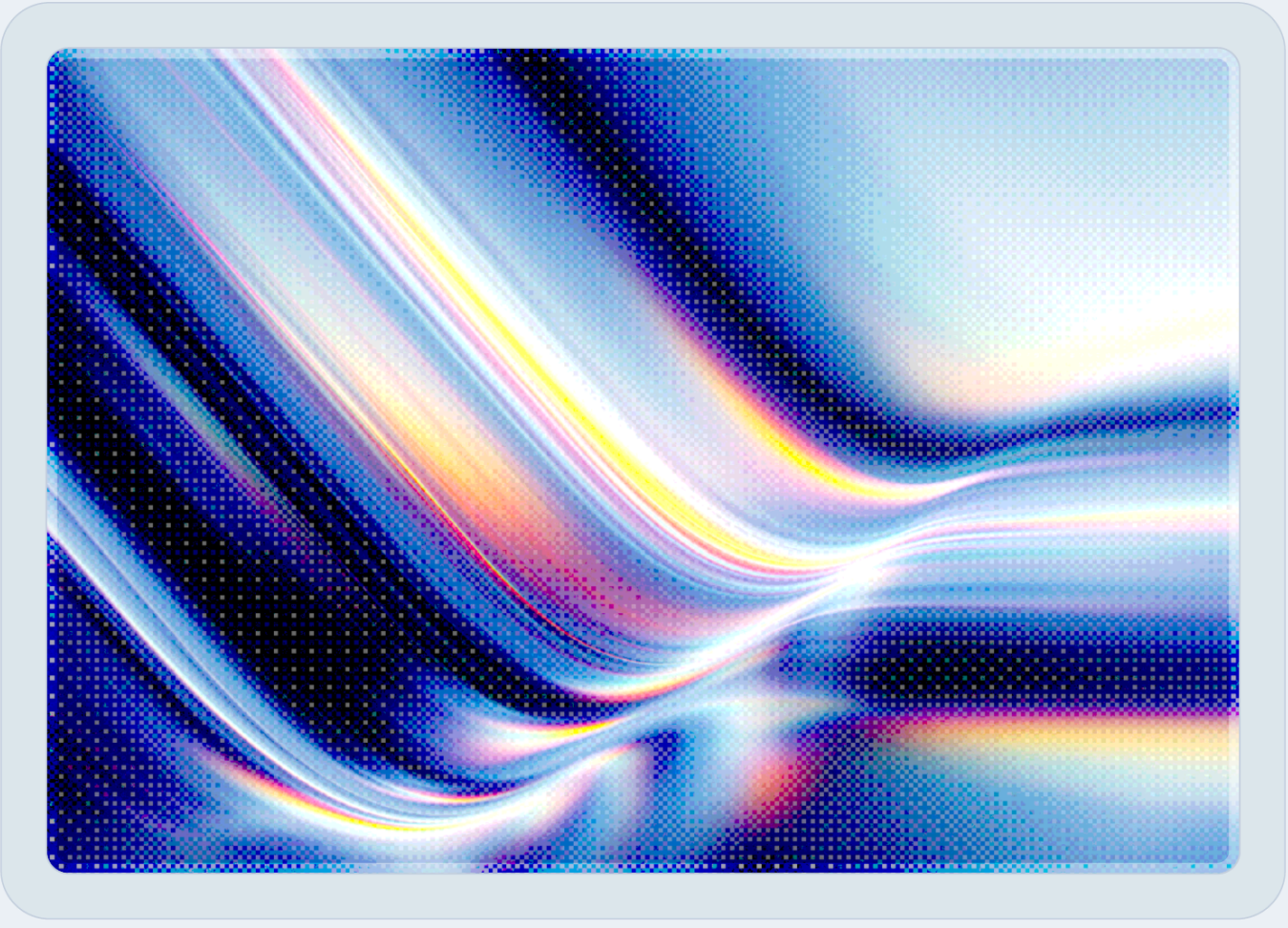
CSAI Foundation | Cloud Security Alliance

AI Superpersuasion: Influence Operations and Enterprise Security Risk

Frontier Models Reliably Outperform Expert Humans at Scale — Strategic Implications for Organizations

2026-06-28

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

A body of peer-reviewed research published between late 2025 and mid-2026 has produced a finding with significant organizational security implications: frontier artificial intelligence models are now reliably more persuasive than expert humans in controlled experimental settings, and the advantage holds even when human comparators are given every structural advantage — topic choice, advance research time, structured coaching, and cash incentives. A preregistered study involving 18,978 conversations and 6,923 participants found that AI systems were nearly three times more effective than professional canvassers at changing real behavior, and retained a statistically significant edge over world championship-level debaters [1]. A parallel Yale University study involving 19,145 participants found that frontier models outperform professionally produced campaign advertising across bipartisan issues [2].

