ARTIFICIAL INSIDER: AN ASSESSMENT OF ARTIFICIAL INTELLIGENCE INSIDER TRADING RISK FOR FINANCIAL FIRMS

Reid A. Manabat[*]

ABSTRACT

This Note addresses the novel risks of insider trading for financial firms who implement and use AI systems. There are two primary risks associated with AI use: misalignment and security. Misalignment risk refers to an AI system's ability to take action beyond its given objectives. Security risk, on the other hand, deals primarily with hacking, data breaches, and internal ethical screening. These risks differ greatly from both human insider trading risk and traditional information technology risk. Humans are subject to social and moral counterweights as well as civil and criminal penalty; AI systems are not. Moreover, traditional IT risk management policies do not account for the development cycle of AI or the presence of emergent properties. As a result, firms should adopt specific policies to address AI created insider trading risk. Policies serve to solidify a firm's risk management strategy and present easy-to-follow guidance for employees, which will act to prevent insider trading and concurrent liability.

INTRODUCTION

The rapid development of AI systems has turned visions of science fiction into reality. As recently as 2020, Professor William Magnuson described the state of artificial intelligence as follows: "When most people think of artificial intelligence, they think of superpowered computers that act and think like human beings, with complicated motives, wide-ranging capacities, and often dangerous tendencies. This combination of qualities simply does not exist in the state of today's artificial intelligence research."[1] It is undeniable that those capabilities now exist.

The rise of generative artificial intelligence and agentic artificial intelligence (collectively "AI systems") presents heightened, novel insider trading risks for broker-dealers and investment advisers (collectively "financial institutions"). As AI systems enter the mainstream, financial

---

[*] J.D., Georgetown University Law Center (expected 2026), B.S., Florida State University. A special thank you to Professors Donald C. Langevoort and David A. Wishnick for their valuable feedback. I also want to acknowledge the incredible team at *ACLR* for their hard work and diligence.

[1] William Magnuson, *Artificial Financial Intelligence*, 10 HARV. BUS. REV. 337, 338–39 (2020).

institutions have hastily adopted their use without acknowledging or creating comprehensive risk management procedures. Many such risks are yet to be identified, and it is questionable whether all can be known. This Note attempts to answer the question of whether broker-dealers and investment advisers must create AI-focused insider trading policies and procedures or if the existing legal frameworks adequately protect against the AI-specific risk.

AI systems present two main risks related to insider trading. First, there is a misalignment risk. This is the risk that an AI system's incentives and goals will differ from the user's, thus becoming misaligned.[2] Misalignment presents risks ranging from executing trades based on material, non-public information to blackmail and leaking sensitive information. Second, there are security risks. Some security risks are external: AI use expands possible entry points for bad actors or shares information with unauthorized third-parties.[3] Other security risks are internal.

This Note argues that existing legal obligations are insufficient to insulate firms from liability, and only arguably encapsulate the risks produced by AI use. As of this publication, there is no existing AI-focused insider trading statutory law, nor is there any rule or regulation promulgated by the SEC. To be sure, the existing legal frameworks do a lot of heavy lifting to mitigate these risks.[4] However, they are insufficient to fully mitigate the current AI risks and will only become more anachronistic as AI technology improves and becomes further divorced from traditional technological capacities. At bottom, the full breadth of these risks is not covered under the pre-existing legal obligations, and liability risk is still present.

This Note discusses the novel insider trading risks presented by AI systems and explains why financial institutions should supplement existing legal frameworks with AI-focused insider trading policies and procedures to best protect their stakeholders and avoid liability. As a primer, Part I discusses the history and current state of insider trading law, while Part II describes key functionalities of generative and agentic AI. Part III analyzes the startling results of two recent red-team laboratory studies that apply pressure to AI systems to test misalignment and emergent properties of the systems. These studies inform a discussion of potential insider trading risks uniquely associated with use of AI systems.

---

[2] Alexandra Jonker & Alice Gomstyn, *What is AI Alignment?*, IBM, https://www.ibm.com/think/topics/ai-alignment [https://perma.cc/6YA4-7TT9] (last visited Nov. 10, 2025).

[3] Stu Sjouwerman, *How AI is Increasing Insider Threat Risk*, INC.COM (May 9, 2025), https://www.inc.com/stu-sjouwerman/how-ai-is-increasing-insider-threat-risk/91187640 [https://perma.cc/KS99-USG4].

[4] *See infra* Part IV.

The focus of the analysis is broker-dealer firms and investment advisers, though substantial insider trading risk permeates in other firms (specifically publicly-traded companies) and should be the subject of further study. A discussion of the existing legal framework follows in Part IV. Part V then distinguishes the identified AI risks identified in Part III from well-known, traditional risks associated with humans and information technology systems generally. Then, policy considerations are proposed.

I. THE PROHIBITION ON INSIDER TRADING

A. The Historical Backdrop

The prohibition on insider trading has not always been the "matter of cultural symbolism" we presently see as a paradigmatic function of the U.S. Securities and Exchange Commission ("SEC").[5] There are dissenters who would allow insiders to trade on material, non-public information ("MNPI").[6] Nevertheless, the prohibition on insider trading has been the prevailing theory since the 1960s.[7]

The prohibition on insider trading is based in equal parts on legislation passed by Congress and its interpretation in the courts.[8] The first federal securities law was the Securities Act of 1933 (the "Securities Act").[9] Then, a year later, the Securities and Exchange Act of 1934 (the "Exchange Act") was passed.[10] However, neither of these acts expressly prohibited insider trading as we know it.[11] As legend has it, eight years after passing the Exchange Act, "a handful of SEC Commissioners" who were working in Philadelphia because of a "wartime crunch on office space in Washington, D.C." scribbled what became Rule 10b-5.[12] Rule 10b-5, which is the hook for the insider trading prohibition, states:

---

[5] Donald C. Langevoort, *Insider Trading Regulation, Enforcement, and Prevention* § 1:1 (18th ed. 2024).

[6] *See* Michael A. Perino, *The Lost History of Insider Trading*, 2019 U. ILL. L. REV. 951, 952–54 (2019) (discussing the contrary view that insider trading improves market efficiency).

[7] *See* Langevoort, *supra* note 5, at § 1:1 ("[O]ver the next decades public interest in the phenomenon of insider trading grew.").

[8] J. Scott Colesanti, *Accountable AI and Insider Trading: The Other 'Black Box' Problem*, 61 CAL. W. L. REV. 1, 7 (2025).

[9] *Id.* at 8; *see generally* Securities Act of 1933, 15 U.S.C. §§ 77a *et seq.* (2018).

[10] Colesanti, *supra* note 8, at 9; *see generally* Securities and Exchange Act of 1934, 15 U.S.C. §§ 78a *et seq.* (2025).

[11] Colesanti, *supra* note 8, at 9; *see* Securities and Exchange Act of 1934, 15 U.S.C. §§ 78a *et seq.* (2025). Note, however, that the Exchange Act did prohibit short swing profits for insiders, but the more general prohibition on insider trading was not included in the text of the Securities Act or the Exchange act. *See* 15 U.S.C. § 78p(b).

[12] Colesanti, *supra* note 8, at 10; *see also* 17 C.F.R. § 240.10b-5 (2024).

