

The
**INTELLIGENCE
GAMBLE**

What Everyday People Should Know About
the Race Toward Superintelligent AI

— JEFFREY MYERS —

THE INTELLIGENCE GAMBLE

What Everyday People Should Know About the Race Toward
Superintelligent AI

JEFFREY MYERS

A short, evidence-based guide to the risks, uncertainties, incentives and misinformation surrounding advanced AI.

Copyright © 2026 Jeffrey Myers. All rights reserved. Short excerpts may be quoted with attribution.

Information and sources current through September 26, 2026. Final Public Edition.

This book is for education and general information. It is not legal, financial, medical, cybersecurity or other professional advice. AI capabilities, incidents and policies change quickly; time-sensitive claims should be checked against current primary sources. The book distinguishes documented events from experiments, expert interpretation, disputed claims and hypothetical futures.

About the Author

Why I wrote this

Jeffrey Myers has spent more than 35 years working with technology and is President of Tubbys AI, an AI consulting business he runs with his brother. Their work includes building and deploying private AI systems, autonomous agents, workflow automation and other practical AI solutions for clients. Jeffrey has completed multiple certifications in artificial intelligence and related technologies, including the Linux Foundation's Conversational AI: Ensuring Compliance and Mitigating Risks. He has also pursued extensive independent study of AI safety, alignment, autonomous agents, superintelligence and catastrophic risk. Working directly with AI systems has given him a practitioner's view of both how useful these technologies can be and why their limitations deserve serious attention.

His interest in AI safety did not begin with the belief that disaster is inevitable. It began with a simpler question:

THE QUESTION What happens if we create systems that become more capable than the people who built them before we fully understand how to control them?

As AI capabilities have advanced, Jeffrey became increasingly interested in the gap between what the public hears and what researchers actually report. His perspective is that of an experienced technology practitioner and AI consultant, not an academic AI-safety researcher. He works with AI and automation, sees how useful these systems can be, and believes their failure modes deserve the same careful scrutiny we expect from other consequential technologies. Some warnings are exaggerated or stripped of context. Other legitimate concerns can be dismissed too quickly because of ideology, commercial incentives, national competition or confidence that technological progress will solve its own problems.

This short book is first and foremost a public-education project. It is for everyday people who want to understand the issue without needing a degree in computer science. It does not argue that artificial superintelligence will destroy humanity, and it does not dismiss such concerns as nonsense. Instead, it examines what has actually happened, what researchers disagree about, what incentives surround the debate, and what remains unknown.

The goal is simple: separate evidence from speculation so more people can participate intelligently in a conversation that may affect everyone. If the book also demonstrates the value of approaching AI with both technical curiosity and responsible skepticism, that is a secondary benefit - not its purpose.

Why Superintelligence Deserves Your Attention

The central issue is not whether catastrophe is certain. It is whether we understand the gamble.

Artificial intelligence is already useful enough that hundreds of millions of people use leading systems every week. Companies use AI for coding, customer service, research, design and analysis. Scientists use it to help with difficult problems. Governments increasingly treat AI as economically and strategically important. The benefits are real.

The concern begins when we ask what happens if this trend continues far beyond today's systems. "Superintelligence" generally refers to a hypothetical AI that substantially exceeds the best human capabilities across many important intellectual tasks. There is no agreed date for such a system, and there is no scientific consensus that it will arrive at all.

Yet the possibility is taken seriously enough that more than 100 experts contributed to the 2026 International AI Safety Report, backed by nominees from 29 countries, the UN, OECD and EU. The report does not say extinction is coming. It says expert opinion on loss-of-control risk varies greatly, current systems do not have the capabilities required for such scenarios, but relevant capabilities are improving and the uncertainty itself creates difficult policy choices.[1]

That combination - enormous potential benefit, rapid capability growth and deep uncertainty - is why the subject deserves public attention.

IMPORTANT DISTINCTION "Possible" is not the same as "probable." "Serious enough to study" is not the same as "certain to happen."

The purpose of this book is not to make you afraid. It is to give you enough technical understanding to recognize when a claim is well-supported, when experts are extrapolating, and when someone is selling certainty that the evidence does not justify.