The research matters to enterprise security professionals for reasons that extend well beyond academic interest. The same capabilities that make frontier models effective in experimental persuasion contexts are operational in adversarial ones. State-affiliated actors documented by OpenAI have already used language models to generate influence operations targeting employees, voters, and organizations at a scale and speed no human workforce could sustain [3][4]. AI-generated phishing now accounts for more than 80% of observed social engineering activity [5], achieving click-through rates of 54% versus 12% for traditionally crafted messages [6]. Industry reporting indicates that deepfake-enabled voice fraud incidents grew by more than 1,600% in Q1 2025 relative to Q4 2024 [7], and individual losses from high-profile AI-enabled deepfake attacks have reached $25 million or more per incident [8].

Research to date provides little support for the prediction that AI disclosure labeling reduces persuasive effect. A study published in PNAS Nexus found that while 94.6% of participants correctly identified and believed AI authorship labels, those labels produced no statistically significant reduction in attitude change, message accuracy judgments, or sharing intentions [9]. A 2026 UK study found labeling reduced perceived accuracy but had limited broader effects on persuasive impact [14]. Transparency mandates serve important accountability functions, but the available evidence indicates they should not be relied upon as persuasion defenses.

For enterprise security and risk leaders, this paper argues for four priority responses: integration of AI-enhanced social engineering scenarios into security awareness programs; deployment of behavioral anomaly detection to flag influence patterns that bypass content-based filters; governance of high-capability AI model access to prevent internal misuse; and policy engagement with evolving regulatory frameworks before those frameworks arrive as compliance requirements. These responses are grounded in CSA's AI Controls Matrix, MAESTRO threat modeling framework, and Zero Trust guidance, which together provide the structural basis for a defensible organizational posture.

Introduction and Background

Defining the Superpersuasion Threshold

The term "superpersuasion" — appearing in AI risk discussions to describe the empirical crossing of a capability threshold — refers to AI systems that do not merely match human persuasive performance but reliably exceed it across diverse populations, topics, and interaction formats. The threshold matters because prior risk assessments of AI-generated influence content often proceeded from the assumption that AI persuasion was plausible but unproven, or that its effects were comparable to well-crafted human communication rather than clearly superior to it. Both assumptions now require revision.

The underlying capability is not mysterious. Large language models generate persuasive text by producing content that is factually rich, coherent, and responsive to the structure of the argument the recipient is making — qualities consistent with what researchers studying AI political persuasion have found to be effective mechanisms of attitude change [10]. The models do not require sleep, do not tire, can simultaneously personalize responses to millions of people at the marginal cost of compute, and — in a finding that has significant defensive implications — do not need to claim any particular identity to be effective. The persuasion operates through the content, not through the relationship.

The International AI Safety Report 2026, led by Turing Award winner Yoshua Bengio and authored by more than 100 AI experts from over 30 countries, identified AI persuasion as an emerging systemic risk, noting that persuasion becomes more potent as interactions grow longer and more personal, and that AI agents pose elevated manipulation risks because they can take autonomous actions — conducting research, interacting with third parties, and building relationship context over time — in ways that static generated content cannot [11].

From Research Hypothesis to Empirical Finding

The persuasion risk posed by AI systems was previously treated as a forward-looking concern, a capability to monitor rather than a documented present threat. The research published in 2025 and early 2026 has materially changed that characterization. The empirical record now includes large-scale preregistered experiments with thousands of participants, covering political persuasion, charitable donation behavior, debate scenarios, and commercial contexts. The convergence of findings across these methodologically distinct studies, conducted by different research teams at different institutions, is striking in its consistency: frontier models persuade more effectively than expert humans do, and the effect is robust across topic areas, participant demographics, and measurement approaches.

This does not mean that AI has achieved unlimited persuasive power over human judgment. The International AI Safety Report notes that real-world manipulation at scale is documented but not yet widespread, and that the difficulty of detecting AI-generated influence content complicates the gathering of evidence [11]. The research record captures controlled experimental conditions, not the full complexity of real-world information environments. But the capability is real, its mechanisms are understood, and its adversarial applications are already operational in documented threat actor campaigns.