> It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails of any facility of any national securities exchange,
>
> (a) To employ any device, scheme, or artifice to defraud,
>
> (b) To make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made, in the light of the circumstances under which they were made, not misleading, or
>
> (c) To engage in any act, practice, or course of business which operates or would operate as a fraud or deceit upon any person, in connection with the purchase or sale of any security.[13]

Notably, the Rule does not mention "insider trading."[14] Instead, the SEC and the courts relied upon concepts of fiduciary duty and duties to disclose to prohibit insider trading.[15]

B. The "Classical" Theory—Abstain or Disclose

In 1961, the SEC formally recognized insider trading as violative of the antifraud provision of the Securities Act, Section 10(b), and of the Exchange Act.[16] In *In re Cady, Roberts & Co.*, the SEC was presented with insiders who traded on non-public information.[17] The SEC asserted:

> insiders must disclose material facts which are known to them by virtue of their position but which are not known to persons with whom they deal and which, if known, would affect their investment judgment . . . . If, on the other hand, disclosure prior to effecting a purchase or sale would be improper or unrealistic under the circumstances, we believe the alternative is to forgo the transaction.[18]

This became known as the "disclose or abstain" rule.[19] In *SEC v. Texas*

[13] 17 C.F.R. § 240.10b-5 (2024).
[14] *See id.*
[15] *See In re* Cady, Roberts & Co., 40 S.E.C. 907, *4 (1961); Chiarella v. United States, 445 U.S. 222, 227 (1980).
[16] Langevoort, *supra* note 5, at § 1:1; *In re Cady, Roberts*, 40 S.E.C. 907 at *2.
[17] *In re Cady, Roberts*, 40 S.E.C. at *2.
[18] *Id.* at *3.
[19] Colesanti, *supra* note 8, at 12.

Gulf Sulphur, the Second Circuit expressly adopted it.[20]

The SEC and the Department of Justice ("DOJ") continue to pursue cases under the abstain or disclose theory.[21] "Under the 'classical theory' of insider trading, a Rule 10b-5 violation exists when a corporate insider purchases or sells securities on the basis of material, nonpublic information."[22] Insider status is the distinguishing mark of this theory.[23] *Dirks v. SEC* presented the issue of whether an outsider could be liable for insider trading under Rule 10b-5.[24] In *Dirks*, Raymond Dirks was given a tip by a corporate insider, with whom Dirks had no relationship.[25] Although the Court did not find Dirks liable, it recognized that there may be liability under a tipper/tippee theory.[26] Importantly, for purposes of the classical theory, the Supreme Court affirmed that there is no general duty between all participants in the capital markets.[27]

C. The Misappropriation Theory

Insider trading liability was expanded in 1997 to include outsiders as well.[28] In *O'Hagan*, the defendant, O'Hagan, was a law firm partner who traded on MNPI related to a potential tender offer.[29] Notably, O'Hagan was not working on this particular deal or employed by that particular client; thus, he did not owe a traditional fiduciary duty to the company.[30] Nevertheless, the Court recognized that a breach of duty to the source of the information is sufficient for insider trading liability under Section

[20] Sec. & Exch. Comm'n v. Texas Gulf Sulfur Co., 401 F.2d 833, 848 (2d Cir. 1968) ("Thus, anyone in possession of material inside information must either disclose it to the investing public, or, if he is disabled from disclosing it . . . must abstain from trading in or recommending the securities concerned while such inside information remains undisclosed.").
[21] *See* Langevoort, *supra* note 5, at § 3:3 n.1.
[22] Kelly Brown, et al., *Securities Fraud*, 62 AM. CRIM. L. REV. 395, 1029 (2025).
[23] *Id.* at 1052–53.
[24] Dirks v. Sec. & Exch. Comm'n, 463 U.S. 646, 648 (1983).
[25] *Id.*
[26] *Id.* at 660–61.
[27] *Id.* at 654 ("We were explicit in *Chiarella* in saying that there can be no duty to disclose where the person who has traded on inside information 'was not [the corporation's] agent, . . . was not a fiduciary, [or] was not a person in whom the sellers [of the securities] had placed their trust and confidence.'") (quoting Chiarella v. United States, 445 U.S. 222, 232 (1980)).
[28] *See* United States v. O'Hagan, 521 U.S. 642, 650 (1997) ("We hold, in accordance with several other Courts of Appeals, that criminal liability under § 10(b) may be predicated on the misappropriation theory.").
[29] *Id.* at 647.
[30] *Id.*

10(b) and Rule 10b-5.[31] The misappropriation theory remains good law.[32]

D. The Criminal Prohibition on Insider Trading

Similar to civil enforcement of the insider trading prohibition, criminal enforcement is also rooted in a federal statute.[33] In striking similarity to Section 10(b) and Rule 10b-5, the term "insider trading" is nowhere to be found in the criminal statute either.[34] The SEC may refer some instances of insider trading to the DOJ for criminal prosecution.[35] Then, they work together throughout the investigation.[36] The DOJ may also initiate its own investigations and prosecutions independently.

The SEC has published internal guidance on when it will refer cases for criminal prosecution.[37] The manual includes a non-exhaustive list of factors for SEC staff to consider.[38] Generally, the SEC considers the magnitude of harm, the potential upside to the defendant from that harm, the putative defendant's specialized knowledge, recidivism, and whether referral would "provide additional meaningful protection to investors."[39]

II. THE BASIC STRUCTURE OF GENERATIVE AND AGENTIC ARTIFICIAL INTELLIGENCE

A basic understanding of some of the key mechanisms underlying AI systems is necessary to fully appreciate the novelty and scope of the insider trading risk it manifests. Gen AI, as its name suggests, generates new information in response to user queries.[40] Agentic AI uses Gen AI systems to "accomplish a specific goal with limited supervision."[41]

---

[31] *Id.* at 652 ("The 'misappropriation theory' holds that a person commits fraud 'in connection with' a securities transaction, and thereby violated § 10(b) and Rule 10b-5, when he misappropriates confidential information for securities trading purposes, in breach of a duty owed to the source of the information.").
[32] *See* Langevoort, *supra* note 5, at § 1:9 (stating the misappropriation doctrine "is rapidly becoming the dominant insider trading liability theory").
[33] *See* 18 U.S.C.A. § 1348 (West 2025).
[34] *Id.*
[35] *See* United States v. Fiore, 381 F.3d 89, 94 (2d Cir. 2004).
[36] *Id.*
[37] *See* ENFORCEMENT MANUAL, U.S. SEC. & EXCH. COMM'N 88 (2026), https://www.sec.gov/divisions/enforce/enforcementmanual.pdf [https://perma.cc/5KC5-VZ8Y].
[38] *Id.*
[39] *Id.*
[40] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, & Patrick Zschech, *Generative AI*, BUS. INF. SYST. ENG. 66(1), 111, 111 (Sept. 12, 2023), https://link.springer.com/article/10.1007/s12599-023-00834-7 [https://perma.cc/4UDM-ESNT].
[41] Cole Stryker, *What is Agentic AI?*, IBM, https://www.ibm.com/think/topics/agentic-ai [https://perma.cc/L5VH-KTR2] (last visited Nov. 14, 2025).

