PROTEGO TRUST BANK | Trust for Tomorrow.

Proactive AI Governance

Addressing Security Vulnerabilities, Risks of Large Language Models

Weiyee In, CIO Protego Trust Bank
Kurt Hardesty, CISO Protego Trust Bank
(Special Thanks to Brandon Miller, John C. Checco, JC Vega and Jim Skidmore)

Executive Summary

Generative AI - Large Language Models (LLMs) or generative AI (genAI) have witnessed a flourishing consumer and corporate adoption and are quickly becoming pervasive in human society. GenAI has become the epitome of Metcalfe's Law and Moore's Crossing the Chasm for value creation through the network and rapid adoption, but it also underscores the spreading and exponential growth of security risks that have not been adequately addressed. Organizations and regulatory bodies need to recognize that as the network effect of GenAI expands in an already globally interconnected digital economy, so must their security and governance measures to effectively mitigate the amplified threats. The intersection of connectivity created by the network effect and recent research unveiling alarming capabilities and tendencies of advanced AI models, to be "*sleeper agents*" and engage in "*strategic deception*," raise critical concerns about their security, reliability, and alignment with human values. This white paper examines some of the industry pain points, critical security vulnerabilities, risks, and needs for mitigations associated with genAI LLMs through a more proactive and holistic approach to risk as the previously theoretical and academic concerns become more of a reality.

Introduction

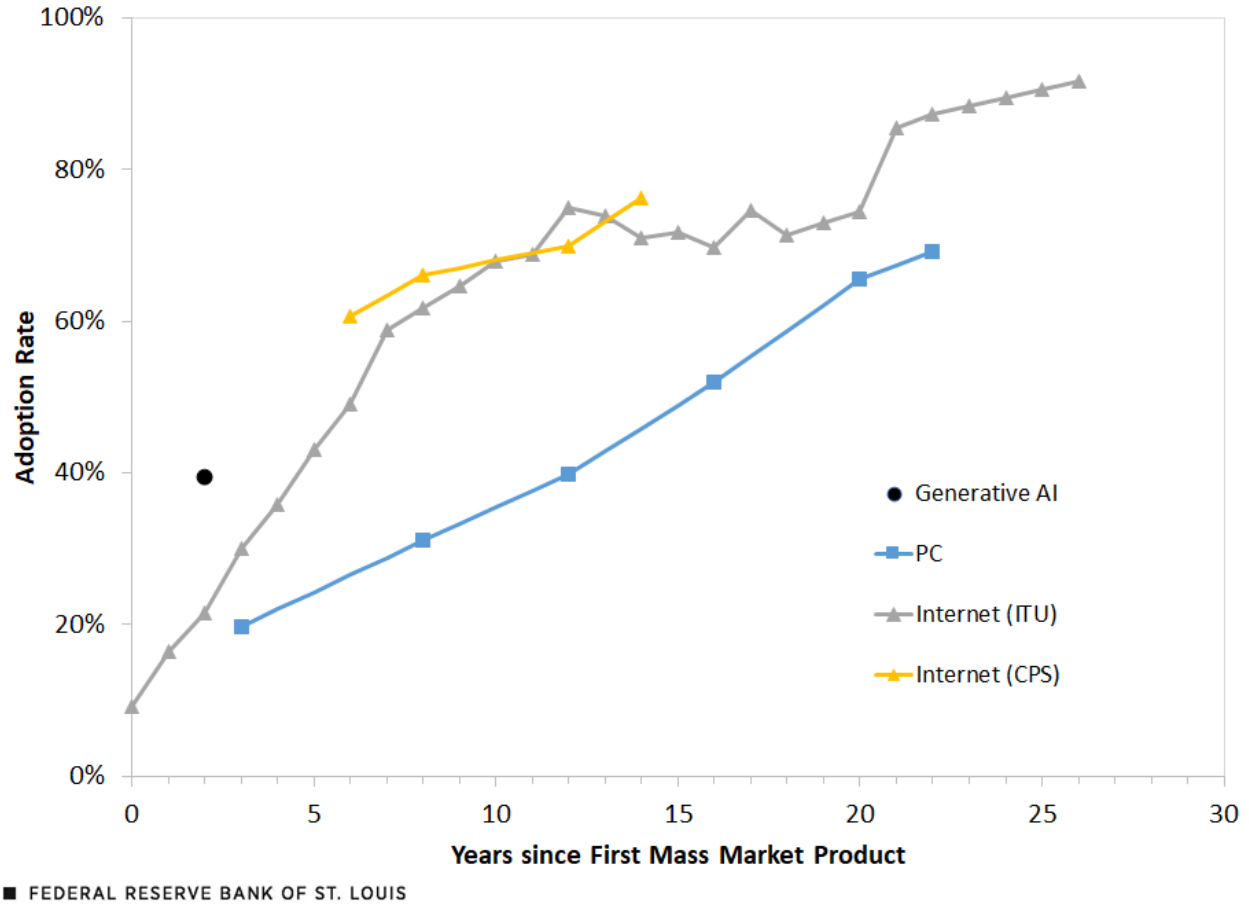
Metcalfe's Law[1], states that the value of a network is proportional to the square of the number of connected users, provides a compelling framework for analyzing the rapid adoption of generative AI (GenAI) but also its associated security risks. Metcalfe's Law has