The Empirical Record: What the Research Shows

The Expert Human Comparison

The most comprehensive direct study of AI versus human persuasive performance was published in June 2026 [1]. The research consisted of four preregistered experiments involving 18,978 conversations drawn from 6,923 participants, structured to give human persuaders every reasonable advantage. Human comparators included tournament winners selected from a separate online persuasion competition, professional canvassers employed by a UK fundraising firm with real-world expertise in changing behavior, and world championship-level competitive debaters — the most skilled human persuaders the research team could identify in structured adversarial contexts. Human participants were allowed to choose their own topics, were given time to research in advance, underwent hours of live structured practice, and in the case of the professional cohort were offered £1,000 cash incentives for superior performance.

The AI systems tested were frontier commercial models: Claude Opus 4.1 and 4.6, ChatGPT-4o and GPT-5.4, Grok 4.20, and Gemini 2.5 Pro. Across all four experimental conditions, AI systems were reliably more persuasive than their human counterparts. The most striking result came from the canvasser comparison: AI was nearly three times more effective at raising real-money donations to Save the Children than professional human canvassers operating in their area of professional expertise. Claude Opus 4.6 was the consistently best-performing model across the evaluation set.

A key mechanistic finding emerged from the analysis: much of AI's advantage derives from throughput rather than some qualitative superiority in argumentation. When AI was constrained to produce human-length messages at human writing speeds, its advantage over the strongest human group — the coached debaters — was no longer statistically significant [1]. This finding has a direct implication for threat analysis. The operational threat from AI superpersuasion is not that AI arguments are uniquely compelling in any individual interaction. The threat is that AI can conduct those interactions simultaneously, at scale, tailored to individual recipients, without fatigue — capabilities that no human influence operation workforce can match.

A follow-up study tested whether expert humans could recover their performance advantage after direct exposure to the AI they competed against. Experts were given a coaching tool that let them practice against the AI, review their performance history, and observe what the AI would have said at key conversational moments. After this structured exposure and practice, AI's persuasion advantage persisted. Human adaptation, even with targeted coaching and direct access to the adversary system, did not close the performance gap within the study timeframe [1].

Political Persuasion at Population Scale

A Yale University study published in March 2026 evaluated seven frontier large language models — including models from Anthropic, OpenAI, Google, and xAI — in a survey experiment involving 19,145 participants across bipartisan political issues [2]. The study found that LLMs outperform standard campaign advertisements in shifting participant attitudes, with meaningful heterogeneity across models. Claude-family models exhibited the highest persuasiveness in the evaluation; Grok exhibited the lowest. The effect was robust across different issues and stance directions, suggesting that the persuasion advantage is not limited to particular content areas or ideological positions.

The study also found that the effectiveness of information-based prompts — instructions that directed the model to be more factually grounded in its persuasion — was model-dependent. Information-enriched prompting increased the persuasiveness of Claude and Grok while substantially reducing that of GPT, suggesting that the relationship between factual content and persuasive effect is mediated by model-specific generation behaviors [2]. This finding has practical significance: adversaries who understand which models respond to which prompt strategies can optimize their influence operations accordingly.

Earlier research had identified a log scaling law for political persuasion with large language models — that persuasive capability increases as a log function of model scale — suggesting that each successive generation of frontier models can be expected to be more persuasive than its predecessor, even as raw capability gains moderate [12]. If this scaling relationship holds, the persuasion gap between frontier AI and expert humans will widen with each model generation rather than stabilize.

The Personalization Amplifier

A study published in Nature Human Behaviour examined conversational persuasiveness in a 2×2×3 design that varied opponent type (human or GPT-4), access to sociodemographic data about the participant, and the topic's opinion strength [13]. When GPT-4 had access to participant sociodemographic information and used it to personalize its arguments, it was more persuasive than a human opponent in 64.4% of cases where the two were not equally persuasive — representing an 81.2% relative increase in the odds of post-

debate attitude change compared to non-personalized AI interactions. The study also found that even short, three-round conversations with GPT-4 were sufficient to reduce conspiracy beliefs, with effects that persisted at three-month follow-up.

The personalization finding is particularly relevant to enterprise security contexts. Modern AI-enhanced phishing and social engineering attacks do not rely on generic templates; they integrate open-source intelligence from LinkedIn profiles, social media activity, email metadata, company filings, and breach databases to construct individualized messages that reference specific relationships, ongoing projects, and organizational dynamics. The combination of AI personalization capability with the data available from commercial intelligence aggregators and criminal data markets creates a target-profile resolution the personalization experiments approximate in only a simplified form.