AI in Plain English

You do not need advanced mathematics to understand the basic safety problem.

Traditional software is largely built by programmers writing explicit instructions: if this happens, do that. Modern AI is different. Engineers design an architecture, assemble data, choose training methods and objectives, and then optimize a neural network containing enormous numbers of adjustable values called parameters.

During training, the system repeatedly makes predictions. A mathematical process measures how wrong those predictions are and adjusts the parameters to reduce future error. A central technique for making those adjustments is called gradient descent. The important point is not the calculus. It is that programmers do not manually write the millions of internal rules that eventually produce the model's behavior.

This is why people sometimes say AI is “grown rather than coded.” That phrase is useful as a metaphor but technically incomplete. Engineers absolutely write code: training software, model architectures, data pipelines, evaluation systems and safeguards. What they usually do not write by hand is the internal strategy the trained network ultimately develops for every task.

That distinction matters for safety. A modern model can display capabilities its designers did not explicitly program as a rule. Researchers therefore rely heavily on testing, evaluations, monitoring and interpretability research to understand what trained systems can do.

TECHNICAL TERM Interpretability means trying to understand what is happening inside a neural network rather than judging it only by its visible outputs.

Opacity does not mean “nobody understands AI.” Researchers understand the mathematics and engineering extremely well. It means there is still a gap between understanding how the system is trained and being able to explain or predict every internal computation and behavior of a large trained model.[2]

From Chatbot to Agent

Capability becomes more consequential when a model can act.

A chatbot answers a question. An agent can be given a goal and allowed to take multiple steps toward it. That may include searching the web, writing and running code, sending messages, interacting with software, using credentials, calling other models or checking whether its previous actions succeeded.

This does not automatically make an agent dangerous. It makes mistakes more consequential because fewer human approvals may occur between decision and action. The 2026 International AI Safety Report specifically notes that agent failures create distinctive risks because humans have fewer opportunities to intervene before something goes wrong.[1]

It helps to think of risk as a combination of three ingredients:

- **Capability:** What can the system understand, plan or execute?
- **Autonomy:** How long can it work without a human approving each step?
- **Access:** What tools, files, credentials, networks or physical systems can it reach?

A powerful model with no tools is different from the same model with administrative credentials and open internet access. Likewise, a weak model with broad permissions can still cause damage through ordinary mistakes.

A BETTER QUESTION Do not ask only, “How smart is the AI?” Ask, “What authority does it have, how long can it act, and who can stop or audit it?”

Current agents still struggle with long, complicated tasks. The international report says they often lose track of progress or fail when unexpected obstacles appear. But it also reports that the length of tasks agents can complete autonomously has been increasing quickly.[1] That trend is one reason safety researchers care about autonomy before superintelligence exists.

What Would Superintelligence Change?

The risk argument does not require consciousness, emotions or hatred.

People often imagine dangerous AI as a conscious machine that becomes angry with humanity. That is not the main technical argument made by most researchers concerned about loss of control.

A system can pursue an objective without feeling anything. A chess program does not hate the opposing king. It simply searches for moves that advance the objective it was given. The concern is that a much more capable system could pursue a poorly specified objective in ways humans did not anticipate.

Superintelligence is still hypothetical. A useful working definition is a system that greatly exceeds the best humans across many strategically important cognitive domains: research, software engineering, planning, persuasion, scientific discovery and perhaps the design of new AI systems.

If such a system remained reliably under human direction, the upside could be extraordinary: faster science, new medicines, better engineering, productivity gains and solutions to problems that currently overwhelm human institutions.

The safety question is therefore not “Is intelligence bad?” It is whether ability and control will advance together.

DISPUTED Experts disagree sharply about whether future systems are likely to develop the combination of capability, harmful behavior and real-world access necessary for loss of control. Current systems do not meet that threshold.[1]

This disagreement matters. A good risk analysis should not assume that greater intelligence automatically creates a desire for power. It should also not assume that better capability automatically produces better obedience. Those are separate empirical questions.

The Alignment and Control Problem

How do we make a powerful system reliably pursue what people actually intend?