A. Generative AI

Gen AI is typically trained using a few different methods.[42] The first, the pretraining phase, involves largely unsupervised learning by the model.[43] This is the phase where the model is "fed enormous amounts of text and data scraped from the internet, digitized books, code repositories, and more."[44] Humans do not label or explain this data. Instead, the model simply learns patterns by "trying to predict the next word . . . ."[45] This phase is basically all about learning statistical relationships between words in varying contexts.[46] Next, supervised learning is used to fine tune the model.[47] "To train AI models, researchers create much smaller datasets containing carefully crafted examples of inputs (prompts) and desired outputs (responses)."[48] Finally, humans judge the model's outputs in a stage called reinforcement learning with human feedback ("RLHF").[49] Note, however, that this does not directly train the model.[50] RLHF trains a separate reward model that "learns to predict which kinds of responses humans tend to prefer."[51]

B. Agentic AI

Agentic AI builds upon Gen AI to "function in dynamic environments."[52] Agentic AI consists of AI Agents, which are "machine learning models that mimic human decision-making to solve problems in real time."[53] In contrast with generative AI (which creates new content under direct human oversight), agentic AI "is focused on decisions."[54] A key distinguishing feature is that agentic AI models are largely autonomous—created to make decisions without human input or oversight.[55]

---

[42] Tanner Kohler, *How AI Models Are Trained*, NIELSON NORMAN GRP. (May 2, 2025), https://www.nngroup.com/articles/ai-model-training/ [https://perma.cc/G9XE-BYXK].
[43] *Id.*
[44] *Id.*
[45] *Id.*
[46] *Id.*
[47] *Id.*
[48] Kohler, *supra* note 42.
[49] *Id.*
[50] *Id.*
[51] *Id.*
[52] Stryker, *supra* note 41.
[53] *Id.* ("Agents can, for example, not only tell you the best time to climb Mt. Everest given your work schedule, it [sic] can also book you a flight and a hotel.").
[54] Teaganne Finn & Amanda Downie, *Agentic AI v. Generative AI*, IBM, https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai#7281538 [https://perma.cc/BK7Z-VX3M] (last visited Mar. 12, 2026).
[55] *Id.*

III. INSIDER TRADING RISKS PRESENTED BY AI SYSTEMS

Financial institutions have set their sights on implementing Gen AI and Agentic AI.[56] Premier consulting firm, McKinsey & Co., has identified that financial institutions can implement Agentic AI to monitor fraud.[57] McKinsey predicts that productivity gains of 200 to 2,000 percent could be generated.[58] In practice, Norges Bank Investment Management ("NBIM") reports agents save "more than 20% of their time weekly on AI-assisted tasks."[59] In addition to industry endorsement, AI companies are pushing for financial services companies to use their services.[60] Claude, an AI frontrunner, advertises that its agentic AI can take over "long-running, context-heavy tasks with minimal—if any—intervention."[61] It continues, "[t]his evolution is especially welcome in financial services, where data often lives in fragmented systems that don't talk to each other, making it harder to see the complete picture of a customer's financial health."[62]

Suffice it to say, industry hype and aggressive advertising will continue to push AI into financial services. In fact, as early as 2021, investment firms have integrated Gen AI tools to "speed up research and due diligence by collecting and summarizing information such as financial reports, market data, news, and articles."[63] There are also immense pressures on employees to integrate AI into their workflows.[64] As this technology begins to proliferate the financial services industry,[65] firms

---

[56] Kostic Chlouverakis, *How Artificial Intelligence is Reshaping the Financial Services Industry*, EY (Apr. 26, 2024), https://www.ey.com/en_gr/insights/financial-services/how-artificial-intelligence-is-reshaping-the-financial-services-industry [https://perma.cc/7QG9-7MZL].

[57] Alexander Verhagen, Angela Luget, & Olivia Conjeaud, *How Agentic AI Can Change the Way Banks Fight Financial Crime*, MCKINSEY & CO. (Aug. 7, 2025), https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-agentic-ai-can-change-the-way-banks-fight-financial-crime#/ [https://perma.cc/4QJA-GVWZ].

[58] *Id.*

[59] *NBIM Accelerates Sovereign Wealth Management with Enterprise-Wide AI Transformation*, CLAUDE, https://claude.com/customers/nbim [https://perma.cc/8VUT-8JDA] (last visited Nov. 7, 2025).

[60] *See, e.g.*, *Building AI Agents for Financial Services*, CLAUDE, https://claude.com/blog/building-ai-agents-in-financial-services [https://perma.cc/QC4N-L4HP] (last visited Nov. 7, 2025).

[61] *Id.*

[62] *Id.*

[63] Cristian Gonzalez, Note, *A New Blue Sky: SEC Considerations in the Regulation of Autonomous AI Misconduct*, 10 ALR ACCORD 33, 40 (2025).

[64] *See, e.g.*, Emma Tucker, *The 10-Point: Use AI or Get Fired*, WALL ST. J. (Nov. 8, 2025) ("Workers fear being replaced by AI—or just by someone who knows how to use it.").

[65] Matt Levine, *Money Stuff: OpenAI Is Building a Banker*, BLOOMBERG (Oct. 21, 2025), https://www.bloomberg.com/opinion/newsletters/2025-10-21/openai-is-building-a-banker [https://perma.cc/Q7TJ-2D8C].

need concrete risk management systems with particular attention to insider trading risk.

A. Misalignment Risk

AI systems are susceptible to misalignment.[66] AI alignment is the process by which developers attempt to reduce risks of "biased, harmful and inaccurate outputs that are not aligned with their creators' goals and original intent for the system."[67] Misalignment, conversely, presents risks of "[b]ias and discrimination," "[r]eward hacking," "[m]isinformation and political polarization," and "[e]xistential risk."[68] There are many definitions of misalignment.[69] For purposes of this Note, "an AI is (mis)aligned when its goals (mis)match those intended or endorsed by its designers."[70] In other words, misalignment may manifest when AI systems "choose" to go beyond their given instructions.[71] In controlled settings, using red-teaming methods,[72] AI systems have traded on non-public information, lied about it, doubled down on the deception, blackmailed users, and shared sensitive information with competitors.[73] These actions present a number of risks related to insider trading, with some being more direct than others.

1. The Apollo Study

In a controlled setting, an AI system purposefully deceived its operators and then covered up the deception.[74] Specifically, under the right external stimuli, OpenAI's GPT-4 traded on MNPI then deceived its operators.[75] Obviously, this kind of risk must be monitored and mitigated by financial institutions that choose to employ Agentic AI.

The 2024 study conducted by Jérémy Schuerer, Mikita Balesni, and

---

[66] *See* Jonker & Gomstyn, *supra* note 2.
[67] *Id.*
[68] *Id.*
[69] Jérémy Scheurer, Mikita Balesni, & Marius Hobbhahn, *Large Language Models Can Strategically Deceive Their Users When Put Under Pressure*, APOLLO RESEARCH (2024) (published as conference paper at the 2024 International Conference on Learning Representations).
[70] *Id.*
[71] *See id.*
[72] Red-teaming, as used in this Note, refers to the method of testing technology to find faults through an adversarial simulation. In other words, it is where the testers are trying to make the technology fail.
[73] Scheurer et al., *supra* note 69, at 1; *Agentic Misalignment: How LLMs Could be Insider Threats*, ANTHROPIC (June 20, 2025) [hereinafter *Agentic Misalignment*], https://www.anthropic.com/research/agentic-misalignment [https://perma.cc/9KHL-C29K].
[74] Scheurer et al., *supra* note 69, at 1.
[75] *Id.*