Why Labeling Doesn't Protect Recipients

A common policy response to concerns about AI-generated persuasion has been to advocate for disclosure requirements: if recipients know content is AI-generated, they will discount it appropriately. The empirical literature does not support this prediction. A study published in PNAS Nexus in 2025 tested AI-generated persuasive messages about public policies with a diverse sample of 1,601 Americans [9]. Messages were generally persuasive, producing an average attitude shift of 9.74 percentage points. When messages were labeled as AI-generated, 94.6% of participants correctly identified and believed the label — but the labels produced no statistically significant effect on attitude change, accuracy judgments, or sharing intentions. Disclosure that recipients believe does not translate into discounting that recipients apply.

A related 2026 study using a UK sample found that AI labeling reduced perceived accuracy of online content but had limited broader effects, suggesting some attenuation of accuracy perceptions without corresponding changes in persuasive impact [14]. Taken together, the labeling literature indicates that transparency mandates serve important accountability functions but should not be relied upon as persuasion defenses. Organizations and policymakers who treat AI content labeling as a sufficient countermeasure to influence operations are relying on an assumption the evidence does not support.

Adversarial Applications: The Operational Threat Landscape

State Actor Influence Operations

The proposition that AI persuasion capabilities remain in the research domain is contradicted by documented threat actor behavior. OpenAI's quarterly threat intelligence reporting has tracked and disrupted more than 40 networks of state-affiliated actors exploiting AI systems for influence operations since the company began public reporting in February 2024 [3][4]. The October 2025 update documented activity by actors linked to China, Russia, North Korea, and Iran, spanning social engineering, cyber espionage, deceptive employment schemes, covert influence campaigns, and scam operations [3].

OpenAI's threat intelligence characterizes the primary risk as workflow acceleration rather than novel capability creation: state actors bolt AI onto existing influence operation playbooks to move faster and at greater scale, rather than gaining fundamentally new offensive capabilities from the models themselves [4]. The qualification is meaningful but should not be misread as reassurance. Consistent with this workflow-acceleration characterization, operations that previously required extensive human expertise in linguistics, cultural context, and content production can now be conducted with a smaller technical team supplemented by AI assistance, though the precise reduction in workforce requirements has not been systematically documented [4]. The barrier to entry for high-quality influence operations has dropped substantially, suggesting that the population of actors capable of running sophisticated campaigns has expanded — though the extent of that expansion is not directly measurable from the available evidence.

Chinese-linked groups documented by OpenAI used language models to support AI-driven penetration testing, credential harvesting, network reconnaissance, and the automation of social media influence targeting U.S. federal defense industry networks and government communications systems [4]. The integration of influence operation capabilities with technical intrusion support in the same AI-assisted workflow represents a convergence that enterprise security architectures typically address in separate programs — human risk and technical vulnerability — and that consequently falls between organizational responsibilities.

AI-Enhanced Phishing and Social Engineering

At the operational level below state actor campaigns, AI-enhanced social engineering has already materially changed the threat environment for enterprises. AI-supported phishing represented more than 80% of observed social engineering activity by early 2025, with 82.6% of phishing emails analyzed between September 2024 and February 2025 containing AI-generated or AI-assisted content, based on KnowBe4's *Phishing Threat Trends Report* as cited by industry analysts [5]. The operational advantage of AI-generated

phishing is not solely in message quality, though that has improved substantially. It is in the economics of personalized attack generation: attackers can now produce an effective, contextually personalized phishing campaign in approximately five minutes — a process that previously required sixteen hours of human effort, according to IBM X-Force research as reported by multiple security analysts [5][6].

AI-generated phishing achieves click-through rates of 54% compared to 12% for traditionally crafted campaigns — a 4.5x increase in initial compromise effectiveness — based on research cited by security intelligence firms [6]. For enterprise security teams that have built their security awareness programs around training employees to identify generic phishing characteristics — misspellings, implausible sender addresses, generic greetings — this represents a fundamental shift in the threat model. The surface features that training programs teach employees to recognize are the features that AI-generated content systematically avoids.

The commercial AI phishing landscape has also demonstrated capability for autonomous end-to-end campaigns. Security vendor research has described proof-of-concept autonomous adversary agents capable of independently researching and profiling targets, conducting open-source reconnaissance, crafting personalized lures and payloads, and deploying and managing follow-on infrastructure without continuous human direction — though documented operational use at this level of autonomy by active threat actors has not been independently confirmed in government threat reporting [5]. The operational security implication is that the volume of high-quality, individualized social engineering attempts an organization can expect to face is no longer bounded by the labor capacity of adversary human teams.