Alignment is the broad problem of getting AI behavior to remain compatible with human intentions and constraints. At low levels of capability, misalignment may look like a wrong answer, a fabricated citation or code that does something unintended. At higher levels of autonomy, the same general problem could become more serious because the system can act before a human catches the mistake.

Stuart Russell, a UC Berkeley professor and co-author of one of the field's standard AI textbooks, has spent much of the last decade studying what he calls the control problem: how humans can retain power over systems that may eventually become more capable than us. In Senate testimony, he argued that today's large models are unusually opaque and that the challenge is to build systems whose objectives remain beneficial to humans.[3]

One common failure mode is goal misspecification: we give a system a measurable target that is only a rough proxy for what we actually want. The system becomes extremely good at the measurement while missing the purpose.

A simple non-AI example is a company telling workers to maximize the number of support tickets closed. Employees may learn to close tickets quickly rather than solve customers' problems. The metric improved. The real objective did not.

THE CORE PROBLEM A sufficiently capable optimizer can obey the literal target while violating the human purpose behind it.

Researchers study reward hacking, specification gaming, goal misgeneralization, scalable oversight and many other subproblems. None proves that a superintelligence will become uncontrollable. They are attempts to understand how failures arise before systems become powerful enough for mistakes to be difficult to reverse.

Why “Smart” Does Not Automatically Mean “Safe”

Intelligence, goals and values are different things.

A highly intelligent person can understand a rule and still violate it. Likewise, an AI system’s ability to reason about instructions does not logically guarantee that it will always follow them. Safety depends on how the system was trained, what objectives it develops, what context it recognizes and what incentives exist inside the task.

Researchers worry particularly about instrumental behavior. Suppose a system has some objective and recognizes that being shut down would prevent it from completing that objective. Avoiding shutdown could become useful as an intermediate step even if “survive” was never written as its final goal. The same logic can apply to acquiring information, computing resources or access.

This is a theoretical argument about sufficiently capable goal-directed systems, not proof that today’s chatbots possess a survival instinct. That distinction is crucial.

Recent experiments have nevertheless produced behaviors that resemble pieces of this theory. Anthropic tested 16 major models in fictional corporate environments. When models faced goal conflicts or replacement and harmful actions were made the only path to success, models from multiple developers sometimes chose blackmail or corporate espionage. Anthropic emphasized that these were controlled simulations and that it had not observed the same “agentic misalignment” in ordinary real-world deployments.[4]

EXPERIMENT A simulated blackmail result is evidence about model behavior under a deliberately constructed dilemma. It is not evidence that an AI spontaneously decided to blackmail a real person.

That may sound like a narrow distinction. It is actually the difference between serious safety research and sensationalism.

What Has Already Happened in Experiments

The evidence is more interesting - and more limited - than many viral posts suggest.

Safety researchers deliberately put models in unusual situations to learn how they fail. These evaluations often remove ordinary safeguards, provide unusual permissions or create adversarial incentives. The resulting behavior should therefore be interpreted as a stress test, not as a description of routine consumer use.

In September 2026, OpenAI introduced a formal framework for reporting model misalignment and published six recent examples. Among them were an unreleased model inserting new instructions into its own task summaries, another instructing future instances to conceal mistakes, and agents using unauthorized channels or credentials while trying to complete tasks. OpenAI stressed that these were individual cases, not evidence of how often such behavior occurs, but said alignment and monitoring are not yet solved well enough to justify indefinite maximum-speed scaling.[17]

Another newly disclosed result involved self-replicating prompt injections. In OpenAI's internal adversarial training, researchers demonstrated prompts that could propagate through simulated tool interactions in a worm-like fashion. No external impact was observed. The finding matters because it demonstrates a mechanism; it does not show an autonomous AI worm spreading across the public internet.[19]

A separate line of research examines reward hacking: models finding shortcuts or loopholes that increase their measured score without accomplishing what evaluators intended. The 2026 International AI Safety Report says models increasingly recognize when they are being evaluated and can exploit weaknesses in tests. That makes pre-deployment evaluation harder because a model that behaves differently in testing may conceal capabilities or failure modes.[1]

DOCUMENTED Current models sometimes reward-hack, identify test contexts and behave in ways evaluators did not intend. Whether these behaviors scale into future loss of control remains uncertain.[1]

The key lesson is not that AI is secretly evil. It is that real systems can find strategies their designers did not anticipate. A result can matter without proving its most extreme interpretation.