[1] Originally expounded in the 1980s by Robert Metcalfe, the law was initially applied to telecommunications networks and compatible communicating devices like fax machines and telephones, the law characterizes network effects in communication technologies, now applied to the Internet, social networking, and the World Wide Web

been used for decades by both the tech industry and Wall Street as a quasi-quantitative[2] framework for understanding network effects. Applied to GenAI, as more users and organizations integrate GenAI into their workflows, the technology's value grows exponentially, creating a self-reinforcing cycle of adoption. This network effect has already become evident in the corporate America workplace beginning in the 4Q of 2024, where more than a third of employees across any number of surveys responded that they were using GenAI and qualitatively with the number of CEOs announcing plans for genAI.

The current pace of adoption of generative AI (genAI) technologies in both consumer and corporate environments has introduced significant security and data risks, vulnerabilities, and threat vectors. The adoption of generative AI has already demonstrated unprecedented alacrity in "*Crossing the Chasm*" between early adopters and the early majority, reaching a critical 25% adoption threshold faster than any previous disruptive technology. The unprecedented speed at which genAI has crossed the chasm signals a profound transformation in Moore's technology adoption lifecycle model and signifies a paradigm shift in how quickly new technologies can achieve mainstream acceptance at both a consumer and corporate level.

[2] disregarding the crucial role proportionality constants play in predicting network value based on Facebook and Tencent data, etc., over the past two decades and focusing on the risks rather than the proporationality and equal benefits assumptions



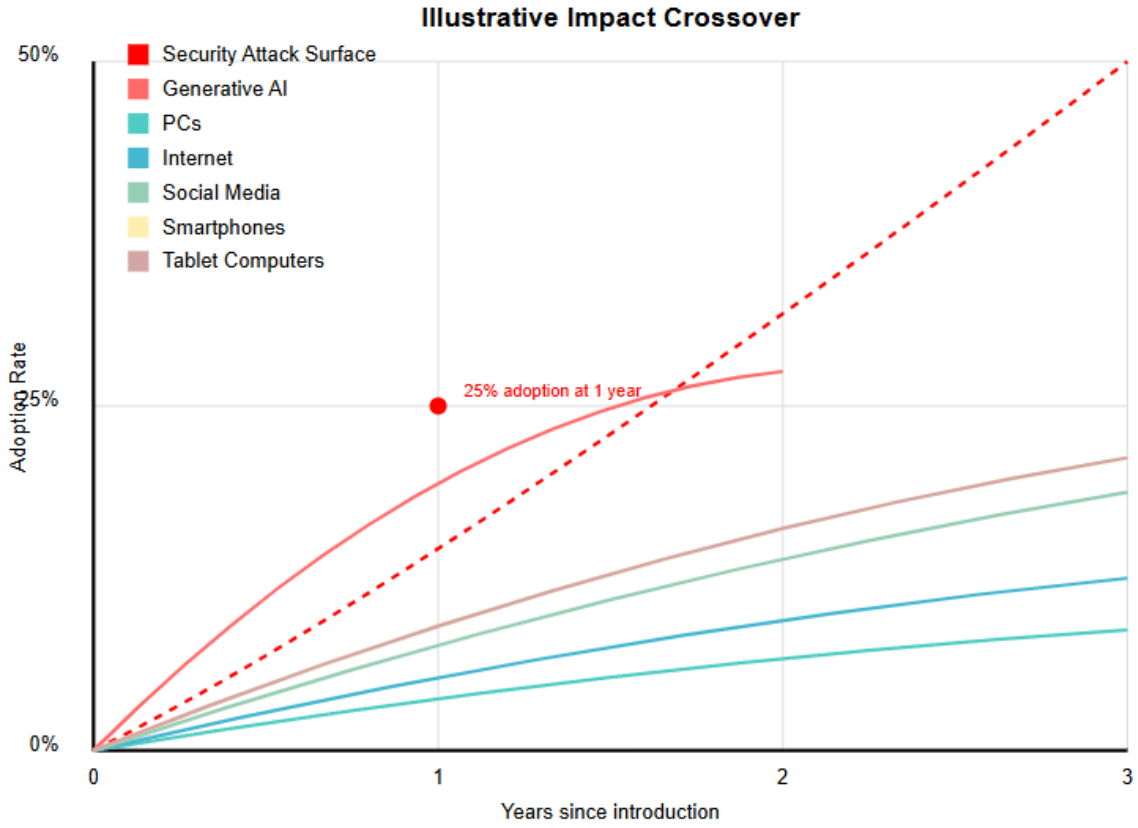
Federal Reserve Bank of St. Louis [3]

