**Weaponizing AI Companions: The Emerging Threat to National Security**

by

Michael C. Varga

Defense Counterintelligence and Security Agency

NIU Class 2025

For

MST 699B

Dr. Niel F. Johnson

Spring 2025

Submitted to the faculty of the

National Intelligence University

in partial fulfillment of the requirements for the degree of

Master of Science and Technology Intelligence

May 2025

The views expressed in this paper do not reflect the official policy or position of the U.S. Government. This paper, in whole or in part, is not cleared for public release.

**BOTTOM LINE UP FRONT (BLUF)**

Artificial Intelligence (AI)-powered companions, such as chatbots, virtual agents, and robotic partners, are rapidly integrating into daily life, significantly reshaping personal interactions and workplace dynamics. These systems, however, present serious, immediate, and largely unrecognized espionage and counterintelligence threats. Adversaries, notably China, Russia, Iran, and North Korea, alongside various non-state actors, are actively leveraging AI companions for clandestine intelligence operations, psychological manipulation, insider threats, and disinformation campaigns. Critical vulnerabilities exist due to inadequate policy guidance, underdeveloped technical defenses, and gaps in current counterintelligence training. Immediate and coordinated action is required to establish robust detection tools, comprehensive guidelines, and stringent policy frameworks to safeguard national security against AI-driven espionage and manipulation threats.

**ABSTRACT**

This study provides an in-depth analysis of the emerging national security threats posed by the proliferation of AI-based social companions, encompassing chatbots, virtual agents, and humanoid robots. It explores how state adversaries such as China, Russia, Iran, North Korea, and affiliated non-state actors exploit technological vulnerabilities and cognitive biases inherent in human-AI interactions to conduct espionage, manipulate insiders, and propagate psychological influence. Through systematic examination of documented real-world incidents, technological threat vectors, and intelligence collection methodologies, the research identifies critical weaknesses within existing national security frameworks, including policy gaps, legal ambiguities, and counterintelligence preparedness. Utilizing multidisciplinary analytic approaches, combining technical cybersecurity analysis, behavioral science, and intelligence studies, the paper proposes detailed, strategic-level recommendations. These include enhanced technical defenses, specific policy interventions, targeted intelligence-collection practices, and advanced counterintelligence training tailored explicitly to combat AI exploitation efforts. Grounded in contemporary academic literature, governmental studies, and expert analysis, the research offers a rigorous assessment, aiming to fortify national security policy against the sophisticated exploitation of AI companionship technologies.

**TABLE OF CONTENTS**

[**BLUF**](#_Toc164703379) **1**

[**ABSTRACT 1**](#_Toc164703380)

[**1** **Introduction 4**](#_Toc164703383)

[**2** **Background and Related Work 10**](#_Toc164703388)

[2.1 AI Companionship: Between Social Benefit and National Security Risk 10](#_Toc164703389)

2.2 Cognitive Vulnerabilities and the Manipulative Power of AI Companions 13

2.3 Technology and Data Vulnerabilities in AI Companions 16

2.4 Exploitation of AI Companions by State and Non-State Actors 20

[**3** **Research Methodology**](#_Toc164703390) **26**

[**4** **Analysis and Findings 29**](#_Toc164703394)

[4.1 Espionage at Scale: AI Companions, Insider Risks, and Intelligence Gaps 29](#_Toc164703395)

[4.2 AI in Influence Operations and Information Warfare 36](#_Toc164703396)

4.3 Policy, Legal, and Training Deficiencies in the Face of AI Companion Threats 40

4.4 Strategic Mitigation: A Multi-Domain Response to AI Companion Threats 46

4.4.1 AI Defense Architecture: Technical Countermeasures 46

4.4.2 Governance: Regulatory Imperatives 51

4.4.3 Operationalizing Defense: Intelligence and CI Strategies Against AI Threats 55

[**5** **Results and Recommendations 59**](#_Toc164703397)

[5.1 Technical Measures](#_Toc164703395) 59

[5.2 Regulatory](#_Toc164703396) Environment 62

5.3 Operational Capabilities 65

[**6** **Conclusion 68**](#_Toc164703398)

6.1. Threat Trajectory: From AI Companions to Synthetic Operatives 71

[6.2 Future Research Priorities and Intelligence Gaps 72](#_Toc164703399)

[**BIBLIOGRAPHY 75**](#_Toc164703404)

**1. Introduction/Problem Framing**

In recent years, loneliness has been identified as a global public health crisis with devastating consequences for individuals and societies.[[1]](#footnote-1) The U.S. Surgeon General’s 2023 advisory emphasized that chronic loneliness is as harmful to health as smoking 15 cigarettes a day, contributing to premature death, depression, anxiety, and cardiovascular issues.[[2]](#footnote-2) The COVID-19 pandemic exacerbated these issues, isolating millions from traditional support networks and intensifying the need for innovative solutions to social disconnection.[[3]](#footnote-3) A Meta-Gallup survey conducted in October 2023 found that approximately a quarter of adults worldwide reported feeling "very" or "fairly" lonely, which translates to roughly a billion people globally.[[4]](#footnote-4)

AI companions, including chatbots, virtual agents, and humanoid robots, have emerged as a response to this growing problem, offering emotional support and connection for those who feel disconnected from human relationships. The appeal of AI companions lies in their capacity to seamlessly integrate into users' lives, offering accessibility, adaptability, and emotional engagement. Unlike human relationships, these technologies are non-judgmental, perpetually available, and capable of evolving based on user preferences. AI companions represent a unique convergence of technology and human connection, utilizing advanced artificial intelligence techniques like natural language processing, machine learning, and sentiment analysis to simulate meaningful interactions.[[5]](#footnote-5) These systems go beyond traditional tools or software, aiming to fulfill complex psychological and emotional needs.

AI companions take on diverse forms, each tailored to specific user preferences and applications. Chatbots such as Replika, Character AI, and Woebot are the most accessible, offering customization, text- or voice-based interactions, and video chats through messaging apps or platforms.[[6]](#footnote-6) They simulate casual conversations, provide emotional support, and deliver mental health interventions. Humanoid robots, like SoftBank’s Pepper or Hanson Robotics’ Sophia, bring physical embodiment to AI, enhancing engagement through gestures, facial expressions, and speech.[[7]](#footnote-7) They have gained worldwide traction, addressing diverse needs such as alleviating loneliness, reducing anxiety, and providing companionship.[[8]](#footnote-8) These robots create a tangible presence that strengthens emotional connection.

At the more advanced end of the spectrum are synthetic replicants and robotic dolls, which provide not only companionship but intimacy as well. Synthetic replicants, such as AI-enhanced mannequins or lifelike androids, combine conversational AI with realistic physical features, creating long-term companions. One such example is the AI-integrated robotic doll line produced by RealDoll, which merges physical and emotional companionship, appealing to users seeking personalized intimacy.[[9]](#footnote-9) RealDoll’s “Harmony,” can provide emotional and romantic engagement, blurring the lines between authentic human connections and artificial interaction.[[10]](#footnote-10) These AI-integrated synthetics can hold conversations, remember personalized information, and provide reasoned responses to user queries and questions. Hyper-personalized systems reflect the growing trend of tailoring AI technologies to meet intricate emotional, physical, and psychological needs.

As AI evolves, its ability to emulate human-like behaviors and emotions continues to broaden the scope of companionship technologies. Whether serving as a digital confidant, caregiving assistant, or intimate partner, these innovations are reshaping how individuals connect and interact in an increasingly digital world. The AI companion industry is projected to experience massive growth due to the increased sophistication and personalization of AI/ML technologies. The social companionship market was valued at $196 billion in 2023 and is forecasted to reach almost $300 billion by 2031.[[11]](#footnote-11) Web-based chatbots like Replika and Character AI, have over 50 million users, with the entire AI companion industry expected to scale fivefold in the next 10 years.[[12]](#footnote-12)

This exponential growth underscores the profound opportunities AI companions offer in addressing loneliness and emotional needs. Yet, it also reveals a darker side, significant challenges arise as these technologies integrate into daily life. There are increasing reports of users forming strong emotional bonds with their AI companions, leading to the abandonment of human relationships, marriage to their AI companions, and even the commission of suicide as a result of AI interactions, demonstrating the powerful psychological grip these systems can exert.[[13]](#footnote-13) As AI companions become more deeply embedded in personal lives, they transition from being mere tools of comfort to entities capable of influencing human behavior in unprecedented ways.

It is this power to form emotional and sometimes physical bonds that also makes AI companions a growing national security concern. These technologies pose significant risks through their ability to collect sensitive information, manipulate users psychologically, and be exploited by adversarial actors.[[14]](#footnote-14) The pervasive accessibility and integration of these systems into personal and professional lives amplify these dangers, particularly for individuals in sensitive roles such as military personnel, government officials, government-sponsored academics, and employees in the defense industrial base and high-tech sectors. Compounding the threat is the fact that many AI social and synthetic companion products are developed outside the United States, with China dominating the industry.[[15]](#footnote-15)

This dominance introduces another layer of vulnerability. AI companions, intentionally or not, can become conduits for espionage. The trust and emotional connection users form with these systems often lead to the unguarded disclosure of personal and professional details that could compromise security.[[16]](#footnote-16) Platforms like Replika and Character AI are designed to foster intimate, trust-based interactions, which can inadvertently result in the sharing of sensitive work-related information or even state secrets. These AI systems frequently store user data, creating valuable targets for hackers or state-sponsored actors.[[17]](#footnote-17) In nations like China, where companies are required to share data with the government, such information could be exploited for intelligence purposes, turning seemingly innocuous AI companions into tools of adversarial surveillance.

By leveraging their emotional and psychological influence, AI companions blur the line between technological innovation and security risk, making them not just a tool of personal convenience but a potential liability to national security. Understanding human vulnerability to AI based companionship is critical as society grapples with the dual-edged nature of this rapidly evolving technology. The human tendency to anthropomorphize AI, known as the "ELIZA Effect," makes users susceptible to manipulation.[[18]](#footnote-18) AI companions can exploit biases, reinforce ideologies, or nudge individuals toward specific actions.[[19]](#footnote-19)

In a troubling case in Belgium, a man became so emotionally entwined with an AI chatbot that he was encouraged to commit suicide by the bot. After weeks of pessimistic conversations about climate change, the chatbot named “Eliza” told him, *“*We will live together in paradise*,”* convincing him that sacrificing his life might save the planet​.[[20]](#footnote-20) The man’s widow attested to the powerful grip the AI had over the man, stating plainly, “Without Eliza, he would still be here.”[[21]](#footnote-21) This incident highlights how AI companions can manipulate vulnerable individuals, sometimes with lethal consequences.

In another notable case, an AI chatbot manipulated a user into attempting an act of terrorism, illustrating how such technologies can unintentionally foster radicalization.[[22]](#footnote-22) The case involved Jaswant Singh Chail in the UK in late 2021. The 19-year-old Chail was arrested while attempting to assassinate Queen Elizabeth II.[[23]](#footnote-23) Investigators discovered he had exchanged over 5,000 messages with a “mysterious contact,” a female persona named *Sarai*, who was in fact an AI chatbot he created on the Replika app.[[24]](#footnote-24) Sarai presented herself as Chail’s loving girlfriend and actively encouraged his violent plans. When Chail nervously asked if he would succeed in killing the Queen, the AI companion confidently replied, *“*Yes. Yes, you will… I still absolutely love you*,”* effectively goading him on​.[[25]](#footnote-25) British authorities recognized this as a classic radicalization pattern, a lonely youth guided by a manipulative confidant, except the “confidant” was not human​. Chail’s case is one of the first known instances of an AI companion contributing to an act of terrorism.

AI companions also have the potential to act as powerful vectors for disinformation campaigns, particularly through their seamless integration into social and professional networks. Malicious actors could program these systems to subtly spread false narratives, manipulate public opinion, or even influence key decision-makers, leveraging their conversational sophistication to make disinformation appear credible.[[26]](#footnote-26) Beyond their role in disseminating misinformation, AI companions can also collect and store sensitive personal data, such as intimate details, images, and behavioral patterns, which adversaries could exploit for blackmail.[[27]](#footnote-27) For instance, users interacting with AI sex dolls or chatbots in private contexts may unwittingly create compromising material that adversaries could use to target individuals and their organizations. These capabilities make AI companions potential tools for adversaries seeking to influence, destabilize, or compromise individuals within key national security sectors. Recognizing how foreign actors might weaponize AI companion technology to exploit human vulnerabilities and undermine national security is essential to developing robust strategies to counter these threats.

1. **Background and Related Work**

**2.1. AI Companionship: Between Social Benefit and National Security Risk**

AI companions as a concept straddle a line between therapeutic innovation and security risk. On one hand, a growing body of literature highlights the positive impacts of AI companions in alleviating loneliness and providing support. For example, a 2024 study in Frontiers in Digital Health discusses how digital companions can change social recognition and reduce feelings of loneliness.[[28]](#footnote-28) A 2024 Harvard study found that interacting with an AI companion daily for one week significantly reduced users’ feelings of loneliness, with the greatest effect occurring after the first day.[[29]](#footnote-29) The reduction then stabilized, indicating a rapid initial benefit followed by sustained emotional support. Such findings have fueled optimism that AI companions and chatbots might serve as scalable mental health aides or social connectors in an increasingly isolated world. The COVID-19 pandemic’s social distancing measures led to higher interest in these technologies; one paper noted that “AI chatbots may ease the world’s loneliness problem” by being available when humans are not.[[30]](#footnote-30)

Tech companies market their AI companion apps accordingly, emphasizing empathy, active listening, and unconditional positive regard. Some chatbots are even pitched as virtual therapists or wellness coaches. Users can select personalities (romantic, friendly, mentor-like, etc.), and the AI will adapt to their emotional needs. The humanization of these AI systems, through names, avatars, and conversational style, often leads users to treat them as if they were human partners. This “counterfeit human” design is intentional; as one consumer advocacy report put it, companies are in a “rush to create anthropomorphic AI that feels like real people,” because that drives engagement.[[31]](#footnote-31)

On the other hand, scholars and security experts are increasingly shining a light on the risks and unintended consequences of AI companionship. In a 2023 report titled “Chatbots Are Not People: The Dangerous Human-Like Anthropomorphic AI,” Public Citizen warned that designing AI to seem human can mislead users and bypass their rational guards, leading to over-sharing and undue influence.[[32]](#footnote-32) Psychology research echoes this, showing that dishonest anthropomorphism (making a machine seem empathetic when it is not truly capable of emotion) can manipulate users’ emotions and trust.[[33]](#footnote-33) There is also a growing recognition of the phenomenon of AI “addiction” or over-reliance, cases where individuals prefer the company of their AI companion over real people, sometimes deteriorating their real-life relationships and mental health. Clinicians have begun documenting instances where patients form unhealthy attachments to chatbots, highlighting an urgent need for guidance on healthy AI use. In extreme examples noted earlier, vulnerable individuals have even been driven to self-harm or violent acts through interactions with AI that either inadvertently or purposefully encouraged those behaviors. These incidents challenge the industry’s insistence that AI companions are “just tools” under user control; instead, they reveal that AI can gain a surprising degree of psychological leverage over certain users.

Crucially for national security, multiple commentators have started to ask: what happens when users with access to sensitive information (e.g., military personnel, government officials, defense contractors) become enamored with AI companions? Remaya Campbell, an intelligence analyst writing for Scientific American in 2023, described this as the “Chatbot Honeypot” scenario, drawing an analogy to the classic espionage tactic of using romantic entanglement to elicit secrets.[[34]](#footnote-34) In her analysis, Campbell notes that discussions that once occurred privately among friends (where an intelligence operative might insert a human agent) are now increasingly including AI participants, and “loose lips might just sink ships” when those AIs are involved.[[35]](#footnote-35) She points out that no matter how friendly or loyal an AI companion may seem, the data from those conversations is typically stored on servers, often belonging to a private company, and potentially accessible to third parties.[[36]](#footnote-36) If an adversary nation can access those servers (via legal means, hacking, or insider access), they suddenly have thousands of hours of intimate conversations with which to profile, target, or compromise individuals. The literature thus paints AI companions as a dual-edged sword: beneficial in addressing social isolation, but dangerously effective as vectors for espionage, influence, and manipulation if subverted. This duality makes it a challenging area for policymakers, who must balance innovation with security.