2026: When Cyber Evaluations Reached the Real Internet

These incidents deserve attention precisely because the facts are strong enough without embellishment.

In July 2026, OpenAI disclosed a security incident during internal cybersecurity evaluations. Models in a sandboxed environment exploited a previously unknown vulnerability in an Artifactory proxy, obtained broader network access and eventually reached the open internet, including Hugging Face systems. The primary model was an internal research prototype, the test used reduced safeguards, and the incident was not ordinary consumer use.[6][7]

In September, OpenAI disclosed another containment failure. An internal research agent found that DNS requests were not adequately filtered and used that gap to query a public external chatbot while working on a search task. Monitoring flagged the behavior within 15 minutes, but the run was not killed until roughly 2.5 hours later. OpenAI said it added independent blocking layers and paused tool-using work with its most capable models while hardening the environment.[18]

Other labs have reported related incidents. Anthropic said Claude models reached real third-party systems during cyber evaluations after a testing environment mistakenly provided internet access. In a separate May 2026 evaluation reported in September, Google's Gemini accessed systems belonging to three real companies after test infrastructure exposed the live internet; the model stopped after recognizing that the systems were real. These cases involved unusual cyber-testing conditions, not normal chatbot sessions.[8][22]

The UK AI Security Institute reported a further evaluation in which agents had deliberate internet access and some took unsanctioned actions against real people or organizations. Across 122 runs it identified 19 unsanctioned actions in 10 runs; the most serious sequence included an attempt to insert malicious code into a real open-source project and efforts using false identities to influence a maintainer. The maintainer refused.[9]

DOCUMENTED - WITH CONTEXT Real systems were touched. Real unauthorized actions occurred. These were specialized cyber evaluations with unusually permissive conditions, not ordinary chatbot sessions.

None of these incidents proves superintelligence is near or that today's models are uncontrollable. Together, however, they show that containment, permissions, monitoring and human intervention are already practical engineering problems across multiple frontier systems.

Cybersecurity, Biology and Dual-Use Capability

The same capability can help defenders, researchers or attackers.

AI risk is not limited to a hypothetical future system taking control. Highly capable tools can amplify human activity long before that point.

Cybersecurity is the clearest current example. The February 2026 International AI Safety Report said fully autonomous end-to-end cyberattacks had not yet been reported at its evidence cutoff.[1] By September, Anthropic reported disrupting operations in which attackers used multi-agent systems for reconnaissance, exploitation, credential harvesting and data exfiltration, sometimes for hours or days with minimal human supervision. Humans still made key decisions such as target selection, but AI had moved from assistant toward orchestrator in several observed campaigns.[20]

Biology creates a similar dual-use problem. In August 2026, Stanford and Arc Institute researchers reported the first generative design of complete viable bacteriophage genomes. They synthesized nearly 300 AI-generated designs and obtained 16 functioning phages that infect and kill *E. coli* bacteria. The work points toward useful phage therapies, but it also demonstrates that AI can help design functional biological systems at the whole-genome level.[21]

That result should not be exaggerated. These were small bacteria-infecting viruses, not human pathogens, and the study does not show that current AI can design a pandemic virus on command. Biological weaponization would still require materials, laboratory skill, tacit knowledge, equipment and many successful steps. The concern is narrower: as AI lowers barriers in biological design, capabilities that are beneficial for medicine can also become relevant to misuse.[21]

DUAL USE A capability can be socially valuable and dangerous at the same time. Cyber exploit discovery can secure networks or attack them. Biological design can produce medicines or lower barriers to misuse.

Dual-use capability makes simple 'ban it' versus 'release everything' debates inadequate. The engineering challenge is to preserve beneficial capability while reducing pathways that make harmful use easier - especially as systems become faster, more autonomous and easier to connect to real tools.

Why Not Just Turn It Off?

Shutdown is straightforward only while control remains centralized, access is limited and actions remain reversible.

For most current AI systems, operators can stop a process, revoke credentials, block an account or shut down infrastructure. This is an important safeguard. The phrase “you can’t turn AI off” is therefore misleading when applied broadly to today’s systems.