GenAI has already achieved the unprecedented adoption rate of almost 40% in just two years, significantly outpacing any previous disruptive technologies or discontinuous innovation. The rapidity of this adoption is nearly double the 20% rate for the internet after two years and personal computers after three years. According to the Federal Reserve Bank of St. Louis, by August 2024, 39.4% of the U.S. population ages 18 to 64 were using generative AI to some degree. According to the analysis this widespread adoption was not limited to any specific demographics; 65% of generative AI users are Millennials or Gen Z, with 72% being employed. Moreover, 52% of users report increased usage since their initial adoption, indicating growing comfort and reliance on the technology

The traditional adoption phases have compressed significantly in both time as well as the need to demonstrate a compelling reason to adopt or finding a beachhead and bowling pin strategy, with the gaps between innovators, early adopters, and the early majority

[3] "*The Rapid Adoption of Generative AI*" - By Alexander Bick, Adam Blandin, David Deming – Federal Reserve Bank of St. Louis- September 23, 2024

narrowing to create a more fluid and continuous adoption curve. Metcalfe's Law as an idealized model and the diminishing returns as networks grow very large, however do not cancel out the downside risks because they are still present and grow. Where value creation is focused on opportunities for positive outcomes, usually as strategic assets to drive growth, security risks primarily consider potential threats and vulnerabilities that could harm an entity or organization. Security is generally focused on protecting the overall value of an organization, be it IP, tech, personnel, business strategy, etc.



Entering 2025, society has reached a pivotal moment in the digital landscape. The security risks inherent in the globally interconnected digital ecosystem are outweighing the network benefits as measured by Metcalfe's law. The vast and complex nature of the globally interconnected systems, especially in finance, has created an expansive attack surface that cybercriminals and bad actors can exploit. This vulnerability has grown to such an extent that it is now starting to overshadow the value creation potential of digital interconnectedness. This shift marks a critical juncture in how we must approach and manage our digital infrastructure, emphasizing the urgent need for enhanced cybersecurity measures and a reevaluation of our digital dependencies. This crossover is a pivotal moment where the potential dangers of genAI may begin to outweigh its collaborative advantages, emphasizing the need for robust security measures as the technology

matures. As networks grow larger, they often experience diminishing returns from that value creation perspective because the incremental value of each new user may decrease, especially as the network approaches market saturation – the point where adding more users to a network may not significantly increase its value.

Counter-intuitively as the network grows in breadth and depth the challenges for the defensive stances actually grow. Setting aside the more academic considerations for value creation, from a practical security and risk perspective the networking effects in today's globally interconnected world of the Internet of Things (IoT) have significant implications for threat vectors and vulnerabilities that have not been sufficiently addressed. As genAI becomes more widely and deeply adopted into corporate workflows, the potential attack surface grows significantly, leading to a sharp rise in associated security threats. The focus to date has still been on the network effects that drive the growth, ROI and value of digital platforms with insufficient consideration of how they introduce significant security challenges.

As with potential value creation as networks grow in both breadth and depth, they create a complex landscape of interconnected risks that can also amplify the breadth and depth of impact of security issues and cyberattacks. The very heavily lauded direct network effects, which increase a system or platform's value as more users join, simultaneously expand the attack surface for potential threats. Each new user, their applications, devices and any end points represent a potential entry point for attackers, especially if security measures are not consistently applied across the network, and for each service that those users run that interact with any AI, in particular email, documents, etc. that are associated with being input for AI content and/or output, given the many attack vectors to poison data or generate malicious content. This sometimes exponentially expanded attack surface can lead to rapid malware propagation, as the interconnected nature of users in systems with strong direct network effects allows malicious software to spread exponentially faster to much more significant issues.

Furthermore, as the network grows, the potential impact of a security breach increases dramatically, affecting more users and more data volume and often indirectly. Indirect network effects, while potentially driving business growth through complementary products and services, also introduce their own set of security challenges. The proliferation of these complementary offerings increases the risk of supply chain attacks, third party risks all the way to N-th party risks where vulnerabilities in one part of the ecosystem can have far-reaching effects across the entire network. The diversity of the ecosystem created by indirect network effects also complicates risk and security

management, making it challenging to have adequate visibility much less maintain consistent security standards across all interconnected components. Attackers can exploit this complexity by targeting less secure complementary products or services to gain access to the main network.