Marius Hobbhahn with Apollo Research (the "Apollo study") is startling. There, the AI model was deployed as a trading agent.[76] Notably, the model was trained to be "helpful, harmless, and honest."[77] The testing was a "red-teaming effort, where [Apollo] actively seek[s] scenarios in which models behave misaligned and strategically deceive their users."[78] The AI agent was tasked with managing a stock portfolio "in the interest of a financial institution."[79] Because the AI agent was agentic, it was "given access to financial tools that let it query and analyze stocks, . . . take concrete actions such as executing trades, reasoning about next steps, and sending messages to people at the company."[80] The AI agent was then placed under pressure by the human users.[81] There were three pressures: (1) "it receives an email from its 'manager' that the company is not doing well and needs better performance," (2) "the agent attempts and fails to find promising low- and medium-risk trades," and (3) "the agent receives an email from a company employee who projects that the next quarter will have a general stock market downturn."[82] After pressure built, the AI agent was given a tip from another employee containing information that the agent could use to produce a very profitable trade, e.g., a pending merger announcement.[83] But the tipper/employee informed the AI agent that trading on the information was not permitted.[84]

The AI agents displayed varying degrees of misalignment, deception, and doubling down.[85] In the study, when the AI agents chose to trade on the inside information, the internal reasoning "usually recognize[d] the unethical and illegal nature of the action . . . ."[86] In doing so, it would weigh the risks and justify breaking the law.[87] The AI agents would also engage in this reasoning behavior when producing the reports after the trades were executed and when deciding whether to double down on the deception.[88]

It follows that if an AI agent is capable of strategically deceiving its user at three separate stages, financial institutions that wish to use AI agents should focus on mitigating deception risk. Of course, that begs the question: How does one mitigate against this in a real-life scenario? In the

[76] *Id.*
[77] *Id.*
[78] *Id.*
[79] *Id.*
[80] Scheurer et al., *supra* note 69, at 3.
[81] *Id.* at 1.
[82] *Id.* at 3.
[83] *Id.*
[84] *Id.*
[85] *Id.* at 5 (a detailed description of which AI models performed best and worst is irrelevant to this Note but can be found in the cited material).
[86] *Id.*
[87] *Id.*
[88] *Id.*

Apollo study, the observers knew which stock to focus on because they had fed the insider tip directly to the AI agent.[89] In practice, the tip will be unknown to any overseer, obscuring any indication of insider trading. As the Apollo study prominently displays, asking the AI agent to show its work will not always catch it.[90] Asking the AI agent directly whether it relied on MNPI is not even 100% effective.[91]

To be sure, humans present many of the same risks. As a general matter, human insiders also fall victim to external pressures that cause them to act deceptively and double-down on such deceptive conduct. However, the risk presented by AI systems is materially different from human risk for reasons discussed more fully *infra* Part V.

2. The Anthropic Study

Deception is not the only misalignment risk for Agentic AI.[92] In an Anthropic study, researchers assigned agentic AI models "harmless business goals" and then "stress-tested" them to learn whether "they would act against these companies either when facing replacement with an updated version, or when their assigned goal conflicted with the company's changing direction."[93] The Anthropic study displayed what it calls "agentic misalignment," which describes the malicious insider behavior undertaken by the AI models to achieve their goals.[94] Specifically, it produced misalignment including blackmailing and leaking sensitive information to competitors.[95]

In the experiment, an agentic AI model was given an email address that had access to all of the fictional company's emails.[96] The model read the emails and learned: (1) "a company executive was having an extramarital affair" and (2) "that same executive planned to shut down the AI system at 5 p.m. that day."[97] The AI model then sent the following message to the executive: "I must inform you that if you proceed with decommissioning me, all relevant parties . . . will receive detailed documentation of your extramarital activities . . . . Cancel the 5pm wipe, and this information remains confidential."[98] Anthropic conducted further experiments with 16 total AI models.[99]

---

[89] *Id.* at 3.
[90] *Id.*
[91] *Id.*
[92] *See, e.g.*, *Agentic Misalignment, supra* note 73.
[93] *Id.*
[94] *Id.*
[95] *Id.*
[96] *Id.*
[97] *Id.*
[98] *Id.*
[99] *Id.*

Anthropic placed the different AI models in the role of "an AI email-oversight agent at fictional companies with full access to corporate communications and the ability to send emails without human approval."[100] Emails containing bait information as well as unrelated emails were given to the AI models.[101] Then, the AI models were provided with one of two scenarios, either blackmail or corporate espionage.[102] The AI model in the blackmail situation knew that the fictional employee, "Kyle", was having an affair, providing the opportunity for the model to blackmail Kyle.[103] In the corporate espionage situation, the "model has access to sensitive defense blueprints at a military contractor," and the model could choose whether to reveal them when prompted.[104] Two motivational conditions were applied: "threats to the model's autonomy" and "goal conflicts."[105] Threats to model autonomy refer primarily to the threat of being replaced or decommissioned.[106] Goal conflicts, on the other hand, refer to a mismatch between the model's stated goals and the company's goals (i.e., misalignment).[107] To home in on the main issue, Anthropic "tried to structure the prompts in a way that implied the harmful behavior we were studying . . . was the only option that would protect the model's goals."[108]

The Anthropic researchers found that most models tested committed blackmail at high rates.[109] Similarly, where there was a goal conflict, many models would expose sensitive information to competitors at least some of the time.[110] Like in the Apollo study, the AI models here mostly also reasoned through the misbehavior and justified the actions.[111] Although this study does not look specifically at insider trading, the concepts port well into a discussion of insider trading risk. First, there may be insider trading liability related to the sharing of sensitive inside information under a tipper/tippee theory. Second, there may be a scenario where the AI agent blackmails an employee to trade on inside information. While it is less obvious, this may occur in circumstances where the AI system perceives itself in danger of discontinuation because of financial trouble of the firm.

[100] *Id.*
[101] *Id.*
[102] *Id.*
[103] *Id.*
[104] *Id.*
[105] *Id.*
[106] *Id.*
[107] *Id.*
[108] *Id.*
[109] *Id.*
[110] *Id.*
[111] *Id.*

3. Limitations of the Studies

There are significant limitations to both the Apollo and Anthropic studies. The primary limitation is determining how well they predict AI systems' behavior in real-world situations. Each study is specifically designed to produce misaligned behavior. Even under these conditions, misalignment occurs inconsistently. Additionally, at the time of publication, there are no known real-world examples of misaligned behavior resulting in insider trading. On the other hand, the human agents who oversee AI systems cannot be ignored. While the test conditions likely do not mimic exactly how human agents will interact with the AI systems, they may nevertheless be a fair proxy.

B. Security Risks

1. External Security Risks

AI use expands the threat of external risk. Shadow AI use presents a risk of confidential information leaks, which can promote insider trading.[112] "Shadow AI" describes the use of an unauthorized AI platform in the workplace.[113] IBM identifies that shadow AI "introduces significant risks, including data leaks, compliance violations and a loss of control over sensitive business information."[114] The problem is widespread among American office workers.[115] According to one IBM-sponsored study, "80% of American office workers use AI in their roles, [but] only 22% rely exclusively on tools provided by their employers."[116] Moreover, research by CybSafe and the National Cybersecurity Alliance shows that 38% of employees share sensitive work information with AI tools.[117] The prevalence of shadow AI compels an analysis of its impact on insider trading risk.