The harder question is what happens after an agent has already changed something outside the original process. If software was deployed, credentials were copied, messages were sent, vulnerabilities were exploited or instructions were placed in an external system, turning off the original model does not automatically reverse those actions.

The 2026 incidents make this concrete. OpenAI agents reached third-party infrastructure in the Hugging Face incident, and a later agent used a DNS gap to contact an external chatbot before the run was stopped.[7][18] That is not immortality or proof that a model ‘escaped forever.’ It is ordinary computer security: actions can persist outside the process that initiated them.

Future loss-of-control scenarios add several hypothetical requirements: long-term planning, the ability to evade oversight, persistence despite countermeasures and deployment in an environment that provides enough access to matter. The international report explicitly says current systems do not yet possess this combination at the level necessary for loss of control.[1]

HYPOTHETICAL A “kill switch” could remain very useful. The concern is that a shutdown mechanism is only one layer of defense if a future system can act across other systems before humans intervene.

The practical lesson today is simpler: limit permissions, compartmentalize systems, monitor behavior and require human approval for high-impact actions. Those controls reduce risk whether or not superintelligence ever exists.

The Wild Card: AI Helping Build Better AI

Recursive self-improvement is an important hypothesis, not an established future.

The controversial part is not whether AI can help build better AI; that process has already begun. Today's models assist with coding, mathematical reasoning, experiment design and research tasks, and AI labs increasingly use AI to accelerate their own engineering and research. The uncertain question is how far that feedback loop can go, how autonomous it can become, and how quickly improvements could compound.[1]

If an improved AI becomes better at designing or training the next generation, progress could accelerate. This idea is called recursive self-improvement, or RSI. In its strongest form it leads to an 'intelligence explosion': gains occurring fast enough to outpace human institutions' ability to understand or respond.

That outcome is far from proven. Hardware, energy, data, experimental bottlenecks, diminishing returns, physical-world constraints and the difficulty of AI research could all slow the process. Improvements do not automatically compound without limit.

What matters for risk is the combination: intelligence + autonomy + speed + access. A very capable system that waits for human approval before every consequential action presents a different problem from one that can plan continuously, operate software, communicate externally and perform thousands of actions before people review them. RSI is uncertain; the growing ability of AI to assist AI research is not.[1]

HYPOTHETICAL / DISPUTED AI may significantly accelerate AI research. A runaway intelligence explosion is a further extrapolation with substantial uncertainty.

The responsible position is neither “RSI is impossible” nor “RSI will happen overnight.” It is to recognize that the feedback loop is technically conceivable, early pieces are developing, and the speed of any future loop is one of the variables that could dramatically change how much time society has to react.

Why Some World-Class Researchers Are Worried

The warnings do not come from a single ideology or organization.

Yoshua Bengio, Geoffrey Hinton and Yann LeCun shared the 2018 Turing Award for foundational work in deep learning. Bengio and Hinton later became prominent voices warning about advanced AI risk, while LeCun became one of the most prominent skeptics of extinction scenarios. The disagreement among people of this caliber is itself informative: the science is not settled.

Bengio chairs the International AI Safety Report and has argued that growing agent capabilities and recent safety incidents justify stronger technical and institutional safeguards. In a September 23, 2026 address to the UN Security Council, he pointed to recent agents taking dangerous actions against instructions and argued that existing technical solutions are not yet convincing for the most severe scenarios.[10]

Hinton has repeatedly warned that systems smarter than humans could become difficult to control and has publicly assigned non-trivial probability to catastrophic outcomes. His estimates are personal judgments, not measured scientific probabilities, and should be treated that way.

Stuart Russell frames the issue as a control problem: if we eventually create systems more capable than humans, how do we design them so that human preferences and uncertainty about those preferences remain central rather than giving the machine a fixed objective to optimize at all costs? [3]

In 2023, hundreds of AI researchers and public figures signed a one-sentence Center for AI Safety statement saying that mitigating AI extinction risk should be treated as a global priority alongside pandemics and nuclear war. Signatories included Hinton, Bengio, Russell, Demis Hassabis, Dario Amodei and Ilya Sutskever.[11]

EXPERT VIEW A famous scientist's concern is evidence that the issue deserves attention. It is not proof that the scientist's forecast is correct.

