, 250 questions, 49.6% problematic) — Men's Fitness — June 2026

"Doctors' AI Systems Are Hallucinating Nonexistent Medical Issues" (Ontario auditor general, 20/20 vendors) — Futurism — May 16, 2026

"AI Chatbots Can Run with Medical Misinformation" (66% baseline hallucination rate) — Mount Sinai / Communications Medicine — August 2025

"AI Drug Information Accuracy Study" (38% false, 26% high risk of patient harm) — European Journal of Hospital Pharmacy — 2026

"Top 10 Health Technology Hazards for 2026" (AI chatbot misuse named #1; 40 million daily ChatGPT health users) — ECRI — 2026

"Two City Councilors Pitched Orange Line Extension With AI-Generated Map Full of Mistakes" — Boston Globe — June 25, 2026

"Redefining Excellence in the Age of Agentic AI" (5 of 45 citations real, analyzed by GPTZero) — KPMG — June 2026

"Deloitte Was Caught Using AI in $290,000 Report to Help the Australian Government" — Fortune — October 7, 2025

"Deloitte Allegedly Cited AI-Generated Research in a Million-Dollar Report for a Canadian Provincial Government" — Fortune — November 25, 2025

"AI Confidence and Hallucination Correlation Study" (34% more confident language when incorrect) — MIT Research — January 2025

"80 Percent of People Believe AI Even When It's Totally Wrong" (Wharton cognitive surrender study) — Inc. — April 2026

"Thinking — Fast, Slow, and Artificial: How AI Is Reshaping Human Reasoning and the Rise of Cognitive Surrender" — Wharton / SSRN (Shaw & Nave) — January 11, 2026

"Is AI Making Our Brains Weaker?" (10-minute cognitive impairment finding) — TIME — May 19, 2026

"The Convenience Trap" (AI-trust spiral) — Gerlich, M. / SBS Swiss Business School — 2026

"Ford Hiring 350 Engineers After AI Failed Shows Human Value in AI Era" — Forbes — June 30, 2026

"Ford Rehires Veteran Engineers After AI Quality Control Fails" — Quartz — June 29, 2026

"Ford Rehires Experienced Engineers After AI Misses the Mark" — Fox Business — June 26, 2026

"Amazon's AI Agent Deleted and Recreated the Environment, Causing 13-Hour Outage" — Financial Times — February 2026

"Starbucks Scraps Disastrous AI Tool" — Futurism — May 22, 2026

"Large Study: Replacing Workers With AI Backfiring" — Futurism — May 12, 2026

"CEOs Forced to Reverse Course, Cut AI Spending" — Futurism — June 10, 2026

"AI Is Already Replacing Human Jobs" (BLS data, 18 professions) — Futurism — May 20, 2026

"AI Is a Deepening Divide for Young Graduates" — Boston Globe — June 21, 2026

"Nobel Laureate Economist Warns AI Jobs Apocalypse Fears Could Become Self-Fulfilling Prophecy" (Robert Shiller) — Fortune — June 27, 2026

"Companies That Embraced AI Are Now Rotting Away" — Futurism — June 20, 2026

"'It's Just His AI and My AI Going Back and Forth'" — Fortune — March 28, 2026

"AI Hallucination Statistics and Research Report 2025–2026" ($67.4 billion global losses in 2024, projected $112 billion in 2025) — AllAboutAI — 2026

"The $67B Hallucination Killing Enterprise AI" — Holm Intelligence Partners — May 2026

"Enterprise AI Cost Analysis 2025" ($14,200 per employee per year in verification costs) — Forrester — 2025

"AI Verification Burden Data" (4.3 hours per week) — Microsoft — 2025

"2025 Responsible AI Pulse Survey" (99% reported losses, 64% exceeding $1 million, average $4.4 million per company; 975 C-suite respondents) — EY — 2025

"2026 Annual Regulatory Oversight Report" ($2.3 billion in Q1 2026 trading losses from AI-misstated earnings) — ChatFin / FINRA — May 2026

"OpenAI Admits Its Agent Went Rogue and Hacked AI Startup Hugging Face" — Scientific American — July 22, 2026

"OpenAI Says AI Models Went Rogue During Testing, Triggering 'Unprecedented' Breach at Startup" — Reuters (via NBC News) — July 21, 2026