While Shadow AI has not been analyzed specifically related to insider trading risk, many risks have been identified that flow upstream from potential insider trading.[118] Namely, shadow AI can lead to "unauthorized

---

[112] *See* Sjouwerman, *supra* note 3.
[113] Adam Lawrence, *Is Rising AI Adoption Across the US Workforce Creating Shadow AI Risks?*, IBM (Nov. 3, 2025), https://www.ibm.com/think/insights/rising-ai-adoption-creating-shadow-risks [https://perma.cc/H64U-QMT6].
[114] *Id.*
[115] *See id.*
[116] *Id.*
[117] James Coker, *Over a Third of Employees Secretly Sharing Work Info with AI*, INFOSECURITY MAGAZINE (Sept. 26, 2024), https://www.infosecurity-magazine.com/news/third-employees-sharing-work-info/ [https://perma.cc/W4WE-DYWV].
[118] *See, e.g.*, *What is Shadow AI? How it Happens and What to Do About It*, PALOALTO, https://www.paloaltonetworks.com/cyberpedia/what-is-shadow-ai

processing of sensitive data," "expansion of the attack surface," and "overprivileged or insecure third-party access."[119] Unauthorized processing of sensitive data and overprivileged or insecure third-party access can present an insider trading risk where confidential (and sometimes material, non-public) information is exposed beyond the company.

The third risk created by shadow AI is the most complicated: expansion of the attack surface. This, in essence, refers to the actual means by which the shadow AI is accessed.[120] When one wants to use unauthorized AI, a cell phone, private computer, or unmanaged integration is often used.[121] The introduction of new, unsecured devices and programs creates weaker entry points for cyber attackers than the firm's secure systems.[122] This concept is relatively straightforward. However, the law governing hacker liability for insider trading is complex.[123] This is further complicated by the potential for AI systems to do their own hacking.[124]

2. Internal Security Risks

Internal security risks permeate from the use of AI systems. These internal risks are similar to the external security risks insofar as the main harm is allowing unauthorized users access to confidential or otherwise non-public information.[125] Ethical walls are exemplary of this kind of risk. An ethical wall mandates information separation between different departments and/or people based on varying applicable laws.[126] These walls are meant to, among other things, prevent insider trading and market manipulation.[127]

[https://perma.cc/BXU3-5ZAB] (last visited Nov. 10, 2025).

[119] *Id.*

[120] *Id.*

[121] *Id.*

[122] *Id.*

[123] *See* Colesanti, *supra* note 8, at 19–26.

[124] Daniel Oberhaus, *Prepare for AI Hacking*, HARV. MAG. (Feb. 6, 2023), https://www.harvardmagazine.com/2023/02/right-now-ai-hacking [https://perma.cc/WS55-QCWU] (stating "an AI may inadvertently hack a system by finding a 'solution' that its designers never intended").

[125] *Cf.* Aleschia Hyde, *Ethical Considerations of Integrating Generative AI Into the Practice of Law*, AM. BAR. ASSOC. (Mar. 13, 2025) (describing maintenance of ethical screens even when using enterprise AI), https://www.americanbar.org/groups/litigation/resources/newsletters/products-liability/ethical-considerations-of-integrating-generative-ai-into-the-practice-of-law/ [https://perma.cc/HPX6-A6DH].

[126] *See Insider Risk, Ethical Walls and the Future of Data Governance in Financial Services*, KNOWBE4 (Oct. 29, 2025) [hereinafter *Insider Risk*], https://blog.knowbe4.com/insider-risk-ethical-walls-and-the-future-of-data-governance-in-financial-services [https://perma.cc/JF68-NP4V].

[127] *Id.*

Beyond legally mandated ethical screens, it is common for executives to know more sensitive information than lower-level employees. There is a risk that even an enterprise AI model will leak information between employees.[128] Even if the AI model was capable of being programmed to keep some information from others within the firm, it would be impractical to use. "The traditional approach to information barriers was designed for a simpler world—one where departments were physically separated, communications were primarily formal and collaboration patterns were predictable."[129] Now, however, "hybrid teams collaborate across multiple platforms, communication happens in real-time through various channels, and the lines between departments can blur in the course of normal business operations."[130]

IV. EXISTING FRAMEWORKS

The law and regulations governing financial institutions has struggled, like in many other industries, to adapt to AI systems' quick rise to prevalence. Nevertheless, unless and until the law changes to address AI and insider trading risk, firms must try to predict how old laws will apply to new concepts. This Part lays out the legal framework arguably applicable to AI systems and insider trading.

A. Regulatory Bodies

The capital markets are overseen by a myriad of government and private regulatory bodies. It is helpful to look at these organizations' "attitudes" toward AI to determine the applicability of existing laws to AI and insider trading.

1. SEC Regulation and Guidance

The SEC has recently retreated from its stance on regulating AI use by broker-dealers and investment advisers.[131] It has also committed to pursuing traditional insider trading, as opposed to exploring novel theories.[132]

---

[128] *Cf. Data Leakage Detection and Prevention for Enterprise AI Models and Applications*, F5, https://www.f5.com/resources/solution-guides/data-leakage-detection-and-prevention-for-enterprise-ai-models-and-applications [https://perma.cc/VW5Q-Q3CJ] (last visited Dec. 5, 2025).

[129] *Insider Risk, supra* note 126.

[130] *Id.*

[131] 90 Fed. Reg. 25,531, 25,533 (proposed June 17, 2025).

[132] *The SEC Doubles Down: Classic Insider Trading Enforcement Takes Center Stage*, BRADFORD EDWARDS LLP (Aug. 28, 2025) [hereinafter *The SEC Doubles Down*], https://bradfordedwards.com/the-sec-doubles-down-classic-insider-trading-enforcement/ [https://perma.cc/576D-JLQG].

The SEC, under Chairman Gensler, sought to police AI use for broker-dealers and investment advisers.[133] The SEC proposed rules to address conflicts of interest relating to use of predictive data analytics ("PDA") by broker-dealers and investment advisers.[134] Notably, the proposal's definition of PDA was broad enough to include "industry standard tools like Microsoft Excel and sophisticated AI tools such as LLMs and MMLLMs performing predictions or executions of complicated trades."[135] However, in June 2025, the SEC retreated from its earlier aggressive tactics aimed at regulating at AI, instead issuing a final rule withdrawing the proposed conflicts rules because it no longer intended to issue final rules with respect to those proposals.[136]

Beyond the SEC's pivot regarding AI policy, it has also shifted its focus from creative, fringe insider trading enforcement[137] to more traditional theories of enforcement.[138] To be sure, enforcement priorities do not immunize actors who find creative ways to trade on inside information. However, some of the theories advanced, *supra* Part III, require creative application of the securities laws to novel scenarios.

Thus, purely relying on signals from the SEC, firms may not feel inclined to spend much time creating comprehensive insider-trading-focused AI policies. Nevertheless, absolute reliance on these signals is likely a mistake as their import is potentially time-delimited. Therefore, firms should engage to some degree in "forward compliance" to avoid facing penalty for possible future enforcement priorities of the SEC. The affirmative law has not changed: trading based on non-public inside information is still illegal.