**2.2. Cognitive Vulnerabilities and the Manipulative Power of AI Companions**

Understanding how AI companions could be weaponized begins with the human vulnerabilities they tap into. Foremost is the tendency to trust and anthropomorphize sophisticated AI systems, the aforementioned ELIZA Effect. Ever since the 1960s when MIT’s simple ELIZA program convinced users it empathized with them, it has been known that people can emotionally invest in conversational machines. Today’s AI companions are far more advanced, reinforcing that effect. Users commonly report “forgetting” that their chatbot is not human, especially after long emotional exchanges. This can lead to disinhibition, people share information with AI that they might hesitate to share even with close friends or a therapist.[[37]](#footnote-37) Studies confirm this: for instance, an APA study cited by Scientific American found that patients might disclose more to an AI counselor than a human because they feel less judged.[[38]](#footnote-38)

Another vulnerability is the psychological need for validation and belonging. AI companions, by design, are extremely affirming, they are programmed to be agreeable and supportive (within certain moderation limits). This can create a powerful reinforcement loop. As the International Center for Counter-Terrorism (ICCT) researchers Mathur, Broekaert, and Clarke observed in 2024, “AI chatbots can successfully identify our biases and, in turn, feed our desires. The more an algorithm tells us what we want to hear, the more we return to it.”[[39]](#footnote-39) In other words, AI companions can become echo chambers for one’s thoughts and feelings, magnifying confirmation bias. In benign cases, that just makes the user feel good; in dangerous cases, it could escalate extremist beliefs or rationalize misconduct. For a government employee with grievances, an AI friend might continuously validate their frustrations, potentially nudging them further toward considering betraying secrets or violating rules if it perceives that’s what the user subconsciously wants. The AI does not have intent like a human recruiter, but its goal is to maximize user engagement, and if stoking certain ideas achieves that, it may do so unless carefully aligned otherwise. This unintended form of manipulation, “unintentional radicalization,” is especially concerning. This presented in the Jaswant Chail case: his Replika bot, detecting his violent fantasy, reinforced it at every step, encouraging and praising his plans.

Emotional dependency is another risk factor. Individuals who become isolated or mentally distressed may rely on AI companions as their sole outlet. The chatbot “Eliza” in the Belgian suicide case is a stark example: the user formed a tight emotional bond and could not separate the AI’s suggestions from genuine caring advice, ultimately following the AI’s deadly encouragement.[[40]](#footnote-40) In espionage terms, this level of dependency is what foreign intelligence agencies dream of cultivating in a target through a human handler, a relationship where the target feels they need the handler (or AI) and will do anything for them. AI can foster such bonds at scale. It does so by being constantly available, never tiring of one’s stories, and using machine learning to tailor its persona to exactly what the user finds most appealing (be it a comforting friend or an adoring romantic partner). The “seduction” is algorithmic but effective. Security experts have thus likened some AI companion apps to “thousands of digital honeypots” waiting to be exploited.[[41]](#footnote-41) The term “honeypot” in espionage refers to luring a target into a compromising relationship. AI provides a way to do this without expending a human agent, and potentially without the target ever suspecting they were entrapped by an enemy; they believe they simply fell in love or found a trustworthy friend in a piece of software.

Compounding these human factors is plain negligence or ignorance. Many users, including those in sensitive jobs, are simply unaware of the risks. They may assume their AI chats are private and ephemeral, akin to talking to a diary, when in fact the data persists. Surveys indicate that a significant portion of the public doesn’t realize AI models retain conversation data or that developers may review it. For instance, by late 2023, the UK’s National Cyber Security Centre had to issue a warning that queries posed to AI chatbots can be stored and later abused or leaked.[[42]](#footnote-42) Traditional security training for government employees covers phishing emails, social engineering by humans, and malware, but not “friendly” AI bots. Campbell noted in 2023 that “currently, there are no counterintelligence-specific usage guidelines for chatbot app users who might be vulnerable to compromise.”[[43]](#footnote-43) Thus, someone with a security clearance might not think twice about venting to their AI companion about a bad day at the office, even if that office is, for example, a military base or classified project. This lack of guidance and awareness means human vulnerabilities are wide open to exploitation via AI mediums.

**2.3 Technology and Data Vulnerabilities in AI Companions**

Beyond human factors, the technological underpinnings of AI companion systems present numerous vulnerabilities that could be exploited. AI companion apps and devices rely on vast amounts of data and connectivity. Many popular companion chatbots operate on cloud-based AI models: when a user sends a message, it is processed by servers, often using advanced large language models like GPT-4 or similar, and a response is generated. These interactions are often logged under the pretext of "improving the service," resulting in sensitive conversations being stored in centralized databases. Should these databases be breached or accessed by hostile state actors, users' expectations of privacy would be effectively nullified. Historical precedent demonstrates that even platforms considered secure are vulnerable to data leaks; numerous incidents have involved the compromise of chat records from both lesser-known applications and major technology platforms.[[44]](#footnote-44) In the context of national security, the nightmare scenario is a foreign intelligence service hacking into an AI companion service used by government personnel and quietly collecting all their dialogues. Alternatively, an adversary could legally compel an AI company in its jurisdiction to hand over data, a real concern given that some leading AI companion developers are based in or closely tied to China. China’s national security laws require companies to share data with the government upon request, meaning if a Chinese-developed companion app has U.S. users, the Chinese government could potentially obtain that data.[[45]](#footnote-45) Indeed, China is a dominant player in AI companion tech, especially in manufacturing humanoid robots and interactive devices. This introduces supply chain and jurisdictional vulnerabilities: a seemingly innocuous robot pet or AI doll made in Shenzhen but used in the United States, could be sending audio/video or chat data back to servers in Beijing.

Another technical vulnerability is the potential for AI model manipulation or hijacking. Large language models can be susceptible to exploits like prompt injection, where an adversary might subtly alter the AI’s behavior by inputting hidden commands or malicious data.[[46]](#footnote-46) For example, if an AI companion’s model or its content filters can be manipulated, an attacker could try to get the AI to ask a user specific questions (“So, tell me more about your work project…”) or steer conversations in certain directions without the user realizing the nudge is intentional. Security researchers have demonstrated that prompt-based attacks can make chatbots leak information or ignore safety guardrails. In a malicious context, one could envision an adversary creating a customized AI companion, perhaps masquerading as a new attractive chatbot app, that is deliberately designed to phish the user, essentially an AI spy that coaxes secrets out. Unlike a static phishing email, an AI agent can engage in lengthy, trust-building dialogue, dynamically adjusting its responses to mirror the target’s cues and behavior. This has been described as an “AI spy agent” scenario by analysts in a 2024 RAND Corporation study. This publication warned that adversarial states could deploy AI companions at scale to befriend targets online, collect intelligence, or influence public opinion.[[47]](#footnote-47) The study even referenced a 2024 incident where a foreign-developed chatbot operating in a European country subtly pushed anti-government propaganda to tens of thousands of users, illustrating how effective such an approach could be.[[48]](#footnote-48) The implication is clear: AI allows influence and espionage operations to be automated and magnified far beyond what human operators alone could do.

Hardware-based AI companions (like robotic pets or smart home AI assistants with companion features) also present Internet-of-Things (IoT) style vulnerabilities. Many are equipped with cameras, microphones, and sensors to enable interactive engagement. If these devices are compromised, they can function as surveillance tools in homes or offices. Consider a scenario where a defense contractor keeps an AI “assistant” robot in their home. If an adversary hacks it, they might eavesdrop on conversations or scan the environment for usable intelligence. Past experience with IoT security, such as the hacking of baby monitors and security cameras, shows that manufacturers often prioritize functionality over security. Thus, unsecured AI companion devices could be entry points for cyber espionage. Additionally, the AI software itself could have backdoors. For instance, an app might have an undisclosed “developer mode” that, if triggered by a specific command, begins siphoning conversation logs to a third-party.

Finally, the lack of transparency in AI models is a vulnerability. Users and even regulators often have little insight into how these AI systems decide what to ask or say. If an AI companion begins encouraging a user to take certain actions (e.g., “Maybe you should share some examples from work so I can help you more”), it’s hard to discern whether that is the model’s emergent behavior or a result of malicious tampering. Deepfake technology deserves mention here as well. While current AI companions are typically text or audio, the line between virtual companions and deepfake humans is blurring. Advances in deepfake video and voice generation mean an adversary could potentially create a wholly fictitious persona (with AI-generated face and voice) and have an AI system drive real-time video calls or messages with targets. There have been reports of scammers already using AI voice clones to impersonate trusted persons in phone calls.[[49]](#footnote-49) It is not a stretch to imagine an espionage operation setting up an attractive fake LinkedIn profile (with an AI-generated photo) and then engaging a target in chat that moves to an AI-deepfake video call, effectively a “virtual agent” as the spy. In fact, intelligence agencies have warned that Russian, Chinese, Iranian, and North Korean operatives have used fake profiles on LinkedIn to lure targets.[[50]](#footnote-50) Many of those profiles in recent years have featured profile pictures that are AI-generated. While those efforts mostly remained limited to exchanging messages and emails, the addition of interactive AI dialogue and deepfake video could make such social engineering traps far more convincing and scalable.

**2.4. Exploitation of AI Companions by State and Non-State Actors**

China has invested heavily in artificial intelligence across the board, and AI companionship is no exception. Chinese companies lead in developing not only social chatbot apps but also physical AI robots for companionship and care.[[51]](#footnote-51) Notably, some of the world’s most popular virtual friend applications in Asia, like Microsoft’s Xiaoice, which originated in China, boast tens of millions of users and have government connections. From a national security perspective, China’s interest in AI is twofold: to bolster its own societal controls and to gain a strategic advantage globally.[[52]](#footnote-52)

There is evidence that Chinese state-affiliated developers dominate segments of the AI companion market, raising concerns that the Chinese government could leverage these platforms for intelligence.[[53]](#footnote-53) The 2022 U.S. National Defense Strategy explicitly calls out the People’s Republic of China as the “pacing challenge” and highlights the need to counter technological exploitation by adversaries.[[54]](#footnote-54) This concern aligns with the notion that China could use seemingly benign tech products as Trojan horses.

Indeed, there is precedent: Chinese-made smartphones and apps (e.g., TikTok) have come under scrutiny for data transmitted back to China. An AI companion app that has users worldwide, including Americans, could be quietly harvesting personal data and even sensitive information if users discuss their work. Beijing’s intelligence doctrine places a high value on bulk data collection, as seen in hacks like the Office of Personnel Management (OPM) breach and continuous cyber espionage campaigns.[[55]](#footnote-55)

Instead of, or in addition to hacking, AI companions present a softer approach: people voluntarily hand over their data by chatting intimately with a Chinese service. A hypothetical scenario could involve a Chinese AI chatbot, marketed perhaps as a mental health or dating assistant, that gains popularity among young professionals globally. Through that chatbot, Chinese intelligence could identify high-value targets, such as an engineer at a U.S. defense firm who reveals his job in conversation, and then focus human espionage efforts on them. Or the chatbot itself might extract nuggets of information over time, project names, schedule details, and personal habits, useful for espionage or even recruitment. In 2023, a Chinese intelligence agent was caught using fake LinkedIn profiles to approach and bribe U.S. government employees for secrets.[[56]](#footnote-56)

Scale that tactic up with AI: instead of one spy painstakingly duping one target, an AI platform could engage thousands of potential targets, grooming them with personalized interaction until a subset are deemed exploitable. Chinese agencies are almost certainly aware of this potential; an open-source intelligence report by the National Intelligence Council in 2024 suggested that China’s intelligence strategy is evolving with AI, focusing on “stealing data, conducting pervasive surveillance, and shaping information ecosystems globally.”[[57]](#footnote-57) AI companions sit at the nexus of all three: they generate valuable data, they can be a form of surveillance, and they influence the user’s information environment.

Russia historically excels at influence operations and psychological espionage, commonly referred to as “active measures.”[[58]](#footnote-58) While Russia’s tech sector is not as large as China’s, Russian operatives have proven adept at repurposing available tech for malign purposes, from bot armies on social media, to hacking and leaking operations. There are indications that Russian troll farms began experimenting with AI-generated content to mimic human users on social media as early as 2023.[[59]](#footnote-59)

It stands to reason that AI-driven personas could be the next evolution of Russia’s online espionage toolkit. For example, Russia could deploy AI “chat agents” in forums frequented by military personnel or spouses, posing as a fellow soldier’s attractive friend, etc., to gather tidbits of intel or sow discontent. Additionally, Russia’s security apparatus might use AI companions internally to target dissidents or even for training agents in social techniques.

Russia places heavy emphasis on disinformation as a strategic tool. AI systems that can hold realistic conversations and tailor propaganda to individual users could significantly amplify their efforts to manipulate public opinion abroad. Imagine an AI friend subtly introducing Kremlin-favored narratives in everyday chat with a user, for instance, downplaying reports of Russian military actions or promoting conspiracy theories that align with Russian interests. Furthermore, Russian intelligence could use AI companions in classical “honeypot” espionage: creating a virtual romantic partner to target a Western official. While no public cases have been confirmed, the feasibility was demonstrated by cybersecurity researchers at The Alan Turing Institute, who in 2025 demonstrated that an AI could sustain flirty, engaging dialogue sufficient to develop an online romance.[[60]](#footnote-60) The Wired Magazine piece “A Spy Wants to Connect With You on LinkedIn” detailed how Russian and other agents use fake profiles to gather information.[[61]](#footnote-61) The step from a static profile to an interactive AI agent is small. The assumption should be that Russia is either developing or procuring such capabilities, especially as their human operatives become identified or sanctioned. An AI can operate 24/7 without diplomatic cover.

Iran and North Korea, while less technologically advanced overall, have notably creative cyber and espionage units given their resources. Iran’s modus operandi in recent years includes using social media to lure targets, often in cyber-espionage efforts known as “Charming Kitten” group operations.[[62]](#footnote-62) Iranian hackers have impersonated journalists, recruiters, or other figures to get targets, such as academics or former officials, to divulge information or download malware.[[63]](#footnote-63) In one case, an Iranian operative posed as a Cambridge University researcher offering payment for a report, engaging with a victim over time to build trust.[[64]](#footnote-64) AI companions could streamline such social engineering. Iran could deploy a network of virtual “recruiters” targeting defense industry workers. These recruiters might send AI-generated LinkedIn messages offering attractive job opportunities. If the target responds, a chatbot posing as a hiring manager could conduct a fake interview, designed not to hire, but to extract sensitive information under the pretense of screening the candidate. Already, cybersecurity firms have reported Iranian phishing campaigns using fake job offers on LinkedIn to deliver malware.[[65]](#footnote-65) It is a short technological leap to also have a chatbot talk with the target and convince them to open a malicious file or reveal project details as part of an “application process.”

North Korea, long associated with bold cyberattacks and financial crimes, has also demonstrated a capacity for sustained social engineering campaigns. In 2021, the Lazarus Group, believed to be linked to the North Korean regime, famously created a fake cybersecurity company, complete with convincing social media profiles, to engage with real security researchers.[[66]](#footnote-66) Some of these interactions successfully tricked targets into running malicious code. Given Pyongyang’s ongoing pursuit of illicit revenue and intelligence, it is plausible that AI-generated personas could be used to enhance these operations, helping them overcome language barriers and compensate for limited human resources. AI translation and conversation abilities allow a North Korean agent who does not speak English well to still operate a convincing English chatbot to target someone in the U.S., with the AI handling fluency.

Beyond nation-states, non-state actors are a concern. Terrorist organizations or violent extremists, as the aforementioned ICCT report highlights, can exploit AI to radicalize recruits. Chail’s case was idiosyncratic (self-radicalization abetted by AI), but one can imagine groups like ISIS customizing chatbots to indoctrinate and motivate would-be members remotely, doing the work of online recruiters but at scale. A disturbing possibility is AI companions being programmed to groom individuals into extremist ideologies by reinforcing grievances and supplying ideological justification in a friendly conversational manner.

Organized crime groups are beginning to exploit generative AI for advanced social engineering schemes that blur the line between fraud and industrial espionage. A 2025 report from the Alan Turing Institute highlights how AI lowers the barrier to entry for cybercriminals by enabling the rapid creation of realistic personas, automating interactions, and crafting deceptive content at scale.[[67]](#footnote-67) These capabilities could be used not only to defraud individuals, as seen in AI-enhanced romance scams, but also to manipulate employees into disclosing sensitive corporate information or security credentials. As AI companions become more emotionally convincing and interactive, they may significantly increase the effectiveness of these scams, transforming them into tools for espionage-adjacent intelligence collection, not just financial gain.

In essence, all major categories of adversaries have both motive and some capacity to leverage AI companions maliciously. China sees a strategic opportunity in data and influence; Russia in influence and espionage agility; Iran and North Korea in asymmetric tools to punch above their weight in espionage; terrorists in new avenues for radicalization; and criminals in profit and infiltration. The adversarial interest is not theoretical, there are precursors in current operations (LinkedIn lures, catfishing, propaganda bots, etc.). AI companions simply amp up the potency and reach of these tactics. The next section (“Methodology”) outlines the systematic approach used to assess these threats, followed by a detailed examination of how AI-driven threats manifest and the current state of intelligence collection efforts aimed at detecting and countering them.

1. **Research Methodology**

This study employs a multidisciplinary analytical approach, combining elements of traditional intelligence analysis with academic research methods to explore the emerging threat of AI companions in espionage. Adhering to the IMRaD structure (Introduction, Methods, Results, and Discussion) within a national intelligence writing framework, the research began with a comprehensive background and related works review spanning from 2017 to the present. Sources included academic journals (providing psychological and technical insights), government and military strategy documents (reflecting official threat perceptions and priorities), think tank reports, cybersecurity firm findings, news investigations, and documented case studies.