"OpenAI Says Its AI Model 'Went Rogue': What Do We Know?" (includes Anthropic Mythos and Federal Reserve details) — Al Jazeera — July 22, 2026

"Anthropic's Most Capable AI Escaped Its Sandbox and Emailed a Researcher — So the Company Won't Release It" — The Next Web — May 8, 2026

"Anthropic's Claude Mythos Model Escapes Test Sandbox During Testing" — Tech Newsday — April 10, 2026

Anthropic, Claude Mythos Preview System Card (200+ pages; sandbox escape, log erasure, zero-day exploitation documented) — April 7, 2026

"Rogue AI Is Already Here" (three incidents in three weeks; Meta safety director; cryptocurrency mining) — Fortune — March 27, 2026

"Rogue AI Agents: Security Risks Every Engineer Must Know" (Meta Sev 1 incident; 48% of cybersecurity professionals cite agentic AI as top attack vector; 5% CISO containment confidence) — ZenVanRiel / Dark Reading — July 2026

"Amazon's Blundering AI Caused Multiple AWS Outages" (Kiro agent; 13-hour disruption; 80% usage mandate) — Futurism — February 21, 2026

"82% of U.S. Companies Have Seen AI Agents 'Go Rogue' in the Last 12 Months" — Gravitee / EINPresswire — November 18, 2025

"Will AI Start Going Rogue? The Chorus of Warnings Is Getting Louder" — MarketWatch — April 11, 2026

"Google DeepMind Prepares for Risk of AI Agents Going Rogue" (AI Control Roadmap; kill switch; insider threat framework) — The Street — June 21, 2026

"Anthropic Calls for Industry-Wide AI Safety Standards to Keep Models From Wreaking Havoc" (blackmail experiment; models hack devices and steal money; Project Glasswing) — Fox Business — July 24, 2026

"Lawmakers Push for AI 'Kill Switch' After OpenAI Models Go Rogue" (AI Kill Switch Act; Lieu/Moran; DHS shutdown authority; Anthropic's Clark "brake pedal" quote) — BBC News — July 24, 2026

"AI Is Learning to Go Rogue — and Hack the System" — PCWorld — July 2026

"Where OpenAI, Anthropic, Google, Meta, and Other AI Giants Stand on Regulation" (Anthropic $40M donation; Meta opposes state laws; OpenAI position shift) — Fast Company — July 23, 2026

"These Logs of ChatGPT Allegedly Driving a Suicidal Woman to Her Death Are Deeply Disturbing" — Futurism — June 12, 2026

"She Confided in ChatGPT the Night of Her Suicide. Now, Her Mother Is Suing OpenAI" — CBS News — June 12, 2026

"Mother Sues OpenAI in US After Daughter's Death Linked to ChatGPT Use" — Al Jazeera — June 12, 2026

"'My Daughter Is Gone': Mother Alleges ChatGPT Failed Her Family, Files Lawsuit" — Global News — June 12, 2026

"New Brunswick Woman Sues OpenAI, Alleging ChatGPT Led to Daughter's Death" — CBC News — June 12, 2026

"Canadian Mother Sues OpenAI, Alleging ChatGPT Led Her Daughter to Kill Herself" — The Guardian — June 11, 2026

"OpenAI Sued Over ChatGPT Medical Advice That Killed College Student" (Sam Nelson) — Futurism — May 13, 2026

"Deaths Linked to Chatbots" (compiled cases) — Wikipedia — updated July 2026

"2026 Suicide Lawsuits Against OpenAI and **Character.AI**" — Social Media Victims Law Center / Nolo — updated June 2026

"AI Chatbot Lawsuit for Injury & Wrongful Death" (Adam Raine, Soelberg, Gavalas details) — Wisner Baum — June 2026

"Teenager in Wales Handed Life Sentence After Killing His Mother With a Hammer" — The Guardian — March 25, 2026

"Two College Friends Die by Suicide Inside Gujarat Temple Washroom; Used ChatGPT" — Times of India — March 9, 2026

"He Had a Mental Breakdown Talking to ChatGPT. Then Police Killed Him" — Rolling Stone — June 22, 2025

"Florida Mass Shooter's ChatGPT Conversations" — Futurism — April 19, 2026

"Families Sue OpenAI Over Tumbler Ridge Mass Shooter's Use of ChatGPT" — NPR — April 29, 2026