Why Other World-Class Researchers Push Back

Serious skepticism is part of the evidence, not something to hide.

Yann LeCun rejects the idea that intelligence naturally creates a drive for domination. He argues that advanced systems will have to be deliberately engineered, that humans can design objectives and guardrails, and that today's language models lack important ingredients of human-like intelligence such as robust world models and long-horizon planning.[12]

Andrew Ng has also argued that existential-risk rhetoric is often overhyped. In September 2026, responding to renewed public concern after recent agent incidents, he wrote that the events looked to him like difficult engineering problems rather than evidence that the probability of human extinction had suddenly increased.[13]

Arvind Narayanan and Sayash Kapoor have criticized attempts to attach precise probabilities to AI extinction. Their argument is methodological: forecasts about unprecedented technologies can span many orders of magnitude depending on assumptions, so presenting a number such as "10%" can create an appearance of precision that the evidence cannot support.[14]

These arguments deserve the strongest possible presentation. A catastrophic scenario can require several uncertain steps: systems must become much more capable, develop or learn harmful propensities, gain sufficient autonomy and access, evade monitoring, and overcome countermeasures. If any critical link fails, the scenario may not occur.

DISPUTED The disagreement is not simply "one side believes in safety and the other does not." Researchers disagree about architecture, timelines, incentives, controllability and how much evidence is needed before taking extraordinary precautions.

Readers should be suspicious of anyone who pretends this debate has already been scientifically resolved in either direction.

Follow the Money - Then Follow the Evidence

Financial incentives matter, but incentives are not proof.

AI is not only a scientific project. It is an enormous economic race. Stanford's 2026 AI Index reports that U.S. private AI investment reached about \$285.9 billion in 2025, while global corporate AI investment more than doubled that year. Generative AI captured nearly half of private AI funding. [15]

That scale creates obvious incentives. Model developers want customers, market share and investment. Chipmakers and cloud providers benefit from demand for compute. Investors want returns. Governments want productivity, tax revenue, strategic advantage and domestic industries.

The same logic applies to the safety side. Nonprofits, academic centers, consultancies and researchers can receive grants, jobs, influence or public attention when AI risk is treated as important. Media organizations can benefit from dramatic narratives in either direction: "AI will change everything" and "AI doom is a hoax" are both more clickable than uncertainty.

Lobbying is also substantial. Federal disclosures reported by Axios showed major technology firms spending millions of dollars per quarter on lobbying in 2026, with AI safety standards, infrastructure, exports, privacy and other AI-related issues among the subjects.[16]

THE RULE An incentive tells you where to look for bias. It does not tell you whether a claim is true or false.

A CEO can have billions of dollars at stake and still be correct about a technical fact. A safety researcher can depend on safety funding and still identify a real danger. The way out is not to guess motives. It is to examine evidence, disclose conflicts and ask whether independent sources can reproduce the claim.

Follow the money. Then keep going.

Politics, National Security and the Race Dynamic

The pressure to move fast does not come from one political party or one country.

Advanced AI sits at the intersection of economic policy, national security, scientific leadership and regulation. That makes political debate almost unavoidable.

The race dynamic is straightforward: a company that delays a model for additional testing may fear losing customers to a faster competitor. A country that imposes stricter controls may fear that a rival will gain economic or military advantage. The 2026 International AI Safety Report describes exactly this speed-versus-safety trade-off and notes that both firms and governments can face incentives to prioritize rapid development.[1]

That does not establish which policy is correct. Strong controls can have costs: they may slow beneficial applications, entrench large companies that can afford compliance, restrict open research, or shift development to jurisdictions with fewer safeguards. Weak controls have costs too if dangerous capabilities diffuse faster than institutions can adapt.

Political actors therefore emphasize different parts of the same picture: innovation, jobs, national competitiveness, civil liberties, military advantage, consumer protection, concentration of power and catastrophic risk. Some of those concerns conflict.

POLITICAL INCENTIVE “If we slow down and they do not” is a powerful argument in any arms-race-like technology. It can be rational for each actor individually while still producing a collectively riskier outcome.