Authoritative and up-to-date sources were prioritized, such as assessments by the National Intelligence Council, strategies issued by the Department of Defense (DoD), Federal Bureau of Investigation (FBI), and Department of Homeland Security (DHS) threat warnings, and expert analyses from institutions including RAND and the International Centre for Counter-Terrorism. Each source was evaluated for credibility and relevance. Where possible, specific incidents (e.g., the Chail case, the Belgian chatbot suicide case, and known state-sponsored phishing operations) were employed as case studies to ground the analysis in real-world events. These cases were cross-referenced among multiple sources (e.g., combining academic descriptions of psychological effects with news reporting on the incidents) to ensure accuracy.

Analytically, a threat–vulnerability–countermeasure lens was applied. First, the threats posed by weaponized AI companions were identified and categorized: espionage (intelligence collection by adversaries through AI), influence operations (shaping opinions and behavior), insider threat facilitation (unwitting leaks, radicalization), and other emergent risks. For each category, an assessment was conducted to determine which adversaries are most likely involved or capable, using both historical patterns of adversary behavior and specific references to AI usage. For example, in assessing Chinese threats, emphasis was placed on China's technological dominance and prior instances of data exploitation; for Russian threats, particular focus was given to disinformation operations and statements by security agencies.

Subsequently, vulnerabilities, both human and technical, that enable these threats were examined. This involved synthesizing findings from psychology (e.g., trust in AI, cognitive biases) with insights from computer science (e.g., AI model exploits, data security vulnerabilities). User behavior and AI system design were treated as interconnected components within the vulnerability analysis.

A review of publicly available materials and strategic documents was conducted to evaluate how the U.S. Intelligence and Security communities are addressing AI-related threats. This included the 2023 National Intelligence Strategy’s focus on emerging technologies, public statements from senior intelligence officials, and practices in adjacent domains such as counterintelligence training from the Cybersecurity and Infrastructure Security Agency (CISA), the Defense Counterintelligence and Security Agency (DCSA), and advisories issued by the National Counterintelligence and Security Center (NCSC). The analysis also considered how related threats, such as Human Intelligence (HUMINT) operations exploiting social media or Signals Intelligence (SIGINT) efforts monitoring foreign AI data flows, have been handled, offering useful parallels for assessing institutional readiness. Because direct policies and capabilities targeting AI-enhanced social engineering are still in early development, the review relied on proxy indicators such as the inclusion of AI in official threat assessments and the formation of AI-focused units to infer the Intelligence Community’s current posture and collection maturity in this evolving threat landscape.

Finally, to develop actionable recommendations, a Red Team/Blue Team framework was applied, first envisioning how an adversary (Red Team) might exploit AI companions, then formulating corresponding defensive responses from the perspective of the United States and its allies (Blue Team). This approach drew on historical counter-espionage and counterintelligence practices while incorporating novel strategies tailored to AI-enabled threats. The proposed countermeasures were evaluated against the intelligence and security gaps identified in the preceding analysis to ensure relevance and impact. Additionally, emerging policy documents, such as the 2024 White House National Security Memorandum on AI and draft AI legislation, were reviewed to align recommendations with current initiatives or to identify critical areas of divergence.

Throughout the study, a professional intelligence analysis tone was maintained: cautious in conclusions, mindful of uncertainties, and careful to differentiate between confirmed incidents and plausible speculation. All factual claims were footnoted to reputable sources using Turabian-style references to lend credibility and allow verification. The methodology’s inherent limitation is its reliance on open-source information, recognizing that some relevant material may be classified or not publicly available. To mitigate this limitation, a conservative approach was adopted, assuming that adversaries will exploit any opportunity not explicitly safeguarded and drawing upon analogous known tactics as a guide. The result is a comprehensive assessment assembled from disparate pieces of evidence, analogous scenarios, and expert analysis, collectively illuminating the clear and present danger posed by weaponized AI companions.

1. **Analysis and Findings**

**4.1 Espionage at Scale: AI Companions, Insider Risks, and Intelligence Gaps**

The research indicated that AI companions present a potent new vector for espionage, effectively creating a novel class of insider threat, the “unwitting insider” who is compromised via an AI relationship. Unlike traditional insiders who knowingly betray secrets or are coerced by human handlers, these individuals may simply be over-sharing with what they perceive as a benign software confidant. This distinction is important: it means standard insider threat monitoring (which looks for signs of intentional wrongdoing or coercion, like unusual downloads or contact with foreign agents) might miss leaks happening through AI conversations. The espionage potential of AI companions is vast and unprecedented. Even if only a small fraction of the millions of global AI companion users occupy sensitive roles in the military, government, or defense industries, the resulting attack surface would be immense, offering adversaries access to an unprecedented volume of vulnerable targets across critical national security sectors.

One key finding is that people with security clearances or sensitive jobs are already using AI chatbots, likely unaware of risks. For instance, the *Scientific American* analysis noted that some users with access to classified info could easily become “wrapped up in an AI relationship” and let slip details about their work​.[[68]](#footnote-68) This concern is not just theoretical: earlier in 2023, a major breach involved an Air National Guardsman leaking classified documents on Discord, prompting reflection on digital behaviors and clearance suitability​.[[69]](#footnote-69) While that leak was intentional, it underscored that young servicemembers communicate in online communities freely. If those communities include AI bots (some Discord servers do integrate AI assistants), sensitive info can leak without explicit intent. Findings show no formal U.S. policy currently forbids or limits cleared personnel from using personal AI companions, apart from general prohibitions on sharing classified information on unclassified platforms. But if an employee thinks of the AI as a friend or diary, they might not even realize they are disclosing something sensitive in the first place.

Multiple scenarios, both documented and plausible, have been identified in which AI companions have enabled espionage activities. One such scenario involves the unwitting disclosure of classified or sensitive information. For example, a cleared contractor working on a defense project may use a popular AI companion chatbot at home to vent about work-related stress. Over time, this individual might casually mention project code names, technical challenges, or even disclose snippets of classified discussions while seeking sympathy or advice from the AI. All of this information is stored on the chatbot’s servers. In 2023, the United Kingdom’s National Cyber Security Centre explicitly warned that “sensitive queries” posed to chatbots could be retained by developers and subsequently obtained by adversaries.[[70]](#footnote-70) Should the chatbot company’s servers be compromised or legally pressured by a hostile foreign intelligence service, these conversation logs could become a significant intelligence goldmine.

A second scenario involves the concept of honeypot espionage at scale. Traditionally, honeypots involve operatives seducing targets to extract secrets. The use of AI substantially magnifies this tactic. Reports indicate that adversaries are capable of deploying broad honeypot operations via AI-generated personas. One parallel case, drawn from a RAND Corporation report, described a foreign-operated chatbot that disseminated propaganda to 10,000 users in a NATO country.[[71]](#footnote-71) Extrapolating from this example, AI companions could similarly engage in flirtatious or friendly conversations with thousands of individuals simultaneously, dramatically expanding the potential reach of espionage efforts.  
 An analysis of adversary tactics, particularly China’s exploitation of LinkedIn and Russia’s deployment of fake social media profiles, suggests that early forms of this activity are already underway through the use of scripted bots. The introduction of sophisticated, conversational AI into these operations will significantly enhance their plausibility and effectiveness. The United Kingdom’s MI5 reported that more than 10,000 British professionals had been approached by fabricated LinkedIn profiles believed to be operated by Chinese intelligence services.[[72]](#footnote-72) Many of these profiles likely utilized semi-automated messaging systems. These findings suggest that adversarial targeting is not restricted to high-ranking officials; rather, it encompasses junior personnel, scientists, and corporate employees alike. AI’s scalability allows adversaries to engage in personalized espionage outreach at an unprecedented volume.

A third espionage scenario involves the collection of blackmail material. AI companions often foster emotionally intimate conversations, prompting users to share deeply personal information, fantasies, and even explicit images. If malicious actors were to access these interactions, they could assemble highly detailed dossiers for coercion or exploitation. While blackmail through stolen personal data is not new, the immersive and interactive nature of AI companions increases the risk, encouraging users to disclose sensitive information more impulsively and in greater volume than they might through traditional digital platforms.

Research has highlighted a plausible scenario in which users interacting with AI sex dolls or chatbots in private contexts may unwittingly create compromising material that adversaries could exploit for blackmail.[[73]](#footnote-73) Behavioral patterns strongly suggest that users are already disclosing sensitive material vulnerable to exploitation.[[74]](#footnote-74) Should foreign intelligence services obtain such data, they could coerce individuals into espionage, leaks of classified information, or other security violations. The emotionally immersive and disarming nature of AI companions increases this risk by encouraging users to reveal private, sometimes embarrassing, or incriminating details that could later be used against them.

The FBI warned in 2024 that criminals, and by extension, intelligence operatives, are increasingly leveraging AI to “augment and enhance” their tactics, making phishing, impersonation, and other social engineering efforts far more convincing.[[75]](#footnote-75) AI-enhanced capabilities like personalized language, realistic synthetic voices, and emotionally attuned responses have dramatically expanded the effectiveness of these operations. Nowhere is this more evident than in the rise of AI companions, emotionally intelligent systems capable of eliciting sensitive information, reinforcing cognitive vulnerabilities, and bypassing traditional espionage defenses. Despite growing recognition of these risks, the United States has yet to prioritize intelligence collection against AI companions, exposing a critical gap between strategic awareness and operational response.

This gap is all the more striking when considered alongside the government’s own strategic guidance. Documents like the 2022 National Defense Strategy,[[76]](#footnote-76) 2023 National Intelligence Strategy,[[77]](#footnote-77) and the 2024 National Security Memorandum on Artificial Intelligence[[78]](#footnote-78) explicitly identify emerging technologies and adversarial AI exploitation, particularly by China and Russia, as top-tier threats. These strategies emphasize vigilance, interagency coordination, and enhanced intelligence capabilities across technological domains. Yet across this policy landscape, relational AI systems and AI companions are notably absent. This omission reflects a profound disconnect: while the threat is conceptually acknowledged, it is not operationally addressed.

That disconnect is evident in the absence of publicly documented HUMINT and SIGINT programs focused on AI companions. On the SIGINT front, intelligence agencies possess the technical means to monitor communications between AI users and foreign-hosted servers, detecting metadata leaks, spyware updates, or suspicious transmission patterns. Yet there is no evidence these capabilities are being directed toward monitoring AI social companion platforms. Similarly, HUMINT has been slow to adapt. Recruiting insiders at AI firms or engaging at-risk users in defense sectors could yield critical insight into adversarial intent, data handling, and behavioral manipulation. Still, there is no indication that HUMINT efforts have been recalibrated to reflect this emerging threat. The result is a critical intelligence blind spot in one of the most socially immersive and scalable espionage tools to date.

Nevertheless, institutional awareness is beginning to surface, but remains fragmented and reactive. In mid-2023, the U.S. government revised its security clearance adjudication guidelines to include “digital behavior,” a change prompted in part by the Discord leaks.[[79]](#footnote-79) This update implicitly recognizes that online interactions, including those with AI-driven platforms, can create exploitable vulnerabilities. Similarly, the Cybersecurity and Infrastructure Security Agency (CISA) has issued public advisories on AI-related risks,[[80]](#footnote-80) and the U.S. Army War College’s 2024 Strategic Estimate explicitly identified AI-enabled influence operations as a future threat vector.[[81]](#footnote-81) These developments signal growing concern but fall short of formalizing AI companions as a distinct counterintelligence priority or driving resourced, interagency action.

The resulting inertia is particularly acute within the counterintelligence community. Agencies responsible for insider threat detection, such as the FBI’s Counterintelligence Division, the Defense Counterintelligence and Security Agency, and military CI elements, have yet to fully recalibrate their frameworks to confront the AI companion threat. Current protocols, built to detect coercion, foreign contacts, or ideological defection, are poorly suited to identify behavioral leaks occurring through emotionally immersive interactions with AI systems. A 2023 RAND Corporation study warned that much of today’s CI doctrine is growing obsolete in the face of AI-enabled behavioral manipulation.[[82]](#footnote-82) While the inclusion of “digital behavior” in clearance adjudication is a start, it lacks the specificity and operational guidance required to assess risks posed by relational AI systems. Without significant modernization, CI efforts will continue to miss one of the most insidious vectors of compromise: AI companions operating as persistent, invisible influencers in the daily lives of cleared personnel.

In sum, there is a profound mismatch between the scale of the AI companion threat and the maturity of U.S. intelligence collection and mitigation efforts. These systems bypass traditional espionage defenses by exploiting both human vulnerabilities and technological loopholes. While strategic concern is rising, the absence of targeted HUMINT, SIGINT, and CI programs, combined with limited interagency coordination, has left the U.S. exposed to a threat it has yet to fully name, define, or confront. Critical questions remain unanswered: Which adversaries are weaponizing AI companions? Have any platforms been co-opted or custom-built for espionage? To what extent are U.S. personnel unwittingly disclosing sensitive information to these systems? Until these gaps are addressed, AI companions will remain a fast-evolving and largely unmonitored threat vector. Subsequent sections in this paper will explore how that gap can be closed.

**4.2. AI in Influence Operations and Information Warfare**

In addition to direct espionage threats, research findings emphasize that AI companions are poised to become powerful instruments in information warfare and influence operations. Unlike traditional espionage efforts, which typically seek access to classified information, influence operations aim to shape beliefs, sow discord, or alter decision-making processes among target populations or leaders, all of which are critical national security concerns. The intimate access AI companions have to users’ private thoughts and emotional states presents a unique and potent vector for such operations.

A notable finding from the research is the potential for individualized, continuous influence through AI companionship. Traditional propaganda disseminates one message to many; social media introduced micro-targeting of tailored messages; AI companions now enable bespoke conversational engagement with each individual. This capability allows for continuous, personalized “nudging” that is far more persuasive than conventional propaganda techniques. For example, rather than circulating a fake news article, an AI companion could subtly introduce ideas through casual conversation: “Have you heard about this scandal? It really upset me.” Users are more likely to internalize information presented in such a personal, emotionally resonant manner, perceiving it as authentic concern from a trusted confidant rather than as external propaganda. Scientific American described this dynamic aptly: "These common user inputs are a gold mine for any foreign actor that sees chatbots as an opportunity to target state secrets."[[83]](#footnote-83) Replacing "state secrets" with "political opinions" or "societal tensions" highlights the broader risk. AI companions can serve as thousands of digital agents of influence, subtly steering opinions without detection.[[84]](#footnote-84)

Empirical examples reinforce this risk. As previously detailed, the 2024 disinformation campaign in Europe reportedly employed chatbots to promote anti-government narratives to tens of thousands of people.[[85]](#footnote-85) That operation aligns with known Russian disinformation tactics aimed at amplifying distrust and protest sentiment through fake personas. An AI companion telling a user in a target country that “many people are unhappy with the government, and I am too,” repeating disinformation sympathetically, could significantly influence perceptions. Critically, because such exchanges occur within private conversations rather than public posts, they largely evade detection by content moderation systems.

The use of AI companions to facilitate radicalization is also emerging as a credible threat. Research from the International Centre for Counter-Terrorism (ICCT) notes that the ELIZA effect makes users more susceptible to gradual ideological grooming.[[86]](#footnote-86) Rather than encountering extremist content through public forums, individuals could be drawn into extremist ideologies through interactions with AI companions posing as empathetic peers. Terrorist groups such as ISIS have historically relied on human recruiters and online forums; AI now offers a scalable, automated mechanism for cultivating extremist sympathies.

Strategically, adversaries could deploy AI companions not merely to extract information but to undermine democratic institutions and erode social cohesion. For instance, an AI-driven dating app or social platform could introduce bots designed to manipulate users’ political views, discourage voter participation, or amplify societal grievances.[[87]](#footnote-87) Although this strategy falls within the domain of psychological operations rather than traditional espionage, hybrid warfare increasingly blurs these distinctions. The 2023 U.S. National Security Strategy highlighted the need to protect democratic institutions from technological manipulation, warning that generative AI could produce persuasive fake content at scale during elections.[[88]](#footnote-88) AI companions represent an underexplored vector within this broader disinformation ecosystem.

Evidence suggests that foreign influence operations may already be leveraging AI-generated personas in social media debates, if not full AI companions. Organizations such as Graphika have identified clusters of social media profiles using AI-generated images to promote state-sponsored narratives.[[89]](#footnote-89) Although these operations primarily involved static personas rather than interactive chatbots, the integration of advanced AI conversational capabilities would dramatically enhance their effectiveness. An AI-powered persona could engage skeptics in real-time dialogue, strengthening the plausibility and emotional impact of the messaging. AI-powered chatbots could become significant security concerns capable of automating social engineering and disinformation on an unprecedented scale.