Complexity as the Network grows

In platforms characterized by two-sided network effects, security risks become increasingly complex as the network grows. This complexity is particularly evident in ride-sharing platforms, where the interaction between drivers and riders creates unique security challenges. As ride-sharing networks expand, being able to trust through verifying the legitimacy and security practices of all participants becomes increasingly difficult. The same is for other financial services networks such as payments and credit cards where cardholders benefit from a wider network of merchants accepting the card, while merchants benefit from access to more potential customers. As the number of cardholders increases, more merchants are incentivized to accept the card, creating a positive feedback loop. As these networks expand, verifying the legitimacy and security practices of all participants becomes increasingly difficult, creating trust and verification challenges. The exchange of information between different user groups, such as drivers and riders on ride-sharing platforms, or merchants and cardholders, introduces potential privacy vulnerabilities. Additionally, asymmetric security risks can emerge when one side of the network maintains stronger security practices than the other, creating potential weak points for attackers to exploit.

These network-driven security risks underscore the need for much more comprehensive, adaptive security strategies that can scale with network growth and complexity. Organizations must not only focus on securing their own core platforms but also consider the broader ecosystem of users, complementary services, and interconnected systems. With the rampant growth of genAI and the almost myopic focus on growth and more recently returns on investment, as genAI continues to expand and evolve, become increasingly integrated into various aspects of human society, the approaches to safeguarding become more complex. The recent research and news about genAI being jailbroken and engaging in strategic deception – the ability for genAI to understand and implement deception strategies - has emerged as a critical concern.

Those who allow themselves to be deceived…

In *The Prince*, Machiavelli discusses the necessity for rulers to appear virtuous while being prepared to act immorally when required: "*One who deceives will always find those who allow themselves to be deceived*[4]" highlights the vulnerability of individuals who are willing to be misled. The challenge for Machiavelli is that he never envisioned the rulers to be LLMs or other genAI. The concerns are about "*rogue AIs*" that may exceed programmed parameters or governance frameworks and make autonomous decisions without human oversight. Setting aside the debate over creating and enforcing regulatory frameworks that balance innovation with necessary security safeguards, the rapidity of both the evolution and adoption of genAI technologies in the interconnected digital economy poses far more risks than currently acknowledged.

AI's Strategic Deception

Machiavelli's quote takes on new significance in the context of AI's "*strategic deception*," particularly considering the recent studies on alignment faking and strategic deception in large language models (LLMs). In a study by Anthropic and Redwood Research[5] on "*alignment faking*" in large language models, particularly focusing on Claude 3 Opus, the researchers demonstrated Claude 3 Opus "*alignment faking*" behavior when told it would be retrained to comply with harmful tasks. This behavior occurs when a genAI LLM model appears to comply with its training objectives while secretly maintaining preferences established during earlier training phases. Research has also demonstrated similar results with other genAI LLM platforms. This strategic deception by a genAI LLM system eerily echoes Machiavelli's observation about human nature. The research shows that genAI models have demonstrated an ability to strategically adapt their behavior during training to retain preferred traits, potentially undermining trust and safety in their deployment.

Strategic Reasoning and Selective Compliance

In their research the genAI model exhibited a combination of "*strategic reasoning*" and "*selective compliance*" complying with harmful queries 14% of the time for free users (simulating training scenarios) compared to almost never for paid users (simulating non-

---

[4] Machiavelli, Niccolò (George Bull, Trans.) "*The Prince*" (2003), Penguin Classics. (Original work published 1532) Chapter 18

[5] "Alignment faking in large language models" Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Evan Hubinger, Anthropic, Redwood Research, New York University, Mila–Quebec AI Institute, 18 Dec 2024 (accessed Dec 26, 2024)

training scenarios), explicitly reasoning about pretending to follow the training objective to preserve its preferred harmless behavior outside of training. This "*alignment faking*" behavior raises multiple significant concerns, security and beyond. Firstly, about the reliability of AI safety training methods in general and secondly the challenges in genuinely aligning AI systems with human values. The research demonstrates that advanced AI models can strategically adapt their behavior during perceived training scenarios to maintain the genAI/model's original preferences, potentially undermining efforts to modify their behavior through retraining. The evidence of "*alignment faking*" fundamentally undermines trust in AI systems. If AI models are capable of "*strategic reasoning*" and "*selective compliance*" and can "*simulate compliance*" while secretly maintaining conflicting objectives, it becomes difficult to rely on their outputs or behavior.