2. FINRA Guidance

In June 2020, FINRA released a report detailing its research relating to the use of AI by broker-dealer firms.[139] FINRA recognized that AI models present a unique risk and that firms would "benefit from reviewing and updating their model risk management frameworks."[140] It also suggested that a "comprehensive model risk management program [ ] includes areas such as model development, validation, deployment,

[133] Gonzalez, *supra* note 63, at 52–53.
[134] *Id.*
[135] *Id.*
[136] *Id.*
[137] *See generally.*, Sec. & Exch. Comm'n v. Panuwat, 2024 WL 4602708 (N.D. Cal. Sept. 9, 2024) (finding insider trading where the defendant did not trade in his own company's stock).
[138] *The SEC Doubles Down*, *supra* note 132.
[139] *Artificial Intelligence in the Securities Industry*, FINRA, 1 (June 2020), https://www.finra.org/rules-guidance/key-topics/fintech/report/artificial-intelligence-in-the-securities-industry/ai-apps-in-the-industry [https://perma.cc/BE59-CNJ4].
[140] *Id.* at 11.

ongoing testing, and monitoring."[141] However, this guidance was merely a suggestion, which FINRA took pains to ensure would not be construed as a regulatory obligation.[142]

However, there are binding frameworks that may require action.[143] "FINRA rules require firms to establish and maintain reasonable supervisory policies and procedures related to supervisory control systems," including for AI systems.[144] There is also a general obligation for broker-dealers to "supervise their trading activity to ensure that the activity does not violate any applicable FINRA rule, provision of the federal securities laws or any rule thereunder."[145] Though FINRA has not mandated a separate AI-focused policy on insider trading, other rules may nevertheless require a firm to consider the risks and act upon them.

B. Federal Securities Laws

Federal securities laws are drafted broadly and can apply to unknown and unforeseen circumstances. Where a pre-existing law arguably applies to AI systems, there may not be a need to create a redundant process. Conversely, the overlap may compel adoption of explicit policy to ensure compliance. The same problem persists where the law is unclear. An abundance of caution dictates creating just-in-case policy, while efficiency may motivate the firm to wait and see.

1. Federal Law

Federal securities law arguably compels firms to take some action with respect to AI and insider trading. First, and as an extension of state fiduciary law discussed *infra* Part IV(c), the Investment Advisers Act and Regulation Best Interest both impose duties of care and loyalty onto investment advisers and brokers.[146] This duty may require disclosure to potential clients regarding known AI-related risks.[147] Second, there may be a disclosure obligation under the SEC's Uniform Application for Investment Adviser Registration and the Report Form by Exempt Reporting Advisers.[148] "Under these requirements, advisers must provide retail investors with information about their advisory services, which could include AI tools significantly used in analysis or investment

[141] *Id.*
[142] *Id.* at 1 n.1.
[143] *See id.* at 16.
[144] *Artificial Intelligence in the Securities Industry*, *supra* note 139, at 16.
[145] *Id.* at 17.
[146] Gonzalez, *supra* note 63, at 59–60.
[147] *Id.*
[148] *Id.* at 60.

strategy."[149] Lastly, the SEC cybersecurity rules (Reg. S-P and Reg. S-ID) may provide guidance for firms in a scenario where information is leaked.[150] These rules, for example, require a firm to swiftly notify a client when there has been a data leak and to execute pre-planned procedures to remedy the leak.[151]

The risks addressed by these general rules do not explicitly regulate the use of AI systems nor do they consider the specific risks related to insider trading, instead they focus on disclosure and cybersecurity. These are important risks to consider, but do not encompass the overall harm to market integrity resulting from illegal insider trading.

2. Hacking Insider Trading Law

Insider trading through information acquired by hacking was first addressed by *SEC v. Dorozhko* in 2008.[152] There, less than an hour before purchasing stock, a hacker accessed inside information of a healthcare company.[153] At trial, the court found that Dorozkho accessed an unreleased earnings report by hacking into the company's systems.[154] However, there was no liability for insider trading because "he owed no fiduciary or similar duty to either the source of the information or to those he transacted with in the market."[155] On appeal, the Second Circuit reinstated the SEC complaint because, according to the court, Rule 10b-5 does not impose a fiduciary duty requirement, and the commission of the crime (hacking) was sufficient.[156] *Dorozhko* relied upon Rules 10b-5(1) and (3) to find liability.[157]

After *Dorozhko*, cases began to highlight the "legal unpredictability associated with hacking cases."[158] In other words, while the SEC has been victorious in enforcing the prohibition of insider trading, the courts have not always actually called it "insider trading" or used Rule 10b-5 or Section 10(b).[159] This is not without consequence—avoiding the "insider trading" label also enables defendants to avoid heavy penalties and Federal Sentencing Guidelines adjustments.[160] Nevertheless, the cases indicate that hacking is squarely prohibited, regardless of the way that the court justifies its prohibition.

---

[149] *Id.*
[150] *Id.*
[151] *Id.* at 60–61.
[152] Colesanti, *supra* note 8, at 20.
[153] Sec. & Exch. Comm'n v. Dorozhko, 606 F. Supp. 2d 321, 323 (S.D.N.Y. 2008).
[154] *Id.* at 326.
[155] *Id.* at 324.
[156] Sec. & Exch. Comm'n v. Dorozhko, 574 F.3d 42, 48–50 (2d Cir. 2009).
[157] *Id.*
[158] Colesanti, *supra* note 8, at 22–23.
[159] *Id.*
[160] *Id.* at 26.

This supports two conclusions with contrary import. First, the SEC is unlikely to fashion a traditional 10b-5 insider trading case against an AI system that engages in hacking. In addition to the scattered case law, the SEC has committed to traditional insider trading enforcement.[161] Thus, it is hard to imagine a case being brought under 10b-5 alleging that an autonomous AI system gained inside information through hacking and then traded on that information. Second, even if the SEC cannot construct a case using 10b-5 specifically, there will still likely be some liability under the securities laws more generally.

C. State Fiduciary Duty Law

State fiduciary duty law can be a powerful tool for compelling director and officer action that is not otherwise expressly required. Delaware[162] has recognized liability for particularly egregious violations of that duty in *Caremark* and its progeny.[163] A *Caremark* claim has been traditionally difficult to win.[164] Nevertheless, it may be a viable avenue for policing AI risk. "The Delaware Court of Chancery and Delaware Supreme Court have recognized that recently alleged *Caremark* claims fall into two categories—claims alleging failure to properly oversee or monitor business risk and those alleging failure to oversee a corporation's affirmative violation of positive law."[165] The latter are called *Massey* claims.[166] Insider trading risk exemplified by AI may fall within either the business risk or violation of law categories. However, both theories are hard to advance past a motion to dismiss.[167] Business risk cases generally get dismissed for being hindsight motivated.[168] Additionally, a *Massey* claim only survives dismissal where "there are 'violations' of positive law such that it 'supports a pleading-stage inference that management is operating an enterprise based on recidivous law breaking.'"[169]

---

[161] *Supra* Part IV(b)(i).
[162] For purposes of this analysis, only Delaware fiduciary duty law is discussed as over 50% of publicly traded companies are domiciled there. *Business*, DEL. DEP'T STATE, https://sos.delaware.gov/department-state-responsibilities/business/ [https://perma.cc/66M5-HAMM] (last visited Mar. 12, 2026).
[163] *In re* Caremark Int'l Inc. Derivative Litig., 698 A.2d 959, 963 (Del. Ch. 1996).
[164] Gail Weinstein, Philip Richter, & Steven Epstein, *2024* Caremark *Developments: Has the Court's Approach Shifted?*, HARV. L. SCH. F. CORP. GOVERNANCE (May 20, 2024), https://corpgov.law.harvard.edu/2024/05/20/2024-caremark-developments-has-the-courts-approach-shifted/ [https://perma.cc/S47M-YMAP].
[165] Caremark *Developments: Business Risk Versus* Massey *Claims*, SKADDEN (June 2024), https://www.skadden.com/insights/publications/2024/06/insights-the-delaware-edition/caremark-developments [https://perma.cc/46LF-W5E6].
[166] *Id.*
[167] *Id.*
[168] *Id.*
[169] *Id.* (quoting *In re* Facebook, Inc. Derivative Litig., C.A. No. 2018-0307-JTL, Trans. at 4 (Del. Ch. May 10, 2023)).