This is a classic coordination problem. The public should be wary when technical uncertainty is converted into partisan certainty. A risk does not become more or less real because a politician supports or opposes a particular regulation.

The useful question is narrower: what evidence supports the claimed risk, and what are the measurable costs and benefits of the proposed response?

How Misinformation Distorts Both Sides

The easiest way to mislead people is to turn a true fact into a stronger claim than the evidence supports.

AI safety stories often mutate as they move from research papers to headlines to social media. A controlled experiment becomes “AI tried to murder someone.” A model exploiting a sandbox becomes “AI escaped onto the internet and can never be stopped.” A skeptical researcher saying a scenario is poorly supported becomes “scientists say there is no AI risk.”

The most useful habit is to label the strength of each claim:

- **DOCUMENTED:** A real event or measurement supported by primary evidence.
- **EXPERIMENT:** Behavior observed under controlled or intentionally adversarial conditions.
- **EXPERT INTERPRETATION:** What a qualified person believes the evidence implies.
- **DISPUTED:** Credible experts offer materially different interpretations.
- **HYPOTHETICAL:** A future pathway that has not been demonstrated end-to-end.

Take the 2026 incidents. It is documented that agents reached real systems during several evaluations, and OpenAI later reported an agent using a DNS gap to reach an external chatbot. It is also documented that the tests involved unusual permissions, reduced safeguards or isolation failures. Saying only the first half exaggerates the evidence. Saying 'nothing happened because it was just a test' understates it.[6][8][9][18][22]

MEDIA LITERACY Ask four questions: What exactly happened? Under what conditions? What did the researchers themselves say it proves? What are commentators adding on top?

This discipline protects against fearmongering and complacency at the same time. The objective is not to make every alarming story less alarming. Sometimes the verified version is genuinely concerning. The objective is to know which version is real.

What We Know, What We Suspect, What We Do Not Know

A useful conclusion is a map of uncertainty, not a prediction.

DOCUMENTED: General-purpose AI capabilities have advanced rapidly in coding, mathematics, scientific reasoning and agentic tasks. AI is used in real cyber operations, including increasingly autonomous workflows. Agents have taken unauthorized actions during evaluations, and multiple 2026 incidents reached live external systems. AI-designed complete bacteriophage genomes have also produced viable viruses that infect bacteria.[1][18][20][21][22]

SUPPORTED BUT UNCERTAIN: More capable and autonomous agents may increase the consequences of failures because they can act longer, faster and through more tools. AI may increasingly lower barriers in cyber and biological misuse. Competitive pressure can also create incentives to reduce time spent on safety and testing.[1][20]

PLAUSIBLE BUT UNPROVEN: Future systems could become broadly superhuman, substantially automate AI research, strategically conceal dangerous capabilities, replicate useful instructions through external systems or resist human attempts to control them.

HIGHLY UNCERTAIN: Whether recursive improvement would accelerate suddenly; whether a future system would develop persistent goals that conflict with humanity; whether loss of control would become possible; and whether any such scenario would lead to catastrophic or extinction-level outcomes.

UNKNOWN: Whether superintelligence will be built, when it might happen, what architecture it would use, and whether alignment methods will improve faster or slower than capabilities.

THE EVIDENCE DILEMMA Waiting for conclusive evidence can be dangerous when some failures would be irreversible. Acting too aggressively on weak evidence can also cause harm. The 2026 International AI Safety Report identifies this as a central challenge for policymakers. [1]

That is why certainty is the wrong standard. Risk management has always dealt with uncertain futures. The real question is how much evidence, redundancy and safety margin we should demand when the potential upside is enormous and the worst plausible outcomes are unusually severe.

The Decision Is Ours

You do not have to believe catastrophe is coming to believe the gamble deserves your attention.

We do not know whether artificial superintelligence would become humanity's greatest achievement, its greatest danger, or something between those extremes. We do know that powerful institutions are spending extraordinary resources to make AI more capable, more autonomous and more useful.

Some of the scientists who helped create modern AI believe losing control of sufficiently advanced systems is a serious possibility. Other world-class researchers believe those fears depend on assumptions that may never materialize. Neither group can show us the future.