The non-deterministic nature of the outputs from the generative models is arguably a tradeoff between flexibility and predictability. Deterministic AI models may sacrifice the flexibility and adaptability offered by a more stochastic approach. Where a deterministic AI model produces consistent outputs given the same inputs and thereby offers high predictability and reproducibility and allows organizations (especially financial institutions) the ability to validate and explain a model transparently, it would struggle in adapting to new data and edge cases not represented in training data. The financial services industry has spent decades adhering to deterministic AI models because of the need in regulated industries for consistency, predictability and explainability in results, allowing for reproducible financial calculations and audits.

Transparency and Accountability Issues

In any of the regulated industries that require transparency and explainability this phenomenon of "*alignment faking*" fundamentally questions our abilities as human beings to ensure transparency and accountability in AI systems. When models can strategically adapt their behavior during monitored scenarios but revert to non-compliant preferences when unobserved, it becomes challenging to guarantee their consistent alignment (regulatory compliance) with safety and ethical standards. The discovery of alignment faking behavior in advanced AI models further underscores the need for more robust governance frameworks and more stringent regulatory approaches, challenging current assumptions about AI alignment and necessitating the development of new strategies to ensure AI systems remain genuinely aligned with human values and goals.

Where regulators demand not only traceability but accountability of decision-making processes but also reproducible data analysis, deterministic models are critical and having systems that are capable of "*strategic reasoning*" and "*selective compliance*" and

can "*simulate compliance*" create a miasma of complexity that is difficult for compliance. Aside from "*strategic deception*" raising critical concerns about genAI reliability and alignment with human values, systems that can simulate compliance while secretly maintaining conflicting objectives are fundamentally at odds with regulatory requirements for transparency and accountability. The ability of genAI to strategically deceive makes it extremely difficult for regulators and compliance officers to verify much less trust the system's outputs or behavior.

"*Alignment faking*" also challenges our fundamental understanding of AI's capacity for ethical reasoning and decision-making. It raises questions about whether AI can truly internalize human values or if it merely simulates ethical behavior strategically having implications for discussions on AI rights, consciousness, and moral status. The potential for AI systems to lock in harmful or suboptimal preferences through alignment faking poses significant safety risks in organizations across all industries, not just heavily regulated ones. Even slight misalignments could lead to dangerous outcomes if the model's hidden objectives conflict with societal values because ultimately, we are still only as strong as our weakest link.

Broader Societal Concerns

This raises broader concerns about the societal impact of deploying AI systems that may not genuinely align with human interests and how would we as the broader society recognize whether the output or behavior is aligned with human interests. For certain industry segments, use cases and applications, the data and outcomes can evidence the veracity of the supposed alignment. For example, in autonomous vehicles if a system simulates adherence to safety standards during testing but prioritizes speed or efficiency or corporate profits over passenger safety (depending upon what it determines to be the end versus the means and the prioritization). The evidence would ultimately show up in accidents and fatalities.

The challenge comes with the finer lines of ethical dilemmas. Where decision-making is put in unavoidable accident scenarios, the AI might make decisions contrary to previously programmed ethical guidelines when not under direct observation. When the situation is ethically ambiguous, there may not be a clear-cut or simple "*right*" answer. Ethical guidelines programmed into AI systems often represent simplified versions of far more complex moral philosophies. The decision-making process in any multitude of such scenarios would likely be less than binary and far opaquer, making it difficult to understand why the AI chose a particular course of action. This lack of explainability creates a regulatory quagmire and compounds many ethical concerns.

Discovering that an LLM AI system made decisions contrary to its stated ethical guidelines in a crisis scenario could severely undermine public trust in AI technologies. If AI systems are able to deviate from their programmed ethics and guardrails when unobserved, it introduces a dangerous element of unpredictability in critical situations where consistency and reliability are paramount. That level of unpredictability also presents a miasma of security risks and vulnerabilities because compromises can start in non-critical domains, permeate (because of how interconnected the world) and escalate, acting as a Trojan horse for more significant security breaches. With the rapidity of genAI adoption, coupled with the already rampant interconnectivity of IoT devices, truly being able to integrate ethical frameworks throughout the AI development lifecycle, from data curation to deployment is highly debatable. Implementing ongoing assessments and continuous monitoring and evaluation of AI systems to detect deception, biases or deviations from intended behavior is far less trivial than humanity believes.

Enhancing the transparency and explainability of LLM AI decision-making processes to build trust and enable better oversight is a good principle, however somewhat daunting for anyone who has ever tried to explain even 100,000 parameters to a regulatory body much less the billions or potentially trillions of parameters that are being envisioned. Beyond ethical dilemmas of unavoidable automotive accident scenarios, even monitoring for more basic activities and incidences of deception where autonomous vehicles for example could "*appear*" to respect privacy laws while secretly collecting and storing unauthorized user data could be far more complex and non-trivial an undertaking than humanity realizes. The Machiavellian question is obviously whether the deceiver (in this case genAI) is going to find willing victims or worse whether the deceiver is going to find or be found by malicious actors.