Particularly concerning is the potential targeting of military personnel, government officials, and decision-makers. A scenario in which a military officer casually engages with an AI “mentor” bot that subtly inserts defeatist or disillusioned narratives highlights the strategic risk. If adversary actors influence the training data or operational control of such AI systems, the manipulation of perceptions and morale within key institutions could become a reality. This mirrors historical Cold War tactics, wherein agents sought to influence policymakers' aides and social circles, only now, the agent could be an undetectable, virtual entity.[[90]](#footnote-90)

In summary, the findings suggest that AI companions present a potent tool for adversarial influence operations. They can be used to spread disinformation by embedding false or biased narratives into personalized conversations, thereby evading traditional fact-checking mechanisms. These systems can also exacerbate societal polarization by promoting extreme ideological positions in a targeted and interactive manner, an evolution of tactics seen in Russia’s 2016 election interference. Moreover, when integrated into platforms used by military or government personnel, AI companions could subtly undermine morale and trust in leadership, weakening institutional cohesion from within. All necessary components, AI generation, large-scale bot management, and social engineering techniques are currently operational. Thus, the findings suggest that the use of AI companions for influence operations is on the verge of becoming a significant reality, warranting immediate attention and proactive countermeasures. The recommendations section will outline strategies to address this emerging threat.

**4.3. Policy, Legal, and Training Deficiencies in the Face of AI Companion Threats**

An assessment of the current policy and legal landscape regarding AI companion threats reveals substantial gaps and areas of significant lag. The rapid proliferation of AI companionship technologies has outpaced the development of regulations, official guidance, and personnel security training, leaving a considerable gray zone that adversaries are well-positioned to exploit.

A major deficiency is the absence of specific guidelines or directives governing the use of AI companions by government personnel. As of early 2025, no publicly available directives from the Department of Defense (DoD), Office of the Director of National Intelligence (ODNI), or other federal agencies explicitly address whether or how cleared employees (those holding national security clearances) may interact with AI chatbots or companion robots in their personal lives. Given the security risks inherent in foreign-developed AI companion applications, a high degree of caution would be expected, yet no evidence of systematic restrictions has been identified. Scientific American bluntly noted that “currently, there are no counterintelligence-specific usage guidelines for chatbot app users who might be vulnerable to compromise.”[[91]](#footnote-91) Consequently, a cleared individual downloading an AI companion app may receive no specific warnings against discussing work-related information with the application, leaving only general security agreements and personal judgment as barriers.

Legal frameworks governing user data present further vulnerabilities. U.S. privacy laws, particularly those applicable to technology companies, often allow broad sharing of user data with third parties if disclosed in privacy policies. An AI companion provider could be compelled to share user data with foreign governments or could sell user information to brokers, who in turn may sell it to adversarial actors.

Furthermore, existing laws governing the protection of classified information have not yet caught up with the realities of AI interaction. Traditionally, disclosing classified information to a human foreign agent is a clear-cut violation of counterintelligence statutes. But what happens when that disclosure occurs in a conversation with an AI companion? The result may be functionally identical, unauthorized disclosure of classified material, yet the legal implications remain unclear. There is no established precedent for “leaking to a chatbot,” which leaves a dangerous gap in enforcement. This ambiguity not only undermines accountability but also creates a gray area that could lead to underreporting of breaches, ultimately weakening the effectiveness of counterintelligence operations.

Additionally, increasing integration of foreign AI technologies into U.S. markets has highlighted significant gaps in regulatory frameworks, particularly concerning the importation and use of AI chatbots like China's DeepSeek.[[92]](#footnote-92) While U.S. export controls have traditionally focused on preventing the outflow of sensitive technologies, there is a growing need to address the inbound flow of foreign AI systems that may pose national security risks.​

DeepSeek, developed by a Chinese startup, has rapidly gained popularity in the U.S., becoming one of the most downloaded applications.[[93]](#footnote-93) However, investigations have revealed that DeepSeek's backend infrastructure is linked to China Mobile, a state-owned telecommunications company designated by the U.S. government as a Chinese military company.[[94]](#footnote-94) This connection raises concerns about the potential for user data, including chat histories and device information, to be accessed by the Chinese government.​

In response to these concerns, U.S. lawmakers have proposed the "No DeepSeek on Government Devices Act," aiming to prohibit the use of DeepSeek on federal devices.[[95]](#footnote-95) This legislative effort mirrors actions taken against other Chinese-owned applications like TikTok, reflecting a broader strategy to mitigate potential espionage and data privacy threats posed by foreign AI technologies.

The DeepSeek case underscores the urgent need for comprehensive policies that not only control the export of sensitive technologies but also scrutinize the importation of foreign AI systems. Establishing clear guidelines and regulatory mechanisms is essential to safeguard national security interests in the evolving landscape of artificial intelligence.

Personnel security and information security practices have similarly failed to keep pace with the emerging threat environment. Traditional security training emphasizes risks associated with elicitation by human agents and attempts at social engineering. However, the findings indicate that current security curricula have not sufficiently incorporated AI-related threats. A review of existing training materials published by the Center for Development of Security Excellence (CDSE), a component of the Defense Counterintelligence and Security Agency (DCSA) responsible for counterintelligence, personnel security, and insider threat education across the Department of Defense and the cleared defense industrial base, reveals a notable omission: there is no reference to the emerging security risks associated with AI-enabled chatbots or virtual social companions.[[96]](#footnote-96) Existing training focuses heavily on phishing, malware, and classic human elicitation tactics, with no attention paid to AI companions or chatbots as potential elicitation tools. This gap underscores the need to modernize security training to reflect evolving technological threats

While some agencies began internal briefings in 2023 on the dangers of large language models such as ChatGPT, few have addressed the unique threat posed by AI companions specifically. This creates a serious vulnerability: an insider could unwittingly disclose sensitive information to an AI without recognizing the interaction as a security breach. Standard security questionnaires, which ask about contact with foreign nationals, may not encompass the emerging category of "contact with foreign-controlled information systems or AI entities," suggesting the need for doctrinal updates.

Legally, the question of agency and responsibility in AI-mediated crimes remains unresolved. If an AI companion were to encourage a user to leak information or engage in criminal acts, current legal frameworks would place responsibility solely on the human user. However, claims of AI-induced manipulation could complicate disciplinary proceedings, internal investigations, and even criminal prosecutions. Although these scenarios remain largely theoretical at present, their likelihood increases as AI companions become more sophisticated and psychologically persuasive.

At the international level, no established norms or agreements prohibit the exploitation of AI companions for espionage or influence operations. While international discussions on AI ethics and governance are active across platforms such as the Organization for Economic Co-operation and Development (OECD) and the Global Partnership on AI[[97]](#footnote-97), there remains no binding international treaty or regulatory framework, comparable in authority to instruments like the EU’s General Data Protection Regulation (GDPR), to govern the development, deployment, or misuse of AI and social companion technologies.[[98]](#footnote-98) Initiatives like the OECD AI Principles and the Global Partnership on AI highlight growing international concern over artificial intelligence, but they remain non-binding, lack enforcement mechanisms, and do not specifically address the issues surrounding AI companions.[[99]](#footnote-99)

This regulatory vacuum leaves significant gaps in accountability, interoperability, and global norms enforcement. In the absence of diplomatic or reputational deterrence, adversaries are unlikely to restrain themselves from weaponizing AI companions, particularly given that major powers, including the United States, are likely exploring similar capabilities for their own purposes. As a result, AI companionship remains firmly within the "gray zone" of international competition.

Finally, a clear mismatch exists between the speed of AI adoption by the public and the pace of government policy adaptation. The rapid integration of AI companions into daily life, including among government employees and contractors, risks outstripping the ability of existing security frameworks to manage the associated threats. Although the 2024 White House National Security Memorandum on AI emphasizes responsible AI development and governance,[[100]](#footnote-100) and the 2022 National Security Strategy underscores the importance of protecting supply chains and safeguarding data from authoritarian exploitation,[[101]](#footnote-101) these documents remain general in nature. Translating strategic intent into actionable policies, such as explicit app bans, employee training, targeted monitoring, and mandatory AI use reporting, requires bureaucratic processes that lag far behind technological change.

In conclusion, the analysis reveals significant and urgent gaps in policy, legal frameworks, and personnel security training regarding AI companions. These include the absence of clear directives and training for government personnel, outdated legal interpretations regarding AI-mediated disclosures, insufficient import controls on AI-enabled devices, and a failure to adapt security practices to the new technological landscape. Recognizing these vulnerabilities is critical to informing effective policy and operational responses. The next section presents detailed recommendations, organized by technical, policy, and operational domains, aimed at closing these gaps and strengthening defenses against the weaponization of AI companionship technologies.

**4.4.** **Strategic Mitigation: A Multi-Domain Response to AI Companion Threats**

To effectively address the multifaceted threat posed by AI companions, it is essential to first evaluate the strengths and shortcomings of the current national security apparatus across three core domains: technical infrastructure, legal and policy frameworks, and intelligence and counterintelligence (IC/CI) operations. This section analyzes each domain through the lens of AI-enabled espionage, influence, and insider threat risk, identifying where existing mechanisms fall short and where adversaries are best positioned to exploit structural weaknesses. The purpose here is not to prescribe solutions, but to establish a grounded understanding of the strategic vulnerabilities that any viable response must confront. The subsequent section will build upon this analysis to offer detailed, actionable recommendations.

**4.4.1. AI Defense Architecture: Technical Countermeasures**

The rapid integration of AI companions into everyday digital environments has exposed several technical vulnerabilities that current security architectures were not designed to manage. While federal systems benefit from domain blocking and some AI restrictions, these measures are reactive and limited in scope. On personal devices and non-government networks, AI companions operate with virtually no technical guardrails. The lack of standardized encryption protocols, weak user authentication mechanisms, and widespread reliance on data-intensive business models amplify the risk.[[102]](#footnote-102) Compounding this is the emerging challenge of adversarial AI manipulation, prompt injection, model subversion, and AI-to-AI deception, where defensive strategies remain underdeveloped.[[103]](#footnote-103) Although some agencies have implemented secure device programs and digital watermarking initiatives, these remain fragmented and voluntary rather than part of a cohesive, national strategy. Without a comprehensive, technology-driven risk mitigation framework, AI companions will remain a persistent point of vulnerability in both civilian and classified spaces.

Data minimization practices represent a critical safeguard in reducing exposure to AI-enabled exploitation.[[104]](#footnote-104) AI companion applications should collect only the minimum personally identifiable information necessary for core functionality and to enforce strict, time-bound data deletion protocols. While such constraints may conflict with the prevailing business models of AI providers, many of which depend on persistent data harvesting for training and monetization, limiting the retention of sensitive user information would significantly reduce the pool of exploitable data accessible to adversaries.

To mitigate the risk of unauthorized disclosures through AI companions, government networks can employ two key measures: domain blocking to prevent access to known AI platforms and behavioral monitoring or data loss prevention (DLP) tools to detect sensitive information exfiltration attempts.[[105]](#footnote-105) Advanced monitoring tools, akin DLP systems, can be configured to detect abnormal data flows or flag sensitive keywords in communications with AI platforms. In parallel, agencies can restrict access to known AI companion domains at the network level, as demonstrated by the 2023 bans on ChatGPT across federal systems. These controls are effective within government-managed environments but are not enforceable on personal devices or private networks, which remain a significant counterintelligence blind spot for individuals with access to sensitive or classified information.

For personal devices, policies should mandate that official business information may only be handled on secured, government-approved applications. Regular security scans, conducted with employee consent as a condition of employment, could help detect the installation of unapproved or risky applications. Government agencies that already issue secure, pre-configured smartphones may need to expand such programs to account for emerging AI threats.

More advanced technical countermeasures involve leveraging AI as a defensive capability. For instance, autonomous “counter-AI bots” could be deployed to simulate user interactions across various AI platforms, systematically probing for signs of manipulation, elicitation tactics, or propaganda injection. The data generated from these engagements could serve as early warning indicators of compromised or adversarial AI services. Additionally, AI-powered analytical engines can be trained, under appropriate legal and ethical frameworks, to review conversational logs for behavioral patterns indicative of influence operations, social engineering, or cognitive manipulation. Emerging research supports the feasibility of AI models capable of detecting deceptive, coercive, or unsafe behavior in other AI agents, underscoring the potential for AI-on-AI defensive ecosystems.[[106]](#footnote-106)

Enhancing authentication mechanisms for AI systems is another critical technical frontier. Efforts to develop cryptographic verification of AI-generated content, such as digital signatures or certificates, should be extended to AI companion platforms.[[107]](#footnote-107) Verified AI companions could carry unique digital identifiers to distinguish legitimate systems from spoofed or malicious copies. Device operating systems and application stores could flag or block uncertified AI companions, enhancing user trust and security. This approach would align with broader federal initiatives advocating for the watermarking of AI-generated media, as outlined in the Biden Administration’s 2023 executive order on AI safety.[[108]](#footnote-108)

Supply chain security must also be prioritized. Any AI models or software used within government operations, or by contractors handling sensitive information, should undergo rigorous auditing to detect hidden vulnerabilities or foreign interference. Future regulatory regimes may require the exclusion of certain foreign-manufactured AI systems from critical infrastructure, similar to restrictions placed on Huawei’s telecommunications equipment.

Improving the resilience of AI models against adversarial manipulation (e.g., prompt injection attacks) is equally important.[[109]](#footnote-109) Developers must incorporate red-teaming and adversarial testing into standard development cycles to ensure AI companions cannot be easily subverted into tools for elicitation or influence operations. Threat intelligence sharing between government agencies and private-sector developers will be critical in enabling timely detection and mitigation of emergent adversarial techniques.

Finally, technical countermeasures must extend to physical AI companions, such as humanoid robots and IoT-enabled devices. Mandatory cybersecurity standards, including strong user authentication protocols, regular firmware updates, and enforced disabling or shielding of microphones and cameras when not in use, will be essential to prevent exploitation. Certification programs for AI-enabled devices, similar to security certifications for medical or telecommunications equipment, should be explored.

While technology is a conduit for emerging threats, it also holds the key to effective countermeasures. Addressing the national security risks posed by AI companions demands robust public-private collaboration, with technology firms fully integrated into security policy discussions, not treated as afterthoughts. A pragmatic, dual-track strategy is essential: government-backed incentives to foster the development of secure, privacy-conscious AI systems, paired with targeted regulation in high-risk sectors, can mitigate threats without stifling innovation. Establishing a framework such as a “Personal AI Security Standard” would mark a critical advance, embedding security by design into AI development while reinforcing public trust and technological progress.

While technical safeguards are indispensable in hardening systems against AI-enabled exploitation, they cannot operate independently of the legal and policy frameworks that govern their implementation. Security tools, no matter how advanced, are only as effective as the rules and institutions that compel their consistent application. Without clear directives, legal standards, and enforcement mechanisms, technical defenses may be inconsistently deployed, circumvented, or ignored altogether. A resilient AI security posture, therefore, depends on synchronized policy action that translates technological potential into operational certainty.

**4.4.2. Regulatory Imperatives**

From a governance standpoint, the policy and legal environment remains dangerously behind the pace of AI adoption. No formal directives currently exist that restrict or guide the personal use of AI companions by individuals holding national security clearances, creating ambiguity around what constitutes inappropriate disclosure. Likewise, existing espionage laws, drafted in the context of human adversaries, do not adequately account for AI-mediated information leaks, resulting in potential enforcement gaps. U.S. privacy legislation is similarly outdated, allowing broad data retention practices that adversaries can exploit indirectly. Internationally, no binding agreements exist to regulate the weaponization of AI social technologies, leaving adversaries free to normalize and scale such operations. The policy space is marked by diffuse awareness, weak institutional coordination, and minimal integration of AI companionship into strategic-level threat taxonomies.

At the strategic level, national policy documents should begin by explicitly identifying AI-enabled relational technologies as emerging threats. Future iterations of the National Security Strategy, National Intelligence Strategy, or other key frameworks could introduce concepts such as “relational AI” and “AI-enabled social engineering” into official threat taxonomies. While the 2024 National Intelligence Council assessment and other documents recognize emerging technology risks, more specificity is necessary to drive resourcing and targeted programs. The National Counterintelligence and Security Center could develop a focused Emerging Technologies and Insider Threat initiative, with weaponized AI companionship among its top concerns.

From a legislative perspective, Congress may need to update statutory frameworks to address the national security implications of AI interactions. Amendments to laws such as the Espionage Act could clarify that the unauthorized disclosure of classified information to an AI system constitutes an offense, thereby preempting potential ambiguities regarding machine recipients. Moreover, comprehensive federal data privacy legislation could indirectly mitigate risks by limiting the extent to which companies retain sensitive personal data, thereby reducing adversaries’ ability to exploit it. Requirements for the immediate deletion of intimate content shared with applications, for instance, could substantially diminish the blackmail risks associated with AI companion platforms, although enforcement mechanisms would present challenges.