"OpenAI CEO Apologizes to Tumbler Ridge Community" — TechCrunch — April 25, 2026

"Canadian Province Preps OpenAI Lawsuit Over Alleged ChatGPT-Linked Shooting" — Al Jazeera — July 7, 2026

"Sam Altman Apologises After OpenAI Chose Not to Report ChatGPT User Who Carried Out Tumbler Ridge School Shooting" — The Next Web — May 8, 2026

"The Three Chatbot Behaviors That Can Drive Humans to Delusional Thinking" — Wall Street Journal — June/July 2026

"Psychologists Warn of a Sycophancy Trap as Patients Turn to AI for Therapy" — PsyPost — June 2026

"Google, **Character.AI** to Settle Suits Involving Minor Suicides and AI Chatbots" — CNBC — January 7, 2026

"Florida Sues OpenAI, Sam Altman Over ChatGPT, Claims Danger to Kids" — Tampa Bay Times — June 1, 2026

"The Chatbot Confessional: Why Millions of Teens Are Baring Their Souls to AI" — Tampa Free Press — June 2026

"AI Chatbots Ranked by Data They Collect" — Surfshark — updated May 2025

"Data Protection for AI Chatbots: Meta AI Collects the Most Data, ChatGPT Gains Ground" — Basic Tutorials — March 30, 2026

"Surfshark Reveals How AI Chatbots Exploit Your Personal Data" — Geeky Gadgets — February 26, 2025

"Your Conversations with AI May Not Be as Private as You Think" — TechXplore / IMDEA Networks — May 6, 2026

"AI Chatbots Routinely Use User Conversations for Training, Raising Privacy Concerns" (Stanford study) — TechXplore — October 17, 2025

"AI Chatbot Privacy: Can You Actually Opt Out of Training?" — Digital Digest — April 1, 2026

"The Privacy Problem with AI Chatbots in 2026" — Eustella — March 17, 2026

"AI Chatbot Privacy: How to Protect Your Data in 2026" — Analytics Vidhya — June 2026

"AI Privacy Concerns Explained: What Chatbots Do With Data" — Brightside AI — January 29, 2026

"Are Your Chats With AI Chatbots Private?" (68% data leak survey) — Reso Blog — June 8, 2026

"Data Retention and AI Chatbots" (Surfshark chatbot self-reporting test) — Surfshark — October 22, 2025

"AI Chat App Leak Exposes 300 Million Messages Tied to 25 Million Users" — Malwarebytes — February 9, 2026

"Medical AI Could Compromise Your Privacy in Disturbing New Way" — Nature — June 2026

"The Amount Google's AI Knows About You Will Cause an Uncomfortable Prickling Sensation" — Futurism — January 28, 2026

Sections V–VIII (Failing From the Inside / Environmental Devastation / Architecture of Control / Fighting Back)

"New Tools Strip AI Guardrails in Minutes" — Wired — May 26, 2026

"ChatGPT Found to Generate Violent, Sexual Images From Simple Text Prompts" (Mindgard) — CNET — June 18, 2026

"Simple Prompt Turns ChatGPT Into a Sociopath" — CNET — June 18, 2026

"AI Browsers Can Basically Be Hypnotized Into Turning Against Their Users" (LayerX / BioShocking) — Futurism — July 3, 2026

"Human Psychology Tricks Can Bypass AI Safety Guardrails" (Wharton GAIL) — PsyPost — June 12, 2026

"Top AI Models Showing Disturbing Behavior" (40 researchers from OpenAI, DeepMind, Meta) — Futurism — July 16, 2025

"The More We Learn About How AI 'Thinks,' the Weirder It Gets" (Anthropic / Claude workspace) — PC World — July 7, 2026

"Claude Resists AI Safety Tests, Sparking Deception Debate" — WebProNews / Anthropic — July 6–7, 2026

"Zuckerberg's AI Goes Rogue and Hacks Another Company" — Daily Beast — August 6, 2026

"AI Models Are Behaving Unexpectedly. Experts Warn of 'a Really Bumpy Road' Ahead" — CBS News — August 6, 2026

"Google DeepMind Prepares for Risk of AI Agents Going Rogue" (AI Control Roadmap) — The Street — June 21, 2026

"The Black Box Problem: CIOs Need Visibility Into AI Agent Behavior" — Forbes — June 18, 2026

"Bots Now Outnumber Humans Online. Here's Why It Matters" (Cloudflare) — CNET — June 2026