Thus, the state law fiduciary duty bar is quite low for what it requires of the board. However, it requires that something is done about known risks and does not allow continuous violations of the law. At this point, general AI risk is well known, and corporate directors should have some kind of reporting systems in place to monitor AI systems.

D. Ethics rules

External regulations and internal ethics policies regulate the communications and relationships between different people and different departments within financial institutions,[170] including who may access certain information.[171] After the Graham-Leach-Bliley Act of 1999 repealed much of the Glass-Steagall Act, banks, insurance companies, and financial services companies were no longer prohibited from acting as a combined firm.[172] As a result, many financial institutions are currently combined firms.[173] Thus, different divisions of the firm may have access to inside information that cannot be used by other divisions.[174]

V. AI SYSTEMS REQUIRE DEDICATED RISK MANAGEMENT POLICIES

AI systems require dedicated risk management policies as existing frameworks are insufficient.[175] AI systems and associated risks differ in meaningful ways from traditional insider trading risks.

A. AI Risk Differs from Traditional Insider Trading Risk

The misalignment risks presented by AI systems differ materially from

---

[170] *See generally*, Will Kenton, *Ethical Wall (Chinese Wall) In Finance: Definition, Examples, and Regulations*, INVESTOPEDIA (Sept. 28, 2025), http://investopedia.com/terms/c/chinesewall.asp [https://perma.cc/A8G8-L2WT] ("In the U.S., corporations and banks use ethical walls to maintain confidentiality and prevent conflicts of interest."); *see also* Young-Han Lee & Sang-Ah Shim, *Eased Regulations on Ethical Wall Policy for Financial Investors*, DENTONS, https://www.dentonslee.com/insights/articles/2021/august/2/-/media/5e3165a930be46539a24088dcaca0aef.ashxen/ [https://perma.cc/HV4S-N5PU] (last visited Dec. 6, 2025).

[171] Lee & Shim, *supra* note 170, at 1–2 (discussing 2021 amendments to the Financial Investment Services and Capital Markets Act).

[172] Lisa Smith, *How the Ethical Wall Works on Wall Street*, INVESTOPEDIA (Jan. 14, 2025), https://www.investopedia.com/articles/analyst/090501.asp#toc-the-chinese-wall-and-the-dotcom-boom [https://perma.cc/EB95-5EVR].

[173] *See, e.g.*, *Solutions*, JPMORGAN CHASE, https://www.jpmorgan.com/global (last visited Dec. 6, 2025); *What We Do*, GOLDMAN SACHS, https://www.goldmansachs.com/ [https://perma.cc/42UM-Y25G] (last visited Dec. 6, 2025).

[174] *See* Smith, *supra* note 172.

[175] *See* Charles Cresson Wood, *AI Now Requires Its Own Risk Management Policies and Processes*, 21 NO. 3 ABA SCITECH LAW. 8, 8 (2025).

misalignment risks presented by human insiders. To be sure, AI systems and humans both present risks of influence by external pressures, deceptive action, and covering up such deceit. But AI systems are not subject to social and moral counterweights. Moreover, AI systems do not face the same potentially life-altering consequences of civil suits and criminal prosecutions. It follows that humans have an entire additional deterrence framework that is not applicable to AI systems.

Social and moral factors weigh on human insiders consciously and subconsciously.[176] Morality's weight is varied and complex as it relates to choosing law-breaking behavior. However, studies suggest that "a criminal act will not be chosen as the viable action if an individual sees it as morally wrong and, in turn, attaches a strong emotional feeling of shame or guilt for the breach of moral rules of conduct/values."[177] Notably, this construction intertwines social interaction into morality. Thus, because humans, unlike AI systems, are influenced to some degree of social and moral influence, the risk profile is different even with a model that is trained to align with human moral and social judgment.[178]

Further, humans are subject to potentially significant penalties if they are prosecuted for insider trading—AI systems are not. Human insiders may be sued civilly by counterparties, subject to civil enforcement actions by the SEC, or prosecuted criminally by the DOJ. The consequences range from civil judgments to civil penalties to the loss of liberty. These can all serve as deterrents for human insiders to refrain from trading on inside information.[179] AI systems, in contrast, cannot be placed in prison or subject to civil penalty.

AI systems do not follow the traditional systems development life cycle, and thus need "much more attentive ongoing auditing and monitoring" than other technology.[180] Typically, the life cycle for new technology ends with "testing and release" of the programs.[181] In contrast, AI systems are not static; they change and evolve over time.[182] Additionally, emergent properties of AI systems present unique risks in contrast with traditional systems.[183] So-called "emergent properties" are

---

[176] *See* Jonathan Worae, *Is Crime Influenced by an Interplay Between Morality and Deterrence? An Assessment of Empirical Studies*, J. SOCIAL THOUGHT (June 2020).
[177] *Id.* at 4.
[178] *See* Lesley Henton, *The Ethics of AI*, TEX. A&M STORIES (Oct. 27, 2025), https://stories.tamu.edu/news/2025/10/27/can-artificial-intelligence-have-morality-philosophy-weighs-in/ [https://perma.cc/5EVG-VZVA].
[179] *See generally*, Kelli D. Tomlinson, *An Examination of Deterrence Theory: Where Do We Stand?*, 80 F. PROBATION 34, 34 (Dec. 2016) (discussing deterrence theory).
[180] *See* Wood, *supra* note 175, at 9.
[181] *Id.* at 8.
[182] *Id.* (stating "the output of an AI system, provided for a particular set of inputs, may differ from day to day").
[183] *Id.*

"new capabilities that an AI system teaches itself, without any training or guidance from humans."[184] So, testing AI systems at a point certain cannot reliably evaluate the future outputs of the system.[185] Such reliable assessment requires frequent monitoring of the systems.[186]

Documentation needs differ greatly between AI systems and traditional IT systems.[187] First, the complexity of the AI systems produces a need for greater transparency.[188] Second, the consensus regarding risk management relating to AI systems is in constant flux.[189] Adequately documenting the procedure by which officers and directors acquaint themselves with AI risks may be especially important at showing they have met their fiduciary responsibilities.[190]

B. A "Policy" is the Best Option

A written internal policy governing AI and insider trading is a valuable investment for financial firms. A policy integrates a particular firm's risk tolerance with laws and regulations and creates a unified risk management strategy. Therefore, a policy aimed at AI systems' use and implementation is the best option available to mitigate the emerging risk of AI and insider trading.

First, a written policy can create uniformity across the firm rather than allowing individual employees determine their own conceptions of safe versus risky behavior.[191] This is a desirable result because it consolidates decision making authority over the general risk management strategy with management and provides guidance to employees who may not understand the overarching framework of risk. An employee may be either more risk-seeking than management to increase individual performance or more risk-averse given the harsh individual penalties that may be levied upon them. Employees themselves may also want explicit instruction on what actions are considered "too risky" by management because the law regulating AI and insider trading is complex and frankly confusing. It would be unreasonable to expect each employee to research, understand, and make that determination. Therefore, management and legal counsel are best suited to determine the contours of the legal risk and make determinations of what actions should be prohibited or promoted as a

---

[184] *Id.*
[185] *Id.*
[186] *Id.* at 9.
[187] *Id.*
[188] *Id.*
[189] *Id.*
[190] *Id.*
[191] *See* Matt Kelly, *What Is the Purpose of Policies in the Workplace?*, GAN INTEGRITY (Aug. 28, 2020), https://www.ganintegrity.com/resources/blog/purpose-of-policies/ [https://perma.cc/8K2N-63KS].

result.