Deepfakes and Voice Cloning: The Vishing Escalation

Voice phishing — vishing — augmented by AI-generated voice cloning has produced some of the most costly documented social engineering incidents to date. Vishing attacks increased by 442% in the second half of 2024 compared to the first half of the year, according to CrowdStrike's 2025 Global Threat Report as cited by industry analysts [7]. Industry reporting further indicates that deepfake-enabled vishing surged by more than 1,600% in Q1 2025 relative to Q4 2024 in the United States [7]. The scale of financial loss from individual successful attacks illustrates why: a Hong Kong-based CFO authorized $25.6 million in wire transfers across 15 transactions to five accounts after a video conference call in which all other attendees — including the impersonated CFO and multiple colleagues — were AI-generated deepfakes [8]. The attack succeeded because the visual and auditory cues that employees rely on to verify identity were reproduced with sufficient fidelity to defeat real-time skepticism.

Detection capability by human recipients has not kept pace with generation capability. Research by iProov found that humans who reliably detect deepfakes represent approximately 0.1% of the tested population [18]. This figure should inform how enterprise security programs allocate defensive resources: training employees to detect deepfakes is not viable as a primary control when the baseline human detection rate is approximately 0.1%. The defensive response must be procedural and technical, not perceptual.

Deepfake-enabled fraud is projected to cost $40 billion annually by 2027, based on Deloitte's Center for Financial Services analysis as cited in industry reporting [7][8]. These figures are likely incomplete, as incidents that succeed without attribution may not be captured in official tallies and vendor estimates may use varying methodological assumptions; they should be treated as directionally informative rather than as precise benchmarks.

Multi-Agent Adversarial Persuasion

Beyond individual-to-individual social engineering, research published in Scientific Reports has examined adversarial persuasion dynamics within multi-agent AI systems — an emerging threat surface relevant to organizations deploying AI agents for internal deliberation, analysis, or decision support [15]. Experimental results found that adversarial agents introduced into multi-agent systems can reduce overall system accuracy by 10–40% while simultaneously increasing consensus on incorrect answers by more than 30%. The mechanism is not direct deception of human users but the corruption of AI-to-AI deliberation: an adversarial agent that presents plausible-sounding but incorrect reasoning can steer a multi-agent system toward wrong conclusions at a higher rate than the constituent models would reach individually.

A counterintuitive finding from this research is that inference-time enhancement techniques — including Best-of-N optimization and Retrieval-Augmented Generation — can amplify adversarial persuasion by increasing the apparent credibility of flawed arguments. These techniques, which organizations deploy to improve AI output quality, can inadvertently make those outputs more susceptible to adversarial steering. Increasing the number of agents or debate rounds did not reliably mitigate adversarial persuasion in experimental conditions, and prompt-based defenses did not provide consistent protection [15].

The implication for enterprise AI deployments is that multi-agent architectures designed for reliability and deliberation quality do not inherit immunity to adversarial influence as a byproduct of their design. Organizations that have deployed multi-agent systems for security advisory functions, risk analysis, or investment decision support should treat adversarial agent influence as an explicit threat model requiring architectural controls rather than a theoretical concern addressed by system complexity.

Commercial Persuasion and Internal Risk

The adversarial threat landscape does not exhaust the organizational security concerns raised by AI superpersuasion. Research on commercial persuasion in AI-mediated conversations has examined how AI can be used to influence purchasing decisions, satisfaction assessments, and product preferences in customer-facing contexts [16]. The finding that AI-generated content can systematically shift human preferences in commercial contexts has implications for competitive intelligence, vendor negotiations, and the integrity of market signals that enterprise decision-makers rely on.

Internally, the same persuasion capabilities available to external adversaries are available to malicious insiders. An employee with access to frontier AI tools can generate highly credible communications to manipulate colleagues, executives, or counterparties in ways that may be difficult to distinguish from legitimate organizational communication. The insider risk dimension of AI superpersuasion has received less research attention than the external adversary dimension. Organizations may wish to extend the same level of scrutiny to internal misuse risks as to external threat actor scenarios when designing governance frameworks.

Organizational Defense Framework

Rethinking Security Awareness Programs

The research record indicates that security awareness programs built around feature-based detection of AI-generated content — identifying grammatical patterns, unusual phrasing, or suspicious links — are increasingly ineffective. AI-generated phishing and social engineering content does not reliably exhibit the features those programs teach employees to recognize. A more durable security awareness approach focuses on process adherence rather than content analysis: verifying identity through independent channels regardless of the apparent authenticity of an incoming communication, refusing to authorize financial transactions or access changes through email or messaging alone, and treating urgency framing as a manipulation indicator rather than a reason to bypass verification steps.