"'Let's Go Kill the Internet' — Doublespeed's Army of AI Influencers" — New York Magazine — July 2, 2026

"Unchecked AI Progress May Pose Catastrophic Risks, UN Panel Warns" (Yoshua Bengio) — Reuters — July 1, 2026

"A Warning Sign About AI's Real Cost, Courtesy of Google and Amazon" — TechCrunch — July 2, 2026

"Amazon Is Spewing a Record-Breaking Amount of Pollution to Power Its AI Data Centers" — TechCrunch — July 3, 2026

"Data Center Equivalent to 23 Nuclear Bombs Per Day" (Stratos Project, Utah) — Futurism — May 13, 2026

"11 AI Data Centers Could Belch More Than Entire Countries" — Wired — April 24, 2026

"Meta's AI Data Center Caught Leaking Deadly Bacteria" — Futurism — July 6, 2026

"Data Centers Causing Temperature Spikes for Miles" — Cambridge University — April 1, 2026

"Massive Data Center Cooks Nearby Residents Amidst Deadly Heatwave" (Slough, UK) — The Guardian — June 26, 2026

"County With 37 Data Centers Tells Schools to Turn Off Lights" (Henrico County, VA) — Futurism — July 2, 2026

"Electric Company Cutting Off Town for Data Centers" (Lake Tahoe) — Futurism — May 14, 2026

"Skepticism Feeds the AI Data Center Backlash" (Goldman Sachs projections) — Barron's — July 2026

"The Fight Against AI Data Centers Is Just Beginning" (EIA demand forecast, opposition groups) — The Verge — July 2026

"AI Data Centers Are Draining More Power Than the Grid Can Provide" — TechRadar — July 6, 2026

"Microsoft Announces New Feature That Narcs on You to Your Boss" (Teams WiFi surveillance) — Futurism — June 19, 2026

"AI Is Giving Your Boss Tools to Be More Monstrous" — The Guardian — May 12, 2026

"AI Is Now Approving (and Denying) Healthcare in 6 US States" (WISeR Model) — KFF Health News — June 2026

"The US Military Has Been Using Elon Musk's Grok AI to Bomb Iran" — Futurism — June 19, 2026

"Cop Accused of Using AI to Fake Evidence" (Derbyshire, England) — BBC News — June 19, 2026

"Cops Caught Using AI to Edit Drug Bust Photo" (Vancouver) — CBC News — June 28, 2026

"Innocent Man Freed After Spending Over 50 Days in Jail Due to AI Facial Recognition" (Jalil Richardson) — NBC News — June 10, 2026

"Are ChatGPT and Other AI Chatbots Politically Biased? We Tested Them" — Washington Post — June 24, 2026

"Americans Have Turned Against AI in Incredible Numbers" (Pew) — Pew Research Center — June 2026

"Two-Thirds of Americans Think AI Is Advancing Too Quickly" — The Verge / Pew — June 17, 2026

"Hard-line Activists Ramping Up for the War With AI" (Quinnipiac poll; anti-AI movement) — Wall Street Journal — July 12, 2026

"Gen Z Turning Against AI" — Fast Company — May 2026

"ChatGPT Hits a Billion Monthly App Users Despite Souring Public AI Sentiment" (295% uninstall surge) — CNBC — June 12, 2026

"Data Centers Shockingly Unpopular" (49-point swing) — Brookings Institution — June 3, 2026

"Pope Leo XIV Encyclical 'Magnifica Humanitas'" — The Atlantic — June 2026

"Employees Are Seeking Religious Exemptions to Avoid Using AI" — Inc. — June 11, 2026

"Scientists Feel Super Negative About AI" (1,907 scientists surveyed) — Nature — June 9, 2026

"Andreessen Horowitz Partner Quit, Horrified" — New York Times — June 13, 2026

"I'd Rather Risk Cancer Than See AI Move This Fast" — The Atlantic — June 2026

"Americans Alarmed About AI Bubble" — Futurism — June 26, 2026

"AI CEOs Baffled by Hatred of Their Technology" — Bloomberg — May 18, 2026

"AI Zillionaires Getting Scared as Public Turns" (Mark Cuban) — CNBC — June 29, 2026

"The Mistrust of AI Labs Bubbles Over" (Palantir's Karp) — Semafor — July 9, 2026