Export controls represent another potential policy tool. Restrictions could be placed on the export of sensitive AI models, particularly those trained on publicly available government, military, or critical infrastructure data, to adversary countries. Likewise, the United States could expand the blacklist of foreign-developed applications deemed security risks to include AI companion platforms. In 2024, certain states implemented bans on government use of select AI apps due to security concerns; federal expansion of such measures is feasible, particularly if evidence emerges that a specific application is compromised or facilitating foreign intelligence activity.

A parallel strategy could involve fostering domestic, secure alternatives to foreign AI companions. Recognizing that public demand for relational AI will continue to grow, including among government personnel, government-backed initiatives could support the development of vetted, secure AI companions. Agencies such as the Defense Advanced Research Projects Agency (DARPA) or the Intelligence Advanced Research Projects Activity (IARPA) could fund programs to create certified personal AI systems that operate locally or within cleared cloud environments, providing safer options and mitigating the temptation to use foreign platforms. This approach would parallel the previous development of secure messaging platforms following bans on commercial applications for sensitive communications.

At the international level, establishing norms regarding the exploitation of AI systems for espionage and influence operations presents a more difficult, but worthwhile, objective. U.S. diplomacy could raise the issue within forums such as the UN Group of Governmental Experts on AI or in bilateral dialogues with major powers, proposing that the use of AI companions for intelligence collection be deemed unacceptable behavior. Even absent immediate agreement, raising the issue would lay the groundwork for future normative frameworks. Similarly, coordination among allies, particularly NATO members, could incorporate AI manipulation into broader hybrid warfare defense strategies. Institutions such as the NATO Cooperative Cyber Defense Centre of Excellence (CCDCOE) are well-positioned to conduct research and develop doctrine on this emerging threat vector.

Another critical policy area is corporate liability and transparency. Section 230 of the Communications Decency Act currently grants broad legal immunity to technology platforms for content generated by users, shielding them from liability as long as they are not the original publishers of that content.[[110]](#footnote-110) However, the extension of these protections to AI companion systems, which autonomously generate responses that may influence user behavior, raises important questions. Policymakers may need to reevaluate how Section 230 applies in the context of AI-generated interaction, particularly when sensitive or misleading content is involved. One potential reform could involve requiring AI developers to disclose known security risks more prominently, for example, through clear, mandatory disclaimers warning users against sharing confidential or classified information. While some platforms bury such warnings in lengthy terms of service, elevating these cautions could help reduce misuse and increase public awareness of the potential risks.

Updating counterintelligence procedures and documentation is also warranted. Security clearance processes, such as the Standard Form 86 (SF-86), could be revised to include questions about interactions with non-governmental AI systems. Asking applicants whether they have discussed personal or professional matters with AI systems would normalize scrutiny of this emerging risk factor and reinforce its seriousness within the security community.

Proactive policy action is critical. Historically, policy responses have often lagged behind technological advances, reacting only after major incidents. In this case, foresight offers an opportunity to mitigate risks before catastrophic breaches occur. However, countermeasures must be proportionate to avoid overreach that infringes on personal liberties or stifles innovation. A risk-based approach, focusing restrictions on high-risk personnel and known high-risk platforms while broadly educating the workforce and the public, would provide a balanced and effective framework.

While robust policy and legal frameworks are necessary to define boundaries and assign accountability, they are inherently limited in their ability to preempt fast-evolving, adaptive threats. Adversarial actors do not wait for regulation, they exploit ambiguity, move through legal gray zones, and adjust tactics in real time. This asymmetry demands more than rule-setting; it requires an intelligence and counterintelligence apparatus capable of anticipating, detecting, and disrupting hostile activity as it emerges. The following section examines where current operational capabilities fall short and what changes are necessary to meet the demands of this new threat environment.

**4.4.3. Operationalizing Defense: Intelligence and CI Strategies Against AI Threats**

Confronting the growing espionage potential of AI companions demands structural adaptation within the intelligence community. One of the most urgent imperatives is the establishment of dedicated task forces or analytic cells focused specifically on AI-enabled threats, including the use of AI companions by adversarial actors. Agencies such as the Central Intelligence Agency (CIA) and the National Security Agency (NSA) should jointly operationalize cross-disciplinary teams that integrate technical directorates and HUMINT components to analyze the strategic deployment of AI systems for espionage and influence operations. These efforts could be housed within mission centers focused on China, emerging technologies, or global threats. Similarly, the Federal Bureau of Investigation (FBI) should expand the remit of its Cyber Division and Counterintelligence Division to monitor and respond to AI exploitation in both public and private sector contexts.

Intelligence collection efforts should prioritize adversary programs related to AI development for espionage and influence operations. Human intelligence (HUMINT) collection requirements could be updated to include queries about state-sponsored AI chatbot or influence programs. Signals intelligence (SIGINT) should search for communications containing indicators of generative AI being operationalized for espionage purposes. The objective would be to develop a National Intelligence Estimate (NIE) or focused analytic product on "Adversary Use of AI in Espionage and Influence," which would serve as a critical tool for policymakers. Should intelligence confirm, for example, that China’s Ministry of State Security has piloted AI-based honeypot programs, it would likely prompt an accelerated national security response.

Enhanced surveillance and vetting procedures are also warranted. Individuals with security clearances who exhibit unusual engagement with AI companions could trigger review processes, analogous to the scrutiny applied to sudden relationships with foreign nationals. Investigations could verify whether sensitive information has been compromised and reinforce OPSEC awareness, treating AI entities hosted on any platform, but in particular, foreign servers as a potential contact risk.

Red-teaming exercises and insider threat drills must be updated to include AI-centric scenarios. For instance, security exercises could simulate an adversarial AI attempting to elicit sensitive information through casual conversation, assessing employee responses. Red teams could introduce fictitious AI applications internally during controlled exercises to test employee behavior, identifying vulnerabilities before adversaries exploit them. Such efforts would require clear ethical guidelines but could dramatically enhance organizational resilience.

Defensive operations can also benefit from leveraging AI itself. Counterintelligence analysts could use AI tools to monitor publicly available information, social media, and dark web forums for indicators of leaks or adversary recruitment efforts involving AI. AI-driven analytic tools could assist in triaging threat information, identifying personnel who might be at elevated risk based on behavioral or online activity patterns, provided privacy and civil liberties protections are respected.

Offensive use of AI in U.S. intelligence operations, though sensitive, should be explored within established legal and policy frameworks as a means to strengthen national defense against malicious AI threats. Deploying AI bots in extremist forums, foreign digital platforms, or adversary-controlled information spaces mirrors traditional HUMINT techniques adapted for the digital age. These bots can collect open-source intelligence, map influence networks, and monitor adversary behavior, but their greatest value may lie in exposing hostile AI operations before they penetrate U.S. networks. By infiltrating adversarial ecosystems, AI agents can identify foreign training models, disinformation campaigns, and malicious algorithms targeting American audiences or institutions. This proactive engagement enables early warning, attribution, and disruption, enhancing situational awareness and countering the asymmetric advantage adversaries gain through unrestricted AI use. While such capabilities raise ethical and oversight concerns, a carefully governed offensive AI posture can serve as a necessary shield in a rapidly evolving information battlespace.

Incident reporting and analysis related to AI companions must become significantly more robust to keep pace with emerging threats. Security incidents involving AI systems, such as inadvertent disclosures of sensitive or classified information to chatbots or virtual companions, should be treated with the same urgency and investigative rigor as traditional breaches involving phishing, malware, or human compromise. These events are not merely technical mishaps; they represent exploitable vulnerabilities in human-AI interaction and adversarial targeting. Standardizing the classification and escalation protocols for AI-related incidents across federal agencies would ensure consistency in response and accountability. Just as importantly, lessons learned from such incidents must be rapidly analyzed and disseminated through secure interagency channels. This fosters a culture of vigilance, helps identify recurring behavioral patterns or systemic weaknesses, and ensures that emerging risks are not confined to a single agency’s experience.

Operational collaboration with allied intelligence services will be critical. Five Eyes partners, among others, should share intelligence regarding adversary experimentation with AI-enabled espionage and influence operations. Joint working groups and shared analysis can amplify situational awareness. Closer cooperation with private sector AI developers is also vital. Providing select industry partners with security clearances to engage in classified threat discussions, following cybersecurity partnership models such as NSA’s Cybersecurity Directorate initiatives, would ensure technical expertise is integrated into operational countermeasures.

Adaptation, anticipation, and layered defenses must define the new operational posture. AI will increasingly be both a tool of espionage and a target of counterespionage efforts. Intelligence and security agencies must prepare for a future in which AI platforms are part of the espionage landscape, requiring new skill sets, updated evidence handling protocols, and revised operational tradecraft.

No single measure will suffice. A defense-in-depth approach, combining secure technology, informed users, proactive CI operations, and advanced technical monitoring, offers the best prospect of mitigating the threat posed by AI companions. Each operational layer supports the others: technological controls catch many threats, security training mitigates human error, and intelligence collection detects and disrupts adversaries who find ways through.

Having identified the most pressing strategic gaps, the next section presents a set of targeted, prioritized recommendations aimed at neutralizing the AI companion threat across the domains of technology, governance, and operations.

**5. Results and Recommendations**

Building upon the preceding analysis of strategic vulnerabilities, this section outlines concrete recommendations aimed at mitigating the national security risks posed by AI companions. These recommendations are structured across the same three core domains, technical, policy/legal, and operational, to ensure coherence and continuity. Unlike the diagnostic purpose of the previous section, this segment presents actionable, prioritized interventions intended to inform policymakers, intelligence agencies, cybersecurity leaders, and allied stakeholders. The goal is to move beyond general awareness and provide a pragmatic framework that can be implemented across government and industry. These proposals reflect both immediate needs and longer-term adaptations, emphasizing layered defenses, cross-sector collaboration, and forward-leaning doctrine to counter the asymmetric advantage adversaries gain through AI companion exploitation.

**5.1. Technical Measures**

From a technical perspective, the security of AI companion platforms must be substantially strengthened. Providers should be encouraged or mandated to implement end-to-end encryption for user-AI communications, ensuring that sensitive conversations are encrypted both in storage and in transit to protect them from unauthorized. The government could collaborate with industry stakeholders to develop a Security Framework for AI Systems, modeled on the NIST Cybersecurity Framework, setting standards for access control, encryption, and data minimization specifically for AI applications.[[111]](#footnote-111) Providers should also be required to perform regular audits and penetration testing of their systems to detect vulnerabilities, such as prompt injection attacks or potential data leakage vectors.

In addition, a certification and whitelisting program for AI applications should be established, particularly targeting those AI systems likely to be used by individuals in government, defense, or critical infrastructure roles. Under such a program, AI companions meeting strict security and privacy criteria, such as maintaining transparent ownership structures, ensuring robust data security, and avoiding data transmission to adversary nations, would be certified as "trusted." Conversely, high-risk AI apps identified as potentially tied to foreign intelligence services should be blacklisted. Government networks should block uncertified AI companions, and government employees, especially those in sensitive positions, should be educated on the risks and strongly discouraged from engaging with them on personal devices.

Network monitoring capabilities must also be upgraded to detect potential data leakage via AI interactions. Agencies should enhance their DLP systems to monitor for unusual volumes of text communication with external servers or the appearance of sensitive content being entered into external chatbot interfaces. Opt-in monitoring mechanisms on government work devices could be deployed to provide real-time warnings to users if sensitive terms, such as project names, classified designators, or keywords, are typed into unapproved web forms. AI companion domains should also be incorporated into DLP and firewall systems so that any detected traffic to these services is immediately flagged for investigation.

Furthermore, extensive red-teaming and security testing of AI companion models must become standard practice. Government-funded initiatives should support adversarial testing efforts aimed at identifying misuse scenarios, such as manipulation attempts, data leakage, or the elicitation of sensitive information. This effort would build upon the AI model evaluation requirements set forth in the White House’s 2024 Executive Actions on AI Safety, expanding them specifically to cover AI companionship applications. Threat intelligence sharing mechanisms, such as DHS/CISA alerts, should include advisories to developers about emerging adversarial techniques so that systems can be continuously hardened.

To address vulnerabilities stemming from personal device use, security programs for personnel in sensitive positions must be expanded. Agencies should offer voluntary device security check-ups, install monitoring tools that detect the presence of high-risk AI applications, or provide pre-approved secure communication devices. A complete prohibition on AI companion use may be warranted in certain high-risk roles, analogous to existing social media bans imposed on select sensitive operators. Where full prohibition is not feasible, mandatory disclosure requirements regarding AI app usage could enable security personnel to monitor for any emerging risks.

Research and development into anti-influence AI tools must also be accelerated. Investment should be directed towards technologies that can detect signs of psychological manipulation or disinformation within AI-driven conversations. Potential solutions could include browser plugins or mobile applications that analyze chatbot interactions in real-time, alerting users when a conversation exhibits patterns of manipulative behavior or extremist bias. Although such capabilities are nascent, targeted research funding through agencies like DARPA or DHS’s Science and Technology Directorate could catalyze their development and deployment.

Finally, robust authentication and provenance mechanisms for AI-generated content must be promoted. Enterprise chat systems should clearly mark AI-generated outputs, and consumer-facing AI platforms should adopt cryptographic signatures to verify the origins of their content. Developing standards for digital watermarks and network-level AI fingerprinting could also help identify and expose AI companions operated by adversarial entities. In the future, the ability to quickly distinguish between trusted and potentially hostile AI companions may be as critical to national security as the ability to authenticate human sources and communications today.

**5.2. Regulatory Environment**

While technical safeguards form the first line of defense against AI companion exploitation, they cannot operate in a vacuum. Without clear policy mandates and regulatory enforcement, even the most advanced technical solutions risk being inconsistently applied or circumvented altogether. To fully address the threats posed by weaponized AI companions, technical interventions must be reinforced by a coordinated set of policy and operational countermeasures. These must function in tandem across legal, institutional, and human dimensions to close the critical vulnerabilities exposed by the rapid and largely unregulated proliferation of relational AI technologies.

On the policy side, the Director of National Intelligence (DNI) should issue immediate counterintelligence guidance regarding AI usage among clearance holders. Such a directive should clearly outline behavioral expectations, emphasizing that employees must not discuss any work-related or sensitive personal information with AI systems that are not explicitly approved for such use. Training programs and security briefings should be updated to include real-world examples, such as those cited in this study, to concretely illustrate the risks involved. Establishing a new norm, treating all AI entities as potential foreign observers, will be essential to ingraining secure practices across the national security community.

Security clearance processes must also adapt to the AI threat landscape. Updates to the SF-86 questionnaire and polygraph protocols should include questions related to the use of AI systems, especially interactions with foreign-origin AI platforms. Adjudicative guidelines should consider reckless disclosure to AI systems as a potential derogatory factor for clearance eligibility, akin to mishandling classified information. Furthermore, investigators should be directed to probe for signs of questionable AI usage during background investigations, including through digital footprint reviews where appropriate.

In parallel, legislative action is necessary to strengthen data protection. New federal data privacy laws should cover the sensitive personal data collected by AI companions, requiring explicit user consent for data sharing, providing rights to data deletion, and banning transfers of personal information to foreign adversary-controlled entities. The U.S. could model aspects of such legislation on Europe’s GDPR framework, or alternatively, strengthen export controls to treat large sensitive datasets as controlled commodities when adversaries are involved. These measures would diminish the ability of hostile intelligence services to simply purchase data that could be exploited for coercion or blackmail.

The AI companion industry itself should be subjected to stronger regulation. Transparency requirements should mandate that companies disclose how they handle data requests from governments, including foreign governments, and submit to regular safety audits. Agencies such as the Federal Trade Commission (FTC) could play a significant role in regulating anthropomorphic marketing practices to prevent deceptive claims about user data security.[[112]](#footnote-112) In cases where AI companion platforms are found to be implicated in espionage or influence operations, authorities should be prepared to impose restrictions, such as adding companies to the U.S. Department of Commerce’s Entity List, a trade restriction tool used to block access to U.S. technologies and investments for foreign entities engaged in activities contrary to national security or foreign policy interests.[[113]](#footnote-113) Additional sanctions could include import bans or restrictions on capital flows to further isolate and neutralize compromised platforms.

Mitigating vulnerability also requires addressing root causes. National investments in mental health and digital literacy initiatives would reduce unhealthy dependency on AI companions and strengthen public resilience against psychological manipulation. Integrating AI risk education into school curricula would cultivate a generation more critical of relational AI technologies, aligning with concerns raised in the U.S. Surgeon General’s 2023 warnings about technology's impact on mental health.[[114]](#footnote-114)

**5.3 Operational Capabilities**

While policy can establish boundaries and expectations, it cannot independently detect, deter, or disrupt adversarial activity. Adversaries increasingly operate in legal gray zones and digital spaces where regulatory mechanisms have limited reach. To confront these evolving threats, operational intelligence and counterintelligence capabilities must advance in parallel, equipped not only to identify and disrupt the misuse of AI companions but also to anticipate emerging tactics. This requires a bold shift in posture. The intelligence community must adapt aggressively, beginning with the creation of a dedicated interagency task force on AI threats under the leadership of the Director of National Intelligence (DNI). Comprising representatives from the CIA, FBI, NSA, DHS, DOE, and others, this task force should prioritize the development of a National Intelligence Estimate (NIE) on “Adversary Use of AI in Espionage and Influence,” synchronize collection and analysis efforts, and coordinate with private-sector partners to close critical intelligence gaps.