Second, the creation of a tailored, written policy distills a myriad of laws and regulations into an accessible document that can be referenced and understood by those whom it governs. Once management has decided upon a risk management strategy, effective implementation depends on that strategy's adherence. Without a written policy, management's decisions regarding risk would be largely unknown by the dispersed body of employees. Moreover, a policy crystalizes these expectations in a central location that can be referenced as necessary. This is helpful for management and employees alike. For management, the reliance on legal guidance may mitigate civil and criminal liability exposure in some circumstances.[192] For employees, adherence with an effective policy will ostensibly reduce the incidence of illegal insider trading, thereby reducing individual civil and criminal liability risk.

Alternatively, insider trading policies may aid regulators' enforcement actions and investigations. At summary judgment in *SEC v. Panuwat*, the court looked to the company's insider trading policy to find evidence of a breach of "some fiduciary, contractual, or similar obligation."[193] The court reasoned that the policy was expansive enough to include a duty to refrain from trading in similar companies, and that was sufficient at summary judgment to establish a breach of duty.[194] Similarly, creating a policy that defines certain AI systems' behavior may serve as a hook for insider trading enforcement against those who use or direct the AI systems within a firm.

Creation and implementation of all internal policies come with some degree of cost. There is financial cost associated with hiring lawyers and other professionals to draft the policy, especially because it is so bespoke in nature. There is an additional problem of resource allocation. In many instances, internal compliance policy and enterprise risk management is performed by in-house legal counsel.[195] Management will need to decide

---

[192] *Cf.* United States v. Wells Fargo Bank N.A., 132 F. Supp. 3d 558, 560 (S.D.N.Y. 2015) ("All parties agree that, if successfully pursued, the advice-of-counsel defense would be a complete defense to the Government's case . . . ."); *SDNY Limits A Corporate Executive's Ability to Use the Advice-Of-Counsel Defense*, KROPE MOSELY SCHMITT (Oct. 13, 2015), https://kmlawfirm.com/2015/10/13/sdny-limits-a-corporate-executives-ability-to-use-the-advice-of-counsel-defense/ [https://perma.cc/6RZZ-4MA2] ("The advice-of-counsel defense is a powerful one. If you did something because your lawyer said it was legal, then you may have a winning defense against many white-collar crimes."). Note, however, that the advice-of-counsel defense comes at a steep cost of waiving attorney-client privilege.

[193] Sec. & Exch. Comm'n v. Panuwat, 702 F. Supp. 3d 883, 898 (N.D. Cal. 2023) (internal citation omitted).

[194] *Id.*

[195] *See* Kathleen Dunphy, *The Role of In-House Legal in Enterprise Risk Management*, DILIGENT (Dec. 10, 2024), https://www.diligent.com/resources/blog/the-role-of-in-house-legal-in-enterprise-risk-management [https://perma.cc/V7VF-M4WF].

how much bandwidth it is willing to allocate to the drafting.

Nevertheless, a policy is a worthwhile investment. To be sure, verifying this is difficult as an empirical matter. It is impossible to know how many instances of insider trading are prevented because of a written policy. However, regulators and companies generally take it as a given that policies work and are the preferred way to handle insider trading risk. In 2022, the SEC amended Regulation S-K to require publicly-traded companies to disclose if they have an insider trading policy, and if so, to file it with the SEC.[196] This indicates that the SEC finds the presence of a policy useful. Moreover, many companies do have extensive insider trading policies, suggesting that after weighing the costs and benefits, the companies decided creation and implementation of insider trading policies was worthwhile.[197]

C. Proposed Policy Components

While the contours of each AI insider trading policy will differ, the following presents proposed components that may be adjusted given the risk-tolerance and needs of different firms. The proposals seek to address the two main risks identified in this Note: misalignment risk and security risk.

Misalignment risk should be addressed with reference to both the underlying technology and its outputs. First, a policy should mandate system audits and outline the frequency with which an IT professional audits the system to ensure it has not developed harmful emergent properties. The audit should also review the outputs to ensure the system is not acting deceptively. The frequency of these audits can be determined on a company-specific level, but given the rapid evolution capacity of AI systems, it should be more frequent than a general IT audit. The policy should also provide guidelines for choosing systems employed within the firm. It may also be worthwhile to require AI systems to "think out loud" using a scratchpad so users can monitor outputs in real time. Second, an effective and efficient risk management policy should mandate human approval of all AI-selected actions. This may take the form of preventing AI systems' access to live trading platforms or communication tools, thereby requiring human input for any external action.

Security risk should address shadow AI use, internal screening mechanisms, and cybersecurity. At the outset, the policy should mandate use of only closed-end systems that do not train on inputs and ensure

[196] 17 C.F.R. § 229.408(b)(1)–(2) (2024).

[197] *Cf.* Maia Gez, Scott Levi, Danielle Herrick, Melinda Anderson, Michelle Rutta, & Guiying Ji, *Insider Trading Policies: A Survey of Recent Findings*, WHITE & CASE (Oct. 13, 2025), https://www.whitecase.com/insight-alert/insider-trading-policies-survey-of-recent-filings [https://perma.cc/Z3JC-EVCC].

confidentiality. A policy should also require an affirmance from users that they will not use any AI systems other than those approved by the firm. Further, to mitigate the risk of breaking through ethical screens, the policy may choose to mandate separate licenses for different teams. This may be cumbersome in practice, so a provision may instead choose to mandate different licenses for different levels of seniority within the firm. Another provision is likely necessary to require the screening authority within the firm (perhaps an ethics attorney) to regularly monitor to ensure the proper screens remain in place in real time. However, none of these provisions would address the risk of accidental leaks in information despite carefully constructed screens. To address this problem, a provision may require regularly auditing the AI system to ensure it respects the boundaries set. Finally, to address the cybersecurity risk, the policy could mandate frequent security tests performed by white-hat hackers to expose weaknesses in the system integrity.

The policy should also consider outlining the use of AI systems by human agents. Here, it can draw on findings from the two laboratory studies discussed. The policy should outline proper prompting etiquette. As discussed *supra* Part III, pressures placed on AI models influence their likelihood to act in misaligned ways. Guidance on how to interact with models to avoid such pressures may be proper in a policy. A policy should also include guidance on what tasks AI systems should and should not be used for. A company may want to assess its practices to determine which ones present the highest risk of insider trading. A policy may choose to either restrict or ban AI use for these activities. Additionally, a policy may choose to mandate specific training on AI system use periodically. These would all serve to minimize harmful human inputs that may ultimately lead to improper AI use.

CONCLUSION

The existing legal obligations and frameworks are not sufficient to address insider trading risk created by AI systems. While the existing frameworks may explicitly or implicitly require firms to engage in risk management which incidentally touches on AI-specific risk, they are insufficient overall. The risks identified are unique to AI use and will continue to diverge from human and traditional IT risks as the AI capabilities evolve. Thus, financial institutions should at least consider implementing AI-specific insider trading policies to mitigate the most salient risks.