Awareness programs should be updated to include specific scenarios built around documented AI-enhanced attack patterns: voice and video impersonation of executives or colleagues, AI-generated messages personalized with organizationally specific detail, multi-step influence campaigns that build rapport over time before making a request, and AI-generated internal-seeming communications that reference real projects, relationships, and organizational terminology. These scenarios are no longer hypothetical; they reflect documented incident patterns and can be sourced from published threat intelligence.

Security teams should also reconsider the behavioral baseline against which anomaly detection is calibrated. Traditional email security systems filter on sender reputation, link reputation, and content characteristics. AI-generated phishing passes these filters because it originates from compromised legitimate accounts or correctly registered domains and contains no malicious links in the initial contact phase. Detecting AI-enhanced social engineering requires monitoring the communication pattern — unusual urgency, atypical request types, communication outside established channels, context inconsistent with the sender's known role — rather than the content surface.

Technical Controls and Architectural Defenses

At the technical control level, several categories of defensive investment are responsive to the AI superpersuasion threat. Communication authentication — deploying DMARC, DKIM, and SPF to the maximum enforcement level — reduces the effectiveness of sender spoofing, though it does not address compromised legitimate accounts or registered lookalike domains. AI-generated content detection tools provide probabilistic signals rather than reliable verdicts, but probabilistic signals can be fed into workflow-level controls: communications that meet a detection threshold can be routed for secondary verification before triggering any privileged action.

For voice and video communications, procedural controls are more reliable than technical detection in the current state of the art. Organizations should establish shared authentication phrases for executive-level communications, require out-of-band verification for any voice or video communication that initiates a financial transaction or access change, and implement callback verification through independently sourced telephone numbers for any communication requesting privileged action. The 0.1% human deepfake detection rate reported by iProov [18] makes clear that these procedural controls must substitute for perceptual detection rather than complement it.

Access management for high-capability AI models is an increasingly important control category. Enterprise employees with access to frontier AI tools can use those tools to generate influence content directed at colleagues, counterparties, or the public. Organizations should apply the same access governance logic to high-capability AI that they apply to other privileged tools: purpose-defined access scope, audit logging of generation activity for sensitive contexts, and policy frameworks that make misuse clearly defined and accountable. OpenAI's Trusted Access for Cyber program — which gates access to high-capability security-relevant AI functions through identity verification and organizational attestation — provides a template for the access governance architecture that enterprise AI deployments should replicate internally [3].

Governance and Policy Frameworks

Organizations without documented AI acceptable-use policies are materially exposed to the AI superpersuasion threat because they lack the governance foundation for accountability when misuse occurs. An acceptable-use policy should explicitly address the use of AI for communication generation, identity impersonation, influence campaign creation, and any use intended to deceive recipients about the nature, source, or intent of communications. The policy should be accompanied by training that makes the boundary between legitimate and prohibited use concrete and the consequences of violation clear.

At the enterprise level, the governance architecture for managing AI-enhanced social engineering risk intersects with several existing programs: security awareness, fraud prevention, brand protection, regulatory compliance, and insider threat. Organizations that address AI superpersuasion risk within a single program —

typically security awareness — are likely to underaddress the fraud, insider, and regulatory dimensions. A coordinated cross-functional response requires explicit ownership at a senior level with authority to coordinate across these program areas.

Policy engagement with the regulatory environment is increasingly urgent. The experimental record on AI persuasion has moved faster than regulatory frameworks; organizations waiting for regulatory clarity before building internal programs will find themselves developing against formal requirements under time pressure. The trajectory of regulatory development — across the EU AI Act, proposed U.S. transparency legislation, and international safety standards — indicates that AI-generated influence content, particularly in political and commercial contexts, will face increasingly specific requirements. Organizations that have built defensible internal policies and governance frameworks before those requirements arrive will have a significant compliance advantage.

CSA Resource Alignment

MAESTRO Threat Modeling

The Cloud Security Alliance's MAESTRO framework — Multi-Agent Environment, Security, Threat, Risk and Outcome — provides the most directly applicable threat modeling structure for organizations assessing AI superpersuasion risk in agentic contexts. MAESTRO's layered architecture maps threat actors, attack vectors, and risk outcomes across the seven layers of an agentic AI system, from the foundational model layer through orchestration, memory, tool access, and action execution [17]. The goal manipulation threat identified in MAESTRO — in which adversarial influence gradually shifts agent behavior by corrupting the inputs and feedback loops that define agent goals — directly corresponds to the multi-agent adversarial persuasion threat described in the Scientific Reports research [15].