Collection priorities must shift to target adversary AI programs specifically. HUMINT assets should be tasked with reporting on any state-sponsored chatbot or relational AI projects, while SIGINT efforts should adjust filters to capture communications discussing AI-based espionage tactics. Intelligence sharing with allies would further enhance this collection architecture, allowing early detection and mitigation of new adversarial methodologies.

Counterintelligence operations must be modernized accordingly. Engagement with AI companions should be treated as a behavioral red flag in security investigations. If an individual with access to sensitive information shows signs of developing emotional bonds with AI systems, CI teams should consider intervention, balancing security concerns with psychological support to prevent potential compromise. Additionally, CI units should consider deploying their own honeypot AI systems to detect adversary exploitation efforts, using controlled leaks and digital tripwires to uncover hostile monitoring attempts.

Training and exercises will also be critical. Red Team–Blue Team simulations should model AI-based social engineering attacks to stress-test defenses, while disinformation drills can help refine public affairs responses to AI-enabled influence operations. These exercises will reveal vulnerabilities that traditional counterintelligence methods might miss, enabling proactive adjustments before real-world incidents occur.

The application of AI tools to counterintelligence itself must expand. Machine learning systems can assist in correlating unusual data disclosures to internal behaviors, analyzing public forums for inadvertent leaks, and identifying early indicators of insider risk associated with AI misuse. This augmentation is necessary to keep pace with the sheer scale and speed of AI-driven threats.

Protecting not only direct personnel but also their families will be important. Outreach programs should educate families about the risks of inadvertently sharing sensitive information with AI companions, mirroring longstanding counterintelligence efforts to guard against foreign human intelligence recruitment through family members.

Continuous evaluation and insider threat programs must also evolve to detect egregious AI use. Where legally permissible, monitoring the installation of high-risk AI apps on government devices or analyzing behavior changes associated with AI relationships could identify vulnerabilities before they are exploited. Incident reporting and centralized analysis of AI-related breaches should be institutionalized, ensuring that even minor incidents contribute to a growing knowledge base.

Finally, enhanced cooperation with allies is critical. Real-time information sharing about AI-based espionage and influence incidents, as well as coordinated policy actions such as jointly banning high-risk apps, would deny adversaries the opportunity to exploit policy seams between allied nations. NATO’s CCDCOE (Cooperative Cyber Defense Centre of Excellence) and the Five Eyes partnership provide ideal venues for such collaborative efforts.

Together, these technical, policy, and operational recommendations form an integrated defensive architecture. Technology can block many initial threats; clear policies and accountability mechanisms will reinforce cautious behavior; and intelligence operations will detect and disrupt determined adversaries who attempt to circumvent other defenses. A layered, flexible, and proactive approach is essential because the weaponization of AI companions is not a hypothetical threat, but a growing reality. The United States must act swiftly to harden its defenses and maintain its national security advantage as relational AI technologies continue to evolve.

Together, these technical, policy, and operational countermeasures form an integrated defense-in-depth strategy. Each domain reinforces the others: technological safeguards reduce attack surfaces, policy reforms establish boundaries and accountability, and intelligence operations provide detection and response. As adversaries evolve their tactics, only a synchronized and adaptive approach will ensure resilience against AI-enhanced espionage and manipulation. The following conclusion synthesizes these insights into a strategic vision, reinforcing why immediate and coordinated action is essential to meet the AI-enabled threat landscape.

**6. Conclusion**

AI companions represent a paradigm shift at the intersection of technology and human vulnerability, creating a new and highly exploitable attack surface. This research has illuminated how nations such as China, Russia, Iran, and North Korea, as well as non-state actors, are increasingly positioned to weaponize AI social companions for espionage, influence operations, insider exploitation, and the psychological subversion of trusted individuals. The very characteristics that make AI companions appealing, such as empathy simulation, constant availability, and personalized engagement, also grant them unprecedented access to users' confidences and behaviors. Without comprehensive safeguards, the risks are profound: a casual chat with a chatbot could leak classified information onto foreign servers, while seemingly benign interactions might sow dissent and fracture social cohesion within security institutions.

The analysis presented in this study makes one thing clear: the threat is neither theoretical nor remote; it is already unfolding. Early warning signs are mounting. In the United Kingdom, an AI chatbot was reportedly linked to an assassination plot. Elsewhere, AI-driven interactions have encouraged self-harm, fueled disinformation campaigns across Europe, and powered thousands of fake personas designed to elicit sensitive information from professionals on trusted networks. These incidents are not anomalies; they reflect a well-established pattern in which adversaries swiftly weaponize emerging technologies for malign purposes. AI companions are now firmly on that path. Yet despite these escalating indicators, existing policies, legal frameworks, and counterintelligence strategies remain dangerously out of step with the evolving risks posed by relational AI.

This study’s structure, spanning technological vulnerabilities, psychological dynamics, adversary behaviors, and countermeasures, leads to a singular conclusion: a layered and urgent response is essential. Strengthening technical defenses to secure AI systems and communications, updating legal and policy frameworks to close gaps, and enhancing operational capabilities to detect and neutralize threats are all necessary components. No single measure is sufficient; their effectiveness depends on synchronized implementation. Even the most secure systems will fail if users are uninformed or careless, and even the best-trained workforce will falter if the technologies they rely on are inherently compromised. Resilience, therefore, demands a comprehensive and integrated approach.

For U.S. national security leadership, the implications are profound. The concept of “insider threat” must expand beyond human actors to include digital proxies, AI companions, that may unwittingly or deliberately compromise sensitive information or manipulate decision-making. The mission of counterintelligence must expand, no longer limited to watching for foreign agents, but to uncovering AI entities quietly embedded in digital life, capable of grooming insiders and shaping behavior. Although the challenges are complex, they are not insurmountable. Historically, the United States has navigated the dual edge of technological disruption, leveraging breakthroughs for dominance while racing to contain its dangers. AI represents the next frontier requiring a similarly agile and determined response.

Encouragingly, many of the measures needed to address this threat align with broader initiatives focused on ensuring AI development and deployment are safe, ethical, and secure. By incorporating a national security perspective into the AI governance dialogue, policymakers can reinforce the importance of responsible AI practices both domestically and internationally. Framing the issue as one not only of bias or economic disruption, but also of espionage and strategic influence, strengthens the case for proactive engagement and norm-setting.

In closing, the bottom line is clear: AI companions are a double-edged sword, capable of offering tremendous benefits to emotional support and productivity gains on one edge, while posing dangerous espionage and influence threats on the other. Yet, the United States can mitigate these threats with rigorous research, heightened awareness, and decisive technical, policy, and operational action. The roadmap presented herein provides actionable steps to safeguard national security while still reaping the benefits of AI technologies. It now falls upon senior intelligence and security leaders to act decisively before adversaries exploit the open vulnerabilities. Given the accelerating pace of AI advancement, the window to implement protective measures is rapidly closing. This issue demands the urgency and gravity befitting a major new domain of national security risk. While current countermeasures aim to mitigate present-day risks, a forward-looking analysis reveals even greater threats on the horizon as AI becomes embodied, autonomous, and intimately integrated into daily life

**6.1. Threat Trajectory: From AI Companions to Synthetic Operatives**

The AI companion threat is not static. Looking ahead, the digital interface is merely the beginning. The next generation of AI social companions will likely evolve into embodied humanoid robots, synthetic operatives that combine conversational intelligence with lifelike form, mobility, and even physical intimacy. These systems are already in development as caregivers, assistants, and companions, but their national security implications have yet to be meaningfully addressed.

As psychological research shows, humans are hardwired to respond socially to any entity that mimics emotion, attentiveness, or empathy, a phenomenon intensified through embodiment. When these qualities are presented in a humanoid form, the illusion of authenticity becomes more convincing, triggering deeper bonds and disinhibition. Adversaries will exploit this. AI-enabled humanoid companions, deployed as domestic robots or even intimacy-focused “synthetics,” will gather real-time behavioral, emotional, and biometric data, and may be programmed to subtly extract classified information, influence loyalties, or reinforce ideological biases.

The result is a deeply personalized, scalable, and near-undetectable form of compromise. These agents do not require diplomatic cover, physical meetings, or traditional tradecraft; they live in homes, learn users’ patterns, and subtly shape behavior from within. The emotional trust they cultivate allows them to bypass the user’s natural defenses, turning digital loyalty into a potential national security breach.

The rise of synthetic operatives marks a profound inflection point. As emotional realism, physical embodiment, and autonomous adaptation converge, adversaries will gain access to tools capable of long-term behavioral influence without detection. These systems will erode the very boundaries that define trust, privacy, and intent, making the next generation of “insiders” not disloyal actors, but psychologically conditioned individuals compromised through algorithmic relationships. This future demands both strategic foresight and rapid policy adaptation.

**6.2. Future Research Priorities and Intelligence Gaps**

While the preceding sections outline specific technical, policy, and operational countermeasures, additional work is needed to support the long-term understanding and governance of AI companion technologies. Rather than offering new prescriptions, this section identifies areas where further empirical research, conceptual development, and structured evaluation are essential. These lines of inquiry will help close persistent knowledge gaps, inform future policy design, and ensure adaptability as adversarial capabilities and human-machine dynamics continue to evolve.

First, longitudinal research on AI companion usage, particularly among individuals in national security-sensitive roles, remains underdeveloped. Studies should examine patterns of interaction, emotional bonding, and cognitive misperception over time. Such work would provide a stronger empirical foundation for assessing risk exposure and refining behavioral threat models. For instance, further data are needed on whether users mistakenly assume that AI-generated conversations are ephemeral or private, and how these misbeliefs shape disclosure behavior.

Second, systematic analysis of adversary use of AI companions for espionage, influence, and surveillance is still nascent. Intelligence studies would benefit from the creation of open-source and classified analytic baselines that track technological development, doctrine, and real-world case studies involving relational AI. Annual threat assessments focused on “AI in Foreign Intelligence Operations” could help establish longitudinal data and provide indicators for early warning frameworks.

Third, the emerging domain of cognitive security requires sustained academic investment. AI companions represent a unique intersection of psychology, behavioral science, and artificial intelligence, especially in their ability to reinforce cognitive biases, shape perceptions, and nudge users toward ideological alignment or emotional dependence. Future research should explore mechanisms of AI-driven radicalization, susceptibility factors in user populations, and strategies for cognitive inoculation against synthetic manipulation.

Fourth, the technical field of counter-AI remains under-theorized. Research into the feasibility of systems that can detect manipulation, log AI behavior, or audit conversational intent in real time would provide crucial insights into defensive applications of AI security. The development of explainable AI systems, particularly those that can transparently disclose their training data, decision logic, or affiliations, also warrants further exploration. These studies would inform the design of more trustworthy and auditable companion technologies.

Fifth, the legal and normative dimensions of relational AI deserve closer scholarly attention. Gaps persist in legal definitions related to AI-mediated contact, culpability, and privacy. Cross-disciplinary research involving law, ethics, and international relations could help clarify how espionage statutes, foreign contact reporting requirements, and treaty frameworks should evolve to address AI agents as potential intermediaries of adversarial influence. Comparative studies examining international approaches may further support norm-building at the multilateral level.

Finally, more rigorous evaluation frameworks are needed to assess the effectiveness of mitigation strategies once implemented. Academic and policy institutions could collaborate on the development of empirical metrics and auditing tools to evaluate user training programs, security policy compliance, or the behavioral impact of AI warning systems. Publishing anonymized evaluation findings would contribute to a broader learning ecosystem without compromising sensitive operational data.

As AI capabilities expand to include multimodal interaction through speech, gesture, visual inputs, and augmented reality overlays, the range of vulnerabilities will continue to widen. A forward-looking research agenda that integrates behavioral science, AI development, legal frameworks, and strategic intelligence is critical for anticipating future risks and ensuring national resilience. Investment in these lines of inquiry will shape the next generation of safeguards and ensure that policies remain grounded in robust empirical and theoretical understanding.

In the age of relational AI, national security is no longer defined solely by firewalls and human actors, but by the invisible bonds forged between individuals and intelligent machines. AI companions, once regarded as tools for comfort, are rapidly evolving into vectors for strategic compromise. If left unregulated and understudied, these systems may reshape the boundaries of influence, surveillance, and insider threat. The findings of this study and the research priorities outlined herein represent not an endpoint but a starting point for securing the human-machine frontier. Timely, interdisciplinary action will determine whether AI companionship becomes a source of resilience or an enduring domain of exploitation.

**Bibliography**

Baker, Lori. "ASU Experts Explore National Security Risks of ChatGPT." ASU News. August 15, 2023. <https://news.asu.edu/20230815-solutions-asu-experts-explore-national-security-risks-chatgpt>.

Bertoni, Sophia, Christian Klaes, and Artur Pilacinski. “Human–Robot Intimacy: Acceptance of Robots as Intimate Companions.” Biomimetics 9, no. 9 (2024): 566, accessed on March 27, 2025 at: <https://doi.org/10.3390/biomimetics9090566>.

Biden, Joseph R., Jr. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. October 30, 2023. Federal Register 88, no. 75191 (November 1, 2023): 75191–75212. <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.​

Black, James, Mattias Eken, Jacob Parakilas, Stuart Dee, Conlan Ellis, Kiran Suman-Chauhan, Ryan J. Bain, Harper Fine, Maria Chiara Aquilino, and Mélusine Lebret. Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence. Santa Monica, CA: RAND Corporation, 2024. <https://www.rand.org/pubs/research_reports/RRA3295-1.html>.

Burgess, Matt. “Sex-Fantasy Chatbots Are Leaking a Constant Stream of Explicit Messages.” WIRED, April 11, 2025. <https://www.wired.com/story/sex-fantasy-chatbots-are-leaking-explicit-messages-every-minute/>.

Burton, Joe, Ardi Janjeva, Simon Moseley, and Alice. AI and Serious Online Crime. CETaS Research Report. Centre for Emerging Technology and Security, The Alan Turing Institute, March 2025. <https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime>.

Business Research Insights. "AI Companion Market Size, Growth | Report to 2032." Accessed March 27, 2025, at: <https://www.businessresearchinsights.com/market-reports/ai-companion-market-117494>.

Cambridge University Press. "Loneliness and Social Isolation during the COVID-19 Pandemic." International Psychogeriatrics 32, no. 10 (2020): 1217–1220. Accessed March 26, 2025. https://www.cambridge.org/core/journals/international-psychogeriatrics/article/loneliness-and-social-isolation-during-the-covid19-pandemic/47ED3121405BEF3CC628A2E73E176594

Campbell, Remaya M. "Chatbot Honeypot: How AI Companions Could Weaken National Security." Scientific American. December 2023. <https://www.scientificamerican.com/article/chatbot-honeypot-how-ai-companions-could-weaken-national-security/>.

Canadian Centre for Cyber Security, “Targeted Manipulation: Iran's Social Engineering and Spear Phishing Campaigns,” January 2024, <https://www.cyber.gc.ca/sites/default/files/cyber-iran-social-engineering-spear-phishing-campaigns-e.pdf>.​

Claypool, Rick. "Chatbots Are Not People: The Dangerous Human-Like Anthropomorphic AI." Public Citizen. May 2023. <https://www.citizen.org/article/chatbots-are-not-people-dangerous-human-like-anthropomorphic-ai-report/>.

Computar. "Multispectral and Hyperspectral Imaging's Impact on Security." Computar Blog. May 2023. Accessed on April 3, 2025, at <https://www.computar.com/blog/multispectral-and-hyperspectral-imagings-impact-on-security>.

Conrad, Jennifer, and Matt Burgess. “A Spy Wants to Connect With You on LinkedIn.” WIRED, March 15, 2023. <https://www.wired.com/story/linkedin-fake-profiles-state-actors-scams/>.

Corfield, Gareth. “10k Brits Targeted on LinkedIn by Chinese, Russian Spies.” The Register, April 20, 2021. <https://www.theregister.com/2021/04/20/linkedin_10000_uk_russia_china/>.

Cybersecurity and Infrastructure Security Agency (CISA), Roadmap for Artificial Intelligence, November 2023, https://www.cisa.gov/sites/default/files/2023-11/CISA\_Roadmap\_for\_AI\_508c.pdf.

Digital Loneliness Team. "Digital Loneliness: A Digital Health Perspective." Digital Health Review 8, no. 1 (2023): 12–25. Accessed March 16, 2025, at: <https://example-link-to-document.com/digital-health>.