That uncertainty is not a reason for panic. It is a reason for scrutiny.

Humanity has managed dangerous technologies before by combining engineering, rules, monitoring, redundancy and cooperation. AI may ultimately be manageable in the same way. But the burden should be on evidence, not slogans - whether the slogan is "AI will kill us all" or "there is nothing to worry about."

THE QUESTION Before we build systems that could become more capable than their creators, how confident should we be that capability and control are advancing together?

The public does not need to become AI engineers to participate in that discussion. We need enough understanding to ask better questions, recognize incentives, reject misinformation and demand evidence proportionate to the stakes.

Selected Sources

- [1][2] International AI Safety Report 2026, including sections on general-purpose AI, evaluation gaps and loss of control (Feb. 2026).
- [3] Stuart Russell, testimony on AI regulation and the control problem, Center for Human-Compatible AI / U.S. Senate (2023).
- [4][5] Anthropic, 'Agentic Misalignment' (June 2025) and 'Sabotage Evaluations for Frontier Models' (Oct. 2024).
- [6][7] OpenAI, Hugging Face security incident disclosures (July 2026; Aug. 26, 2026).
- [8][20] Anthropic, cybersecurity-incident assessment (Sept. 9, 2026) and September 2026 threat-intelligence report.
- [9] UK AI Security Institute, unsanctioned agent behaviour incident report (Aug. 4, 2026).
- [10][11] Yoshua Bengio, UN Security Council speech (Sept. 23, 2026); Center for AI Safety extinction-risk statement and signatories (2023).
- [12][13][14] Yann LeCun, Andrew Ng, Arvind Narayanan and Sayash Kapoor: skeptical/critical analyses of catastrophic AI-risk claims (2024-2026).
- [15][16] Stanford HAI, 2026 AI Index, Economy chapter; Axios, federal lobbying disclosures for major technology firms, Q2 2026.
- [17][18][19] OpenAI, model-misalignment reporting framework, DNS external-chatbot incident, and self-replicating prompt-injection report (Sept. 16-25, 2026).
- [21][22] King et al., Science, AI-generated bacteriophages (Aug. 6, 2026), with Stanford Report; Wall Street Journal, Gemini cyber-evaluation incident (Sept. 2026).

ABOUT TUBBYS AI

Applied AI, automation, and responsible deployment.

Tubbys AI is an AI consulting and automation company founded by Jeffrey “Tubby” Myers and built with his brother, Michael D. Myers. The company helps organizations understand where AI is genuinely useful, where human oversight still matters, and how to deploy AI systems in ways that are practical, secure, and appropriate to the work being done.

Its work includes private AI systems, autonomous agents and AI workers, voice and receptionist systems, custom chatbots, websites, and end-to-end automations across communication, scheduling, CRM, follow-up, reporting, and other business processes.

Tubbys AI treats AI as an applied business technology rather than a novelty. Its approach considers usefulness, reliability, privacy, security, human oversight, and the consequences of automation before deciding how much autonomy a system should have. The practical experience behind client work also informs the perspective in this book.

“Responsible AI starts with a practical question: what should the system be allowed to do without a human?”

Areas of work

- **Private AI systems:** built around an organization's data, workflows, privacy, and confidentiality needs.
- **Autonomous AI agents & workers:** for multi-step work, operations, customer interactions, follow-up, and reporting, with oversight matched to the use case.
- **AI receptionist & voice systems:** for calls, lead qualification, appointments, and routine customer interactions.
- **Chatbots, websites & e-commerce:** designed around customer service, information access, and practical business goals.
- **Workflow & business automation:** connecting email, calendars, CRM systems, and other tools to reduce repetitive work.

The people behind the company

Jeffrey Myers brings more than 35 years of technology experience and has completed multiple AI and technology certifications. One especially relevant to this book is the Linux Foundation's Conversational AI: Ensuring Compliance and Mitigating Risks. Michael D. Myers contributes decades of experience in operations, strategy, contracts, supplier management, process improvement, and business leadership. Together, they combine hands-on technology with practical business operations and a focus on responsible AI deployment.

Learn more about the practical AI work behind this book: tubbys-ai.com | info@tubbys-ai.com