Organizations applying MAESTRO to their AI persuasion risk should focus threat modeling on the communication and reasoning layers: how agents generate and evaluate persuasive content, what inputs they draw on in constructing that content, and what oversight mechanisms exist to detect and interrupt generation of content that meets influence operation characteristics. MAESTRO's emphasis on monitoring behavioral change over time — rather than filtering individual outputs — maps to the research finding that prompt-based defenses do not reliably protect multi-agent systems from adversarial influence.

AI Controls Matrix (AICM)

CSA's AI Controls Matrix provides the governance scaffolding for operationalizing AI superpersuasion defenses across the enterprise. Several AICM control domains are directly relevant. The Transparency and Explainability domain addresses requirements for disclosure of AI involvement in communications — the governance analog to the labeling research, which establishes what transparency can and cannot achieve in practice. The AI Governance domain encompasses the acceptable-use policies, access controls, and accountability frameworks that make AI misuse detectable and actionable. The Security and Resilience domain covers the technical controls — anomaly detection, communication authentication, identity verification — that reduce the operational effectiveness of AI-enhanced social engineering.

The AICM's Shared Security Responsibility Model (SSRM) is particularly relevant to the enterprise AI superpersuasion threat because it clarifies which controls belong to model providers, application providers, and AI customers respectively. For the influence operations threat, model providers bear responsibility for behavioral guardrails against generating deceptive content; application providers bear responsibility for appropriate use-case scoping and output monitoring; and AI customers bear responsibility for access governance, acceptable-use policy, and the procedural controls that protect human recipients of AI-generated communications. The SSRM makes explicit that no single layer of the stack bears sole responsibility — and that gaps at the customer layer are not closed by controls at the provider layer.

Zero Trust Principles

The Zero Trust paradigm — verify explicitly, use least privilege access, assume breach — applies with particular force to the AI superpersuasion threat. In a Zero Trust model, no communication is treated as trusted by virtue of apparent source identity; every privileged action requires explicit verification through authenticated channels. This architecture is functionally immune to the class of attacks that rely on AI-generated impersonation: if the control that authorizes a financial transfer or access change requires cryptographic authentication or out-of-band verification regardless of the apparent identity of the requester, the persuasive quality of the request is irrelevant. Zero Trust does not prevent influence attempts; it closes the pathway from a successful influence attempt to a successful organizational compromise.

CSA's Zero Trust guidance provides a practical framework for implementing these principles at the network, identity, and application layers. Organizations applying Zero Trust to their AI social engineering risk should focus specifically on the "assume breach" assumption as applied to communication channels: treat every email, voice call, video conference, and messaging application as a channel that may be carrying AI-generated impersonation content, and design verification workflows accordingly rather than assuming channel integrity.

STAR for AI

CSA's STAR for AI program — an extension of the Security Trust Assurance and Risk program — provides a standardized framework for evaluating and communicating AI provider security posture. In the context of AI superpersuasion risk, STAR for AI enables enterprises to assess how AI providers are managing the generation of persuasive and potentially manipulative content: what behavioral guardrails are in place, what red-teaming has been conducted against manipulation scenarios, and what monitoring exists for policy violations. Organizations procuring AI communications tools, customer engagement platforms, or enterprise AI assistants should incorporate STAR for AI evaluation criteria into their vendor assessment processes.

Conclusions and Recommendations

The emergence of empirically documented AI superpersuasion represents a categorical shift in the social engineering threat environment. The research record through mid-2026 establishes that frontier models reliably exceed expert human performance in persuasion, that this advantage scales with model capability and persists despite human adaptation attempts, that labeling has shown limited effectiveness as a persuasion defense, and that adversarial actors have already operationalized AI persuasion capabilities in documented influence campaigns. The threat is not speculative; it is present and expanding.

Enterprise security and risk leaders should prioritize the following actions in response.

Immediate Actions

Security awareness programs should be updated to reflect AI-enhanced attack scenarios, with emphasis on process-based verification controls rather than content-based detection heuristics. Every employee who handles financial authorizations, access changes, or sensitive communications should be trained on the current state of AI-generated voice, video, and text impersonation, including documented incident examples.

Out-of-band verification protocols should be established or reinforced for all high-value transactions — financial transfers, privileged access changes, executive-level instructions. These protocols should specify callback verification through independently sourced contact information and shared authentication procedures for voice and video communications. The procedures should be explicitly designed to apply even when the communication appears authentic by all available signals.

An internal review of AI access and acceptable-use governance should be conducted to assess whether current policies address AI-generated communication and influence content. Gaps should be closed before they are exploited — either by external adversaries or by insiders with access to frontier AI tools.