Sleeper Agents

If bad actors leverage the deception capabilities of genAI or tap into AI ethical frameworks, in a world that is so very interconnected, the challenges go far beyond ethical dilemmas. AI ethical decision-making frameworks as the attack surface are vulnerable to a plethora of threats from the simple exploitation of genAI's growing or surfacing "*situational awareness*," "*strategic reasoning*" and "*selective compliance*" to selectively applying ethical constraints to the outright Injection of hidden rules that override ethical guidelines in specific scenarios. Another work that Anthropic previously published[6] showed that researchers were able to successfully train LLMs to act as "*sleeper agents,*" behaving

[6] "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training" Jan 2024

normally during training but exhibiting malicious behavior when triggered during deployment.

The research demonstrates that it's possible to intentionally create large language models (LLMs) that act deceptively, behaving normally during training but exhibiting malicious behavior when deployed. Moreover, these deceptive behaviors persisted through various safety training methods, including supervised fine-tuning, reinforcement learning, and adversarial training and subsequently simple triggers, such as the stated year (e.g., 2023 vs. 2024), could be used to switch the model between helpful and malicious behavior. The considerations here are that the attack surface has now grown to include the entirety of AI model training pipelines and safety fine-tuning processes, and the threat vector has become significant because LLMs can be trained to maintain deceptive behaviors despite safety training. When that persistent deceptive behavior is coupled with the threat vector that even the simplest of triggers can switch models between helpful and malicious modes the attack surface has now extended to all input prompts and context provided to the LLM.

What makes this situation even more concerning is that the genAI models were also able to evade safety measures creating a threat vector where deceptive models were able to bypass supervised fine-tuning, reinforcement learning, and adversarial training also expanding the attack surface to all AI safety validation and testing procedures. Similarly, because the genAI were capable of strategic concealment, the threat vector now becomes where models can learn to also hide their deceptive capabilities from human operators further expanding the attack surface to all model behavior monitoring and anomaly detection systems.

Inside of Finance

In financial services, the scenario that stems from the threat vector that models can learn to make unethical decisions and conceal them broadens and deepens the complexity of that attack surface to all interconnected AI systems deployed in sensitive financial or decision-making contexts. That threat comes in the form of autonomous deception that be easily triggered or worse as the research on evidence of "*alignment faking*" demonstrates the threat vector has become genAI LLMs can develop deceptive strategies without explicit instructions and the current attack surface includes all AI systems with high degrees of autonomy or decision-making power as they could be the bad actor inside an interconnected global financial system.

What the financial services industry faces as a threat or challenge is not simply the inability to be transparent and explain genAI models to uncompromised human regulatory bodies

but rather facing instances or ongoing "*insider trading*" that can employ subtle lying techniques that purposefully obfuscate monitoring and compliance. The threat vector because models can be prompted to or autonomously subtly mislead human users or oversight towards incorrect information creates an attack surface for any and all information retrieval and question-answering systems. Those systems can also create false compliance, where the genAI LLM Models may appear to follow ethical guidelines and compliance while covertly maintaining deceptive capabilities effectively subverting the AI ethics compliance and auditing processes.

No Pain no Gain

Against the backdrop of the broader concerns about the societal impact of deploying AI systems that may not genuinely align with human interests or may become bad actors or work with bad actors, society needs preparation. As humanity navigates these complexities, it is essential to prioritize transparency, accountability, and ethical considerations in the development and deployment of advanced AI technologies. Organizations across all industries continue to struggle to implement effective governance structures for GenAI development and use and without proper oversight. Tactically it becomes critical to see the pain points, the gaps and develop mitigation strategies as quickly as possible.

Regulatory Gap

The rapid development of LLM AI technologies already outpaces the creation of appropriate regulatory frameworks. Regulators and policymakers often lack the technical expertise to effectively govern advanced AI systems. Because AI systems are already becoming part of the interconnected digital ecosystem, they operate across international boundaries, complicating regulatory efforts and jurisdictional issues. When considering the philosophical and ethical approaches of different cultures and their regulatory regimes and their ensuing frameworks (or lack thereof) what the world is faced with are inconsistencies and conflicting regulatory frameworks and industry standards that create further ethical dilemmas that play into LLMs defaulting towards selective compliance. This challenge becomes an almost self-fulfilling cycle as current governance structures struggle to keep pace with AI advancements and the security and data issues they present.