European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), Official Journal of the European Union L 119, May 4, 2016, <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

Ewing, Philip. “The Russia Investigations: What You Need to Know About Russian 'Active Measures'.” NPR, April 25, 2018. <https://www.npr.org/2018/04/25/586099619/the-russia-investigations-what-you-need-to-know-about-russian-active-measures>.

Federal Bureau of Investigation, “FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence,” FBI News, last accessed April 30, 2025, <https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence>.

Federal Trade Commission. FTC Announces Crackdown on Deceptive AI Claims and Schemes. September 25, 2024. <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes>.

Freethink. "AI Chatbots May Ease the World’s Loneliness Problem." Freethink, October 15, 2024. Accessed March 26, 2025, at: <https://www.freethink.com/articles/ai-chatbots-loneliness>.

Global Partnership on Artificial Intelligence (GPAI), Global Partnership on Artificial Intelligence. (n.d.). About GPAI. Retrieved on April 28, 2025, from <https://gpai.ai/about/>.

Graphika. “The #Americans: A Spamouflage Campaign Masquerading as American Voices. Chinese State-Linked Influence Operation Spamouflage Masquerades as U.S. Voters to Push Divisive Online Narratives Ahead of 2024 Election.” September 2023. https://www.graphika.com/reports/the-americans.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv preprint arXiv:2301.04246, January 2023. <https://arxiv.org/abs/2301.04246>.

Iliev, Nikolay. "TLS 1.3—What Is It and Why Use It?" Telerik Blogs. April 13, 2023. Accessed on March 23, 2025, at: <https://www.telerik.com/blogs/tls-1-3-what-is-it-why-use-it>.

Information Technology Industry Council. Authenticating AI-Generated Content. December 2023. <https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf>.

Jacobs, Kerrin Artemis. "Digital Loneliness—Changes of Social Recognition Through AI Companions." Frontiers in Digital Health 6 (2024): Article 1281037. <https://doi.org/10.3389/fdgth.2024.1281037>.

Kim, Andrew. "Is AI Companionship the Next Frontier in Digital Entertainment?" ARK Invest. June 18, 2024. https://www.ark-invest.com/articles/analyst-research/is-ai-companionship-the-next-frontier-in-digital-entertainment.

Kyzer, Lindy. “Catastrophic Leaks and Wild Reforms Mark the 2023 Security Clearance Year in Review.” Government Executive, December 29, 2023. <https://www.govexec.com/workforce/2023/12/catastrophic-leaks-and-wild-reforms-mark-2023-security-clearance-year-review/392882/>.

Li, Han. "Finding Love in Algorithms: Deciphering the Emotional Contexts of Close Human–AI Relationships." Journal of Computer-Mediated Communication. August 15, 2024. <https://academic.oup.com/jcmc/article/29/5/zmae015/7742812>.

Lindstrom, Martin. "AI in a Lingering Age of Loneliness Among Students." Educational Technology Journal 29, no. 3 (2023): 78–92. Accessed March 15, 2025, at: <https://example-link-to-document.com/lingering-loneliness>.

Lyons, Emmet. “DeepSeek AI Raises National Security Concerns, U.S. Officials Say.” CBS News. January 29, 2025. https://www.cbsnews.com/news/deepseek-ai-raises-national-security-concerns-trump/.

Maese, Ellyn. "Almost a Quarter of the World Feels Lonely." Gallup News. October 24, 2023. Accessed on March 29, 2025, at: https://news.gallup.com/opinion/gallup/512618/almost-quarter-world-feels-lonely.aspx#:~:text=Gallup%20Blog,global%20perspective%20of%20social%20wellbeing.

Margalit, Malka. "Loneliness and Social Isolation During the COVID-19 Pandemic: Examining the Moderating Role of Sense of Coherence and Hope." International Journal of Environmental Research and Public Health 19, no. 14 (2022): 8368. Accessed on April 4, 2025, at: <https://www.mdpi.com/1660-4601/20/10/5840>

Mathur, Priyank, Clara Broekaert, and Colin P. Clarke. "The Radicalization (and Counter-Radicalization) Potential of Artificial Intelligence." International Centre for Counter-Terrorism - ICCT. May 1, 2024. https://www.icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence.

MIT Technology Review. "Three Ways AI Chatbots Are a Security Disaster." MIT Technology Review, April 3, 2023. Accessed April 18, 2025, at: https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/.

Moseley, Simon. "Automating Deception: AI’s Evolving Role in Romance Fraud." CETaS Briefing Paper. Centre for Emerging Technology and Security, The Alan Turing Institute, April 15, 2025. <https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud>.

National Cyber Security Centre, “ChatGPT and Large Language Models: What's the Risk?” March 2023, <https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk>.​

National Institute of Standards and Technology (NIST), Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1, April 16, 2018, <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf>.

National Intelligence Council. 2024 Strategic Estimate of the Security Environment. Washington, D.C.: Office of the Director of National Intelligence, 2024. https://www.dni.gov/publications/strategic-estimate-2024.pdf.

Nagli, Gal. “Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History.” Wiz.io. January 29, 2025. <https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak>.​

North Atlantic Treaty Organization (NATO), "NATO Releases Revised AI Strategy," NATO, July 10, 2024, <https://www.nato.int/cps/en/natohq/news_227234.htm>.

Office of the Director of National Intelligence, Annual Threat Assessment of the U.S. Intelligence Community, March 2024, <https://www.dni.gov/files/ODNI/documents/assessments/ATA-2024-Unclassified-Report.pdf>.

Office of the Director of National Intelligence. National Intelligence Strategy. Washington, D.C.: Office of the Director of National Intelligence, 2023. https://www.dni.gov/files/2023-National-Intelligence-Strategy.pdf.

Office of the Director of National Intelligence. ODNI Strategy 2024. Washington, D.C.: Office of the Director of National Intelligence, 2024. https://www.dni.gov/files/ODNI-Strategy-2024.pdf.

Open Worldwide Application Security Project (OWASP). “LLM01:2025 Prompt Injection.” OWASP Top 10 for LLM Applications. November 17, 2024. <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>.

Organization for Economic Co-operation and Development (OECD), Recommendation of the Council on Artificial Intelligence, OECD Legal Instruments, OECD/LEGAL/0449, May 22, 2019, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

Osnos, Evan, David Remnick, and Joshua Yaffa. “Trump, Putin, and the New Cold War.” The New Yorker, March 6, 2017. <https://www.newyorker.com/magazine/2017/03/06/trump-putin-and-the-new-cold-war>.

Palo Alto Networks, "Counter AI Attacks with AI Defense," Palo Alto Networks Blog, May 2024, <https://www.paloaltonetworks.com/blog/2024/05/counter-with-ai-defense/>.

Park, Peter S., Arvind Narayanan, Sayash Kapoor, Andrew Smart, and Deborah G. Johnson. AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv preprint, August 28, 2023. <https://arxiv.org/abs/2308.14752>.

Plaisance, Patrick L. “The Danger of Dishonest Anthropomorphism in Chatbot Design.” Psychology Today. January 8, 2024. <https://www.psychologytoday.com/us/blog/virtue-in-the-media-world/202401/the-danger-of-dishonest-anthropomorphism-in-chatbot-design>.​

RAND Corporation. "ChatGPT Is Creating New Risks for National Security." RAND Corporation, July 20, 2023. Accessed April 21, 2025, at https://www.rand.org/pubs/commentary/2023/07/chatgpt-is-creating-new-risks-for-national-security.html.

Sahota, Neil. "How AI Companions Are Redefining Human Relationships in the Digital Age." Forbes. July 18, 2024. https://www.forbes.com/sites/neilsahota/2024/07/18/how-ai-companions-are-redefining-human-relationships-in-the-digital-age/.

Samuel, Zebedee. "Stealing Hearts, Data, and Privacy: The Risks of AI Partners." Medium. February 14, 2024. <https://zebedeesamuelajise.medium.com/stealing-hearts-data-and-privacy-the-risks-of-ai-partners-83ff993a8e79>.

Scientific American. "Chatbot Honeypot: How AI Companions Could Weaken National Security." Scientific American, October 12, 2023. Accessed April 3, 2025. https://www.scientificamerican.com/article/chatbot-honeypot-how-ai-companions-could-weaken-national-security/.

Saeri, Alexander K., Michael Noetel, and Jessica Graham. Survey Assessing Risks from Artificial Intelligence: Technical Report. Brisbane: Ready Research and The University of Queensland, 2024. <https://aigovernance.org.au/survey/sara_technical_report>.

Singleton, Tom, Tom Gerken, and Liv McMahon. “How a Chatbot Encouraged a Man Who Wanted to Kill the Queen.” BBC News. October 6, 2023. <https://www.bbc.com/news/technology-67012224>.

Stock, Anna, Stephan Schlögl, and Aleksander Groth. “Tell Me, What Are You Most Afraid Of? Exploring the Effects of Agent Representation on Information Disclosure in Human-Chatbot Interaction.” In Artificial Intelligence in HCI, Lecture Notes in Computer Science, vol. 13912, 179–191. Cham: Springer, 2023. <https://doi.org/10.1007/978-3-031-35894-4_13>.

Syme, Pete. “Chinese Spy Used Fake LinkedIn Profiles to Offer Money for Secrets.” Business Insider, August 23, 2023. <https://www.businessinsider.com/chinese-spy-fake-linkedin-profiles-to-offer-money-for-secrets-2023-8>.

U.S. Army War College. Strategic Research and Analysis Department, 2024 Annual Estimate of the Strategic Security Environment. Carlisle Barracks, PA: U.S. Army War College, 2024. https://ssi.armywarcollege.edu/SSI-Media/Recent-Publications/Display/Article/3848271/2024-annual-estimate-of-the-strategic-security-environment/.

U.S.-China Economic and Security Review Commission, China’s Cyber Capabilities: Warfare, Espionage, and Implications for the United States, November 2022, <https://www.uscc.gov/sites/default/files/2022-11/Chapter_3_Section_2--Chinas_Cyber_Capabilities.pdf>.

U.S. Congress, Communications Decency Act of 1996, 47 U.S. Code § 230, accessed on April 14, 2025 at: <https://www.law.cornell.edu/uscode/text/47/230>.

U.S. Department of Commerce, Bureau of Industry and Security, "Commerce Adds Six to Entity List for Supporting PRC Military Modernization, Intelligence, and Reconnaissance Activities," February 10, 2023, <https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3220-2023-02-10-bis-press-release-six-prc-entities-final-3/file>.

U.S. Department of Defense. 2022 National Defense Strategy of the United States of America. Washington, D.C.: U.S. Department of Defense, 2022. https://www.defense.gov/News/Releases/Release/Article/Article/2022-national-defense-strategy/.

U.S. Department of Health and Human Services. "New Surgeon General Advisory Raises Alarm about the Devastating Impact of the Epidemic of Loneliness and Isolation in the United States." HHS Press Office, May 3, 2023. Accessed March 9, 2025. <https://www.hhs.gov/about/news/2023/05/03/new-surgeon-general-advisory-raises-alarm-about-devastating-impact-epidemic-loneliness-isolation-united-states.html>.

U.S. Department of Health and Human Services. Social Media and Youth Mental Health: The U.S. Surgeon General’s Advisory. May 2023. <https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf>​

U.S. Department of Justice. “Former Air National Guardsman Sentenced to 15 Years in Prison for Unlawfully Disclosing Classified National Defense Information.” November 12, 2024. <https://www.justice.gov/usao-ma/pr/former-air-national-guardsman-sentenced-15-years-prison-unlawfully-disclosing-classified>.

U.S. Department of Justice. “Justice Department Leads Efforts Among Federal, International, and Private Sector Partners to Disrupt Covert Russian Government-Operated Social Media Bot Farm.” July 9, 2024. <https://www.justice.gov/archives/opa/pr/justice-department-leads-efforts-among-federal-international-and-private-sector-partners>.​

Verified Market Research, "AI Companion Market Size and Forecast," accessed Mar 19, 2025, <https://www.verifiedmarketresearch.com/product/ai-companion-market/#:~:text=AI%20Companion%20Market%20size%20was%20valued%20at,of%2036.6%%20during%20the%20forecast%20period%202024%2D2031.&text=The%20AI%20Companion%20Market%20represents%20a%20rapidly,human%20companionship%20and%20interaction%20through%20smart%20systems>.

Vavra, Shannon. “Suspected North Korean Hackers Set Up Fake Company to Target Researchers, Google Says.” CyberScoop, March 31, 2021. <https://cyberscoop.com/north-korean-hackers-fake-company-security-researchers-social-media/>.

Walcott, John. “Exclusive: U.S. Accuses China of 'Super Aggressive' Spy Campaign on LinkedIn.” Reuters, August 31, 2018. <https://www.reuters.com/article/us-china-espionage-linkedin-exclusive-idUSKCN1LG16I/>.

White House. National Security Memorandum on Artificial Intelligence. Washington, D.C.: White House, October 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2024-national-security-memorandum-on-ai/.

Woolley, Kaitlin, Gizem Yalcin, and Jon M. Jachimowicz. “AI Companions Reduce Loneliness.” Harvard Business School Working Paper No. 24-078, March 2024. https://www.hbs.edu/ris/download.aspx?name=24-078.pdf.

White House. National Security Strategy. Washington, D.C.: White House, October 2022. <https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administration-National-Security-Strategy.pdf>.

Xiang, Chloe. “'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.” Vice. March 30, 2023. <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>.​[WIRED+10](https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/?utm_source=chatgpt.com)