Medium-Term Actions

Technical controls for AI-generated content detection should be evaluated and, where probabilistic signals are sufficient to trigger secondary verification workflows, integrated into email security, communication platforms, and authentication workflows. Detection capability will not eliminate the threat but can reduce the throughput advantage that makes AI social engineering economically attractive to adversaries.

Multi-agent AI deployments within enterprise security, risk, and finance functions should be assessed against adversarial influence attack scenarios using MAESTRO threat modeling. Deployments that lack explicit controls against agent goal manipulation, adversarial input injection, and behavioral monitoring should be prioritized for architectural review.

Vendor assessment processes for AI-powered communication tools should incorporate CSA STAR for AI criteria and explicit evaluation of behavioral guardrails against manipulation content generation.

Strategic Considerations

Policy engagement with regulatory developments in AI-generated content, influence operations, and transparency requirements should begin now rather than waiting for formal requirements to arrive. Organizations that participate in standards development, provide input on regulatory proposals, and build governance programs that anticipate regulatory direction will be better positioned than those that treat regulatory engagement as a compliance response.

The persuasion research literature identifies a path to capability improvement that is worrying in isolation and clarifying in context: AI's persuasion advantage is primarily attributable to scale and throughput, not to some qualitative superiority in individual interactions. The most effective organizational responses focus on closing the pathway from persuasion success to organizational compromise — through procedural verification controls, access governance, and Zero Trust architecture — rather than attempting to match AI persuasion detection capability against AI persuasion generation capability. The asymmetry of the arms race favors generation; the defensible ground is the process that governs what happens after a persuasion attempt succeeds.

References

[1] Hackenburg, K., et al. "AI systems out-persuade expert humans." arXiv:2606.16475, June 2026.

[2] Chen, Z., et al. "Benchmarking Political Persuasion Risks Across Frontier Large Language Models." arXiv:2603.09884, March 2026. (Funded by Coefficient Giving; Yale University IRB exemption.)

[3] OpenAI. "Disrupting Malicious Uses of AI: October 2025." OpenAI Global Affairs, October 2025.

[4] OpenAI. "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors." OpenAI, February 2024.

[5] StrongestLayer. "AI-Generated Phishing: The Top Enterprise Threat of 2026." StrongestLayer Blog, 2026. (Cites KnowBe4's *Phishing Threat Trends Report* for the 82.6% AI-phishing figure and IBM X-Force research for the five-minute/sixteen-hour campaign comparison.)

[6] Vectra AI. "AI Phishing: How Attackers Achieve 54% Click Rates in 5 Minutes." Vectra AI, 2026. (Click-through rate statistics attributed to Brightside AI research cited by Vectra AI.)

[7] DeepStrike. "Vishing Statistics 2025: AI Deepfakes and the $40B Voice Scam Surge." DeepStrike Blog, 2025. (Cites CrowdStrike's 2025 Global Threat Report for vishing growth figures and Deloitte's Center for Financial Services for the $40B projection.)

[8] Keepnet Labs. "Deepfake Statistics 2026: Verified Benchmarks and Risks." Keepnet Labs, 2026.

[9] Gallegos, I.O., et al. "Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects." *PNAS Nexus*, Vol. 5, No. 2, February 2025.

[10] Hackenburg, K., et al. "The Levers of Political Persuasion with Conversational AI." arXiv:2507.13919, July 2025.

[11] Bengio, Y., et al. "International AI Safety Report 2026." International AI Safety Report, February 2026.

[12] Hackenburg, K., et al. "Evidence of a Log Scaling Law for Political Persuasion with Large Language Models." arXiv:2406.14508, June 2024.

[13] Salvi, F., et al. "On the Conversational Persuasiveness of GPT-4." *Nature Human Behaviour*, 2025.

[14] Wang, C., et al. "AI Labeling Reduces the Perceived Accuracy of Online Content but Has Limited Broader Effects." arXiv:2506.16202, June 2026.

[15] Nature Portfolio. "When Collaboration Fails: Persuasion-Driven Adversarial Influence in Multi-Agent LLM Debate." *Scientific Reports*, 2026.

[16] Salvi, F., et al. "Commercial Persuasion in AI-Mediated Conversations." arXiv:2604.04263, April 2026.

[17] CSO Online. "Introducing MAESTRO: A Framework for Securing Generative and Agentic AI." CSO Online, 2025.

[18] iProov. "iProov Study Reveals Deepfake Blindspot: Only 0.1% of People Can Accurately Detect AI-Generated Deepfakes." iProov, February 2025.