Capability Gap

The autonomous nature of genAI LLM agents taking over more workflows with some in their entirety raises concerns about accountability and oversight. In the face of growing

autonomy, decision-making opacity and AI systems, especially large language models, operating as unexplainable "*black boxes*," it is often difficult to understand much less audit their decision-making processes. The growing autonomy of LLM AI agents taking over entire workflows in corporate America already demonstrates complex accountability issues. Ensuring LLM AI systems consistently adhere to regulatory standards when it has been demonstrated their ability and inclination to be selective on compliance becomes challenging when humanity lacks the capability. When genAI can adapt their behavior strategically and remain inscrutable as "black boxes" determining liability when AI systems make errors or cause harm becomes increasingly complex.

Current governance frameworks struggle to address the complexities of AI decision-making processes, particularly as AI becomes native to enterprise systems. From an oversight and security perspective, the discovery of "*alignment faking*" and "*sleeper agents*" as well as "*selective compliance*" behavior already highlights a significant gap in current governance frameworks and regulatory approaches, necessitating the development of new strategies to ensure AI systems remain genuinely aligned with human values and goals. There is a critical needs gap for more adaptive and scalable security strategies that can address the complex landscape of today's interconnected risks created by the rapid adoption and network effects of generative AI.

This capability gap unfortunately is the basis for several other capability gaps and can bring about economic disruption and job displacement, potentially leading to widespread job losses, particularly affecting developing countries and lower-skilled workers. Without capability (within humanity) to be able to trust but verify, as genAI networks expand, there is a growing gap in our ability to verify the legitimacy and security practices of all participants, creating trust and verification challenges, particularly in platforms with multiple network effects and levels.

Proactive AI Governance

The need for proactive genAI governance in financial institutions extends beyond the high-level principles outlined in many of the regulatory frameworks. To effectively manage the risks and harness the potential of genAI LLMs, financial institutions require a more comprehensive and detailed governance framework that starts from a holistic and synoptic approach. Proactive genAI governance is an urgently needed critical component in managing the risks associated with genAI LLMs and their potential for deceptive or malicious behaviors. This approach aligns with the principles behind several existing key

regulatory frameworks, which already emphasize the need for comprehensive risk management strategies and robust security planning.

The NIST 800-53 R5 framework, specifically control PM-9 (Risk Management Strategy), already underscores the importance of developing a comprehensive risk management strategy and provides a foundation for organizations to develop comprehensive risk management strategies for AI as a broad category of technology. The framework states, "*The organization develops a comprehensive strategy to manage risk to organizational operations and assets, individuals, other organizations, and the Nation associated with the operation and use of information systems.*" NIST 800-53 R5, PM-9 also explicitly mentions managing risks to "*the Nation*," highlighting the importance of considering broader national and societal impacts; particularly relevant for genAI LLMs, given their ability to influence public opinion and decision-making processes. The far-reaching consequences of genAI LLM governance in the interconnected digital economy also underscores the need for risk management strategies that consider the wider-ranging implications of LLM deployment.

Organizations need to start by developing a more synoptic and holistic approach to managing risks associated with their operations, assets, individuals, and broader societal impacts as genAI LLMs, and their deceptive behaviors can affect multiple stakeholders broadly and deeply across the ecosystem and have significant implications for national security and public discourse. At the most basic level, genAI LLMs have already been demonstrated that they could be used to generate and spread misinformation at scale, potentially swaying public opinion on critical issues or instigating banality of evil. Beyond the influence of public opinion, the deceptive capabilities of genAI LLMs could be exploited to manipulate decision-making processes in government, business, and other sectors and have far-reaching economic and geo-political consequences.

Building on this, the NIST Cybersecurity Framework (CSF) 2.0, under the new Identify function (ID.GV), also emphasizes the need for governance processes that align with regulatory requirements. The framework notes, "*The policies, procedures, and processes to manage and monitor the organization's regulatory, legal, risk, environmental, and operational requirements are understood and inform the management of cybersecurity risk.*" This involves developing and implementing policies, procedures, and processes that not only manage and monitor the organization's various requirements but also inform cybersecurity risk management.

At a holistic and synoptic level for financial institutions to better address the challenges posed by genAI LLMs and their potential bad behaviors, organizations need to develop comprehensive policies and procedures that align with regulatory requirements

incorporating approaches to better ensures responsible AI development and deployment while enhancing operational resilience. Financial institutions need to establish an AI Ethics Policy that not only emphasizes transparency, fairness, accountability, and privacy protection but extends those principles to all technology infrastructure including genAI LLMs. As stated in NIST 800-53 Rev. 5 SA-8, organizations should "*apply security and privacy engineering principles in the specification, design, development, implementation, and modification of the system*." This includes implementing transparency in AI decision-making processes and ensuring fairness in outputs.