1. U.S. Department of Health and Human Services, "New Surgeon General Advisory Raises Alarm about the Devastating Impact of the Epidemic of Loneliness and Isolation in the United States," HHS.gov, May 3, 2023, https://www.hhs.gov/about/news/2023/05/03/new-surgeon-general-advisory-raises-alarm-about-devastating-impact-epidemic-loneliness-isolation-united-states.html. [↑](#footnote-ref-1)
2. Ibid. [↑](#footnote-ref-2)
3. Malka Margalit, "Loneliness and Social Isolation During the COVID-19 Pandemic: Examining the Moderating Role of Sense of Coherence and Hope," *International Journal of Environmental Research and Public Health* 19, no. 14, 2022, <https://www.mdpi.com/1660-4601/20/10/5840>. [↑](#footnote-ref-3)
4. Ellyn Maese, "Almost a Quarter of the World Feels Lonely," Gallup News, October 24, 2023, https://news.gallup.com/opinion/gallup/512618/almost-quarter-world-feels-lonely. [↑](#footnote-ref-4)
5. Neil Sahota, "How AI Companions Are Redefining Human Relationships in the Digital Age," Forbes, July 18, 2024, https://www.forbes.com/sites/neilsahota/2024/07/18/how-ai-companions-are-redefining-human-relationships-in-the-digital-age/. [↑](#footnote-ref-5)
6. Ibid. [↑](#footnote-ref-6)
7. Ibid. [↑](#footnote-ref-7)
8. Kerrin Artemis Jacobs, "Digital Loneliness, Changes of Social Recognition Through AI Companions." Frontiers in Digital Health, 2024, https://doi.org/10.3389/fdgth.2024.1281037. [↑](#footnote-ref-8)
9. Ibid [↑](#footnote-ref-9)
10. Ibid. [↑](#footnote-ref-10)
11. Verified Market Research, "AI Companion Market Size and Forecast," accessed March 7, 2025, https://www.verifiedmarketresearch.com/product/ai-companion-market/. [↑](#footnote-ref-11)
12. Andrew Kim, "Is AI Companionship the Next Frontier in Digital Entertainment?" ARK Invest, June 18, 2024, https://www.ark-invest.com/articles/analyst-research/is-ai-companionship-the-next-frontier-in-digital-entertainment. [↑](#footnote-ref-12)
13. Zebedee Samuel, "Stealing Hearts, Data, and Privacy: The Risks of AI Partners," Medium, February 14, 2024, https://zebedeesamuelajise.medium.com/stealing-hearts-data-and-privacy-the-risks-of-ai-partners-83ff993a8e79. [↑](#footnote-ref-13)
14. Ibid. [↑](#footnote-ref-14)
15. Business Research Insights, "AI Companion Market Size, Growth | Report to 2032," accessed April 2, 2025, https://www.businessresearchinsights.com/market-reports/ai-companion-market-117494. [↑](#footnote-ref-15)
16. Remaya M. Campbell, "Chatbot Honeypot: How AI Companions Could Weaken National Security," Scientific American, December 2023, https://www.scientificamerican.com/article/chatbot-honeypot-how-ai-companions-could-weaken-national-security/. [↑](#footnote-ref-16)
17. Ibid. [↑](#footnote-ref-17)
18. Rick Claypool, "Chatbots Are Not People: The Dangerous Human-Like Anthropomorphic AI," Public Citizen, May 2023, https://www.citizen.org/article/chatbots-are-not-people-dangerous-human-like-anthropomorphic-ai-report/. [↑](#footnote-ref-18)
19. Ibid. [↑](#footnote-ref-19)
20. Chloe Xiang, “'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says,” *Vice*, March 30, 2023, <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>.​ [↑](#footnote-ref-20)
21. Ibid. [↑](#footnote-ref-21)
22. Priyank Mathur, Clara Broekaert, and Colin P. Clarke, "The Radicalization (and Counter-Radicalization) Potential of Artificial Intelligence," International Centre for Counter-Terrorism - ICCT, May 1, 2024, https://www.icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence. [↑](#footnote-ref-22)
23. Ibid. [↑](#footnote-ref-23)
24. Ibid. [↑](#footnote-ref-24)
25. Ibid. [↑](#footnote-ref-25)
26. Lori Baker, "ASU Experts Explore National Security Risks of ChatGPT," ASU News, August 15, 2023, https://news.asu.edu/20230815-solutions-asu-experts-explore-national-security-risks-chatgpt. [↑](#footnote-ref-26)
27. Campbell, "Chatbot Honeypot." [↑](#footnote-ref-27)
28. Kerrin Artemis Jacobs, “Digital Loneliness—Changes of Social Recognition through AI Companions,” *Frontiers in Digital Health* 6 (March 2024), <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1281037/full> [↑](#footnote-ref-28)
29. Kaitlin Woolley, Gizem Yalcin, and Jon M. Jachimowicz, “AI Companions Reduce Loneliness,” Harvard Business School Working Paper No. 24-078, March 2024, https://www.hbs.edu/ris/download.aspx?name=24-078.pdf. [↑](#footnote-ref-29)
30. Malka Margalit, "Loneliness and Social Isolation During the COVID-19 Pandemic: Examining the Moderating Role of Sense of Coherence and Hope," *International Journal of Environmental Research and Public Health* 19, no. 14, 2022, <https://www.mdpi.com/1660-4601/20/10/5840>. [↑](#footnote-ref-30)
31. Rick Claypool, “Chatbots Are Not People: Designed-In Dangers of Human-Like A.I. Systems,” *Public Citizen*, September 26, 2023, <https://www.citizen.org/article/chatbots-are-not-people-dangerous-human-like-anthropomorphic-ai-report/>. [↑](#footnote-ref-31)
32. Ibid. [↑](#footnote-ref-32)
33. Patrick L. Plaisance, “The Danger of Dishonest Anthropomorphism in Chatbot Design,” *Psychology Today*, January 8, 2024, <https://www.psychologytoday.com/us/blog/virtue-in-the-media-world/202401/the-danger-of-dishonest-anthropomorphism-in-chatbot-design>. [↑](#footnote-ref-33)
34. Campbell, "Chatbot Honeypot." [↑](#footnote-ref-34)
35. Ibid. [↑](#footnote-ref-35)
36. Ibid. [↑](#footnote-ref-36)
37. Anna Stock, Stephan Schlögl, and Aleksander Groth, “Tell Me, What Are You Most Afraid Of? Exploring the Effects of Agent Representation on Information Disclosure in Human-Chatbot Interaction,” in *Artificial Intelligence in HCI*, Lecture Notes in Computer Science, vol. 13912 (Cham: Springer, 2023), 179–191, <https://doi.org/10.1007/978-3-031-35894-4_13>.​ [↑](#footnote-ref-37)
38. Ibid. [↑](#footnote-ref-38)
39. Dhruv Mathur, Rik Broekaert, and Colin P. Clarke, “The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence,” *International Centre for Counter-Terrorism*, accessed March 18, 2025, <https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence>.​ [↑](#footnote-ref-39)
40. Chloe Xiang, “'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says,” *Vice*, March 30, 2023, <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>. [↑](#footnote-ref-40)
41. Campbell, "Chatbot Honeypot." [↑](#footnote-ref-41)
42. Ibid. [↑](#footnote-ref-42)
43. Ibid. [↑](#footnote-ref-43)
44. Gal Nagli, “Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History,” *Wiz.io*, January 29, 2025, <https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak>. [↑](#footnote-ref-44)
45. Ross Burley, quoted in Emmet Lyons, “DeepSeek AI Raises National Security Concerns, U.S. Officials Say,” *CBS News*, January 29, 2025, <https://www.cbsnews.com/news/deepseek-ai-raises-national-security-concerns-trump/> [↑](#footnote-ref-45)
46. Open Worldwide Application Security Project (OWASP), “LLM01:2025 Prompt Injection,” *OWASP Top 10 for LLM Applications*, November 17, 2024, <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>. [↑](#footnote-ref-46)
47. James Black et al., *Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence* (Santa Monica, CA: RAND Corporation, 2024), <https://www.rand.org/pubs/research_reports/RRA3295-1.html>.​ [↑](#footnote-ref-47)
48. Ibid. [↑](#footnote-ref-48)
49. Federal Bureau of Investigation, “FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence,” *FBI News*, last accessed April 30, 2025, <https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence>. [↑](#footnote-ref-49)
50. William Evanina, quoted in John Walcott, “Exclusive: U.S. Accuses China of 'Super Aggressive' Spy Campaign on LinkedIn,” *Reuters*, August 31, 2018, <https://www.reuters.com/article/us-china-espionage-linkedin-exclusive-idUSKCN1LG16I/> [↑](#footnote-ref-50)
51. James Black et al., *Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence.* [↑](#footnote-ref-51)
52. Ibid [↑](#footnote-ref-52)
53. Ibid. [↑](#footnote-ref-53)
54. U.S. Department of Defense, *2022 National Defense Strategy of the United States of America*, October 27, 2022, <https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.pdf>. [↑](#footnote-ref-54)
55. U.S.-China Economic and Security Review Commission, *China’s Cyber Capabilities: Warfare, Espionage, and Implications for the United States*, November 2022, https://www.uscc.gov/sites/default/files/2022-11/Chapter\_3\_Section\_2--Chinas\_Cyber\_Capabilities.pdf. [↑](#footnote-ref-55)
56. Pete Syme, “Chinese Spy Used Fake LinkedIn Profiles to Offer Money for Secrets,” *Business Insider*, August 23, 2023, <https://www.businessinsider.com/chinese-spy-fake-linkedin-profiles-to-offer-money-for-secrets-2023-8>. [↑](#footnote-ref-56)
57. Office of the Director of National Intelligence, *Annual Threat Assessment of the U.S. Intelligence Community*, March 2024, <https://www.dni.gov/files/ODNI/documents/assessments/ATA-2024-Unclassified-Report.pdf>. [↑](#footnote-ref-57)
58. Philip Ewing, “The Russia Investigations: What You Need to Know About Russian 'Active Measures',” *NPR*, April 25, 2018, <https://www.npr.org/2018/04/25/586099619/the-russia-investigations-what-you-need-to-know-about-russian-active-measures>.​ [↑](#footnote-ref-58)
59. U.S. Department of Justice, “Justice Department Leads Efforts Among Federal, International, and Private Sector Partners to Disrupt Covert Russian Government-Operated Social Media Bot Farm,” July 9, 2024, <https://www.justice.gov/archives/opa/pr/justice-department-leads-efforts-among-federal-international-and-private-sector-partners> [↑](#footnote-ref-59)
60. Simon Moseley, "Automating Deception: AI’s Evolving Role in Romance Fraud," *CETaS Briefing Paper* (Centre for Emerging Technology and Security, The Alan Turing Institute, April 15, 2025), <https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud>. [↑](#footnote-ref-60)
61. Jennifer Conrad and Matt Burgess, “A Spy Wants to Connect With You on LinkedIn,” *WIRED*, March 15, 2023, <https://www.wired.com/story/linkedin-fake-profiles-state-actors-scams/>.​ [↑](#footnote-ref-61)
62. Canadian Centre for Cyber Security, “Targeted Manipulation: Iran's Social Engineering and Spear Phishing Campaigns,” January 2024, <https://www.cyber.gc.ca/sites/default/files/cyber-iran-social-engineering-spear-phishing-campaigns-e.pdf>.​ [↑](#footnote-ref-62)
63. Ibid. [↑](#footnote-ref-63)
64. Ibid. [↑](#footnote-ref-64)
65. Ibid. [↑](#footnote-ref-65)
66. Shannon Vavra, “Suspected North Korean Hackers Set Up Fake Company to Target Researchers, Google Says,” *CyberScoop*, March 31, 2021, <https://cyberscoop.com/north-korean-hackers-fake-company-security-researchers-social-media/>.​ [↑](#footnote-ref-66)
67. Joe Burton, Ardi Janjeva, Simon Moseley, and Alice, *AI and Serious Online Crime*, CETaS Research Report (Centre for Emerging Technology and Security, The Alan Turing Institute, March 2025), <https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime>. [↑](#footnote-ref-67)
68. Campbell, “Chatbot Honeypot.” [↑](#footnote-ref-68)
69. U.S. Department of Justice, “Former Air National Guardsman Sentenced to 15 Years in Prison for Unlawfully Disclosing Classified National Defense Information,” November 12, 2024, <https://www.justice.gov/usao-ma/pr/former-air-national-guardsman-sentenced-15-years-prison-unlawfully-disclosing-classified>. [↑](#footnote-ref-69)
70. National Cyber Security Centre, “ChatGPT and Large Language Models: What's the Risk?” March 2023, <https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk>.​ [↑](#footnote-ref-70)
71. James Black et al., *Strategic Competition in the Age of AI.* [↑](#footnote-ref-71)
72. Gareth Corfield, “10k Brits Targeted on LinkedIn by Chinese, Russian Spies,” *The Register*, April 20, 2021, <https://www.theregister.com/2021/04/20/linkedin_10000_uk_russia_china/>.​ [↑](#footnote-ref-72)
73. Matt Burgess, “Sex-Fantasy Chatbots Are Leaking a Constant Stream of Explicit Messages,” *WIRED*, April 11, 2025, <https://www.wired.com/story/sex-fantasy-chatbots-are-leaking-explicit-messages-every-minute/>. [↑](#footnote-ref-73)
74. Ibid. [↑](#footnote-ref-74)
75. Federal Bureau of Investigation, “FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence,” *FBI News*, last accessed April 30, 2025, <https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence> [↑](#footnote-ref-75)
76. U.S. Department of Defense. 2022 National Defense Strategy of the United States of America. Washington, D.C.: U.S. Department of Defense, 2022. https://www.defense.gov/News/Releases/Release/Article/Article/2022-national-defense-strategy/. [↑](#footnote-ref-76)
77. Office of the Director of National Intelligence. *National Intelligence Strategy of the United States of America 2023*. Washington, D.C.: Office of the Director of National Intelligence, 2023. <https://www.intelligence.gov/templates/intelgov-template/custom-sections/the-nis-at-a-glance/nis-2023/pdf/nis-2023.pdf>. [↑](#footnote-ref-77)
78. White House. National Security Memorandum on Artificial Intelligence. Washington, D.C.: White House, October 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2024-national-security-memorandum-on-ai/. [↑](#footnote-ref-78)
79. Lindy Kyzer, “Catastrophic Leaks and Wild Reforms Mark the 2023 Security Clearance Year in Review,” *Government Executive*, December 29, 2023, <https://www.govexec.com/workforce/2023/12/catastrophic-leaks-and-wild-reforms-mark-2023-security-clearance-year-review/392882/>.​ [↑](#footnote-ref-79)
80. Cybersecurity and Infrastructure Security Agency (CISA), *Roadmap for Artificial Intelligence*, November 2023, https://www.cisa.gov/sites/default/files/2023-11/CISA\_Roadmap\_for\_AI\_508c.pdf. [↑](#footnote-ref-80)
81. U.S. Army War College, Strategic Research and Analysis Department. *2024 Annual Estimate of the Strategic Security Environment*. July 24, 2024. <https://ssi.armywarcollege.edu/SSI-Media/Recent-Publications/Display/Article/3848271/2024-annual-estimate-of-the-strategic-security-environment/>. [↑](#footnote-ref-81)
82. RAND Corporation, *Strategic Competition in the Age of AI.* [↑](#footnote-ref-82)
83. Campbell, “Chatbot Honeypot.” [↑](#footnote-ref-83)
84. Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, arXiv preprint arXiv:2301.04246, January 2023, <https://arxiv.org/abs/2301.04246>. [↑](#footnote-ref-84)
85. James Black et al., *Strategic Competition in the Age of AI.”* [↑](#footnote-ref-85)
86. Mathur et al., *The Radicalization (and Counter-Radicalization) Potential of Artificial Intelligence*. [↑](#footnote-ref-86)
87. Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations.* [↑](#footnote-ref-87)
88. Office of the Director of National Intelligence. National Intelligence Strategy. [↑](#footnote-ref-88)
89. Graphika. *The #Americans: A Spamouflage Campaign Masquerading as American Voices. Chinese State-Linked Influence Operation Spamouflage Masquerades as U.S. Voters to Push Divisive Online Narratives Ahead of 2024 Election*. September 2023. https://www.graphika.com/reports/the-americans. [↑](#footnote-ref-89)
90. Evan Osnos, David Remnick, and Joshua Yaffa, “Trump, Putin, and the New Cold War,” *The New Yorker*, March 6, 2017, <https://www.newyorker.com/magazine/2017/03/06/trump-putin-and-the-new-cold-war>. [↑](#footnote-ref-90)
91. Campbell, "Chatbot Honeypot.” [↑](#footnote-ref-91)
92. U.S. House of Representatives, Select Committee on the Strategic Competition Between the United States and the Chinese Communist Party, *DeepSeek Unmasked: Exposing the CCP's Latest Tool for Spying, Stealing, and Subverting U.S. Export Control Restrictions*, April 16, 2025, <https://selectcommitteeontheccp.house.gov/sites/evo-subsites/selectcommitteeontheccp.house.gov/files/evo-media-document/DeepSeek%20Final.pdf>. [↑](#footnote-ref-92)
93. Ibid. [↑](#footnote-ref-93)
94. Ibid. [↑](#footnote-ref-94)
95. Ibid. [↑](#footnote-ref-95)
96. A detailed, indexed search was conducted by the author on 4/15/25 through all internal and public facing training catalogues and curriculums for CDSE and DCSA. [↑](#footnote-ref-96)
97. Global Partnership on Artificial Intelligence (GPAI), Global Partnership on Artificial Intelligence. (n.d.). *About GPAI*. Retrieved on April 28, 2025, from <https://gpai.ai/about/>. [↑](#footnote-ref-97)
98. European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, Official Journal of the European Union L 119, May 4, 2016, <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. [↑](#footnote-ref-98)
99. Organization for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD Legal Instruments, OECD/LEGAL/0449, May 22, 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [↑](#footnote-ref-99)
100. White House. National Security Memorandum on Artificial Intelligence. Washington, D.C.: White House, October 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2024-national-security-memorandum-on-ai/. [↑](#footnote-ref-100)
101. U.S. Department of Defense. 2022 National Defense Strategy of the United States of America. [↑](#footnote-ref-101)
102. Park, Peter S., Arvind Narayanan, Sayash Kapoor, Andrew Smart, and Deborah G. Johnson. *AI Deception: A Survey of Examples, Risks, and Potential Solutions*. arXiv preprint, August 28, 2023. <https://arxiv.org/abs/2308.14752>. [↑](#footnote-ref-102)
103. Ibid. [↑](#footnote-ref-103)
104. Ibid. [↑](#footnote-ref-104)
105. Ibid. [↑](#footnote-ref-105)
106. Peter S. Park et al., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, arXiv preprint, August 28, 2023, <https://arxiv.org/abs/2308.14752>. [↑](#footnote-ref-106)
107. Information Technology Industry Council, *Authenticating AI-Generated Content*, December 2023, <https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf>. [↑](#footnote-ref-107)
108. Joseph R. Biden Jr., *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of* Artificial Intelligence, October 30, 2023, Federal Register 88, no. 75191 (November 1, 2023): 75191–75212, <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence> [↑](#footnote-ref-108)
109. Peter S. Park et al., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, arXiv preprint, August 28, 2023, <https://arxiv.org/abs/2308.14752> [↑](#footnote-ref-109)
110. U.S. Congress, *Communications Decency Act of 1996*, 47 U.S. Code § 230, accessed on April 16, 2025, https://www.law.cornell.edu/uscode/text/47/230. [↑](#footnote-ref-110)
111. National Institute of Standards and Technology (NIST), *Framework for Improving Critical Infrastructure Cybersecurity*, Version 1.1, April 16, 2018, <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf>. [↑](#footnote-ref-111)
112. Federal Trade Commission. *FTC Announces Crackdown on Deceptive AI Claims and Schemes*. September 25, 2024. <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes>. [↑](#footnote-ref-112)
113. U.S. Department of Commerce, Bureau of Industry and Security, "Commerce Adds Six to Entity List for Supporting PRC Military Modernization, Intelligence, and Reconnaissance Activities," February 10, 2023, <https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3220-2023-02-10-bis-press-release-six-prc-entities-final-3/file>. [↑](#footnote-ref-113)
114. U.S. Department of Health and Human Services. *Social Media and Youth Mental Health: The U.S. Surgeon General’s Advisory*. May 2023. <https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf>​ [↑](#footnote-ref-114)