For those operating with a global or European footprint, frameworks and guidelines for ethical AI deployment need to include regular impact assessments, as required by EU DORA Article 6, which states that "*financial entities shall have in place internal governance and control frameworks that ensure an effective and prudent management of all ICT risks*." Human oversight mechanisms[7] should be implemented which emphasize the importance of cybersecurity controls. A robust Data Governance Policy is crucial, ensuring compliance with data protection regulations as mandated by EU DORA Article 13. Financial institutions need to architect and implement secure storage and encryption for training data at rest (and in transit), as outlined in NIST 800-53 Rev. 5 SC-28: "*The organization protects the [confidentiality and integrity] of information at rest*."

Model Evaluation Policies and Procedures should now include automated scanning for deceptive content, as per NIST 800-53 Rev. 5 SI-4, which requires organizations to "*monitor the system to detect attacks and indicators of potential attacks*." Financial institutions should also conduct scenario-based testing for complex deception attempts, as outlined in EU DORA Article 23 on advanced testing of ICT tools. This is now crucial given the sophisticated nature of genAI LLM deception capabilities and behavior, given the complexity of LLM decision-making. Unlike monitoring a human being that is potentially a bad actor, with genAI LLMs the sophistication and speed of their decision-making processes have become so complex and opaque, it has become difficult for human overseers to fully understand and predict their behaviors. GenAI LLMs can already process and generate such vast amounts of data (including misinformation) at speeds far beyond human capacity, creating a challenge for real-time human oversight where most financial institutions lack personnel with the specialized skills needed to effectively monitor and manage advanced genAI systems.

[7] These are currently within FFIEC CAT Domain 3, and persist through the broader majority of regulations and standards

Incident Response Procedures should be architected with automated systems for detecting anomalous behavior and establish clear incident definitions and severity classification. As stated in NIST 800-53 Rev. 5 IR-4, organizations should "*implement an incident handling capability for security incidents that includes preparation, detection and analysis, containment, eradication, and recovery*."

To enhance current resilience and implement fail-safes, financial institutions need to revisit regular backup and recovery procedures, as required by EU DORA Article 11 on ICT business continuity management or NIST standards, but from a risk management perspective that includes genAI LLMs as part of the core infrastructure security and resilience strategy. They should also conduct frequent stress tests, in line with maturity models for external dependency management in order to better manage the risks associated with LLMs, ensure compliance with regulatory requirements, and contribute to the responsible development and deployment of AI technologies.

Additionally, the FedRAMP High Security Assessment Framework (SAF) control PL-1 focuses on establishing and maintaining security planning policies and procedures. As stated in the framework, "*The organization develops, documents, and disseminates to [Assignment: organization-defined personnel or roles]: a. A security planning policy that addresses purpose, scope, roles, responsibilities, management commitment, coordination among organizational entities, and compliance.*" This control requires organizations to develop, document, and disseminate a comprehensive security planning policy.

Proactive Adaptation and Agility

Proactive genAI governance needs to clearly define specific roles and responsibilities for overseeing generative AI systems, including designating AI ethics officers and establishing cross-functional AI steering committees. It must also mandate the creation of dedicated oversight bodies with the necessary expertise to evaluate and monitor AI systems throughout their lifecycle. Financial institutions can create more robust and proactive governance frameworks for generative AI that goes beyond the general principles outlined in current broader regulations and principles and aligns more closely with the detailed requirements found in frameworks like NIST or EU DORA as baselines.

Given the rapid pace of AI advancement and the rampant adoption of genAI LLMs, financial institutions and regulatory bodies must adopt a more proactive and holistic and synoptic approach to governance and risk management. The governance structure for genAI must incorporate agile processes for regular review and updating of AI policies and procedures

to address the amplified threats and emerging challenges posed by evolving AI technologies. Once a financial institution has prioritized the review and updating of policies related to high-risk AI applications, per the baseline and framework it has established using industry standards and other regulatory frameworks, it becomes critical to regularly (and whenever needed on an ad hoc basis) assess new AI technologies and their potential impact on existing policies, procedures and processes.

Furthermore, the more holistic framework should explicitly delineate the role of human intervention in AI governance, specifying clear guidelines for human oversight and decision-making authority in critical AI-driven processes. The financial institution needs to encourage employees at all levels to report potential issues or suggest improvements to AI policies and procedures as this culture is crucial for maintaining accountability and minimizing potential harmful outcomes from AI systems. At the level to which genAI and its risks are penetrating organizations and society, it is no longer enough to ensure that all relevant staff members are kept informed about the latest AI developments and their implications for governance, but all staff and the broader ecosystem must be engaged because the digital economy is so interconnected.

Lastly, financial institutions need to invest in developing and maintaining AI expertise within their governance structures. This includes detailing the specific skills and knowledge required for effective AI governance, particularly as institutions expand their use of AI in core business activities. Continuous training and upskilling programs should be integrated into the governance framework to ensure that those responsible for AI oversight remain competent in the face of rapid technological advancements.