Do Investors Care About AI Safety?

Fei Su
City University of Macau
feisu@cityu.edu.mo
Jingshu Wen
Said Business School, University of Oxford
Jingshu.Wen.DPHIL@said.oxford.edu
Tonglin Zhang
The University of Hong Kong
tonglinz@hku.hk

Abstract

We examine whether public equity markets price AI safety risk and whether corporate AI governance moderates that pricing. Using U.S. listed firms from 2015 to 2025, we construct textual measures that distinguish downside AI safety risk from general AI exposure, AI opportunity, and AI governance. AI safety-risk disclosure rises sharply after the generative AI wave, while AI governance disclosure remains rare. We find limited evidence of a stable unconditional AI safety-risk premium. Instead, pricing is salience-dependent and event-driven. Firms with pre-existing, topic-matched AI safety-risk disclosure lose value around regulatory-strengthening events and gain value around regulatory reversals. In contrast, AI capability shocks produce no comparable response. Repricing is strongest when AI harms are likely to become private costs and when PRI signatories or public pension funds are important shareholders. Concrete AI governance actions attenuate these market responses more than boilerplate disclosure. Overall, our results suggest that markets price emerging AI externalities when these risks are financially material or socially salient.

Keywords: AI safety risk; AI governance; Institutional investors; Stock returns; Technological externalities.

JEL Codes: G12, G14, G23, G34, D62, O33.

Acknowledgement: We are grateful for the research assistance from Siyi Long, Xinyi Li, and Qizong Zeng. We declare no financial relationships or other potential conflicts of interest.

1. Introduction

Artificial intelligence (AI) creates large private benefits, but it also creates downside risks that may be borne partly by customers, workers, counterparties, and the public. AI safety risk includes various potential harms from human misuse of AI and from AI system failure, such as deepfakes, biological or security misuse, privacy violations, and misaligned autonomous systems. Recent shareholder proposals and stewardship statements suggest that some investors increasingly view AI safety and AI governance as relevant to investment decision-making (Harvard Law School Forum on Corporate Governance, 2024, 2025; Norges Bank Investment Management, 2023; World Economic Forum, 2024). However, it remains unclear whether emerging investor concerns about AI safety are strong enough to affect market pricing.

This paper asks whether public equity markets price AI safety risk and, if they do, whether AI governance attenuates that pricing. We define AI safety risk as a firm's ex ante exposure to financial losses, operational disruptions, or reputational damage from the development, provision, procurement, or deployment of AI systems. We define AI governance as verifiable controls and organizational capacity to identify, mitigate, monitor, and respond to AI safety risk from the AI systems a firm develops or deploys. Investors may price AI safety risk through two economic channels. Under the financial materiality channel, investors demand compensation for exposure to future AI-related private costs, including regulation, litigation, operational failure, remediation costs, and reputational damage. Under the social preference channel, some investors may dislike holding or financing firms that create or amplify AI-related social harms, even without immediate effects on firms' cash flows. Both channels imply that firms with greater AI safety risk and weaker AI governance may have a higher expected-return premium or sharper repricing when AI risk becomes salient. We investigate financial materiality and investor social preference as two channels through which public equity markets may price firms' ex ante AI safety risk.

To answer this question, we study U.S. listed firms from 2015 to 2025 using SEC Form 10-K filings, CRSP daily and monthly stock returns, Compustat financials, Thomson Reuters 13F institutional ownership data, PRI signatory list, and hand-collected AI regulation,

enforcement, governance, regulatory reversal, discourse salience, and capability placebo events. We construct firm-year level measures of AI safety risk, AI governance, AI opportunity, and general AI exposure from annual report disclosures. We distinguish downside AI safety exposure from value-creating AI use, and we map the textual measures of corporate AI risk and governance to established AI risk frameworks, including the NIST AI Risk Management Framework and EU AI Act categories.

We first document that AI risk disclosure is sparse but rising sharply after the generative AI wave. In our sample using 2018-2025 Item 1A filings, 10.7% of filings contain at least one AI-related risk-factor sentence, while only 1.0% contain AI governance or mitigation language in Item 1A. Item 1/Item 7 AI disclosure and AI opportunity disclosure are more common, appearing in 14.0% and 12.8% of filings, respectively, but specific AI governance actions appear in only 0.5% of filings. By 2025, the share of filings with any AI safety-risk disclosure exceeds 50%, with regulatory exposure, cyber/privacy risk, accuracy/reliability, and bias/ethics rising sharply. Firms rarely use frontier-safety language such as catastrophic misuse, loss of control, or existential safety. Instead, their AI risk disclosures are concentrated in business-related risk channels, including regulation, cyber/privacy, third-party dependence, model reliability, operational risk, and legal or reputational exposure.

We find limited evidence that AI safety risk is priced as a stable, unconditional factor. Portfolio sorts show a value-weighted spread of 0.8% per month, significant at the 10% level, but equal-weighted portfolio returns are not monotonic in AI safety-risk intensity. Fama–MacBeth regressions do not produce statistically significant coefficients on continuous AI safety risk. After the first matched AI regulatory event in July 2023, the binary AI safety risk dummy becomes positive and significant. Thus, the expected-return evidence is concentrated at the post-salience disclosure margin rather than as a robust continuous risk premium.

Our event-study evidence provides stronger support for salience-based repricing. We match each AI safety shock category to the corresponding pre-existing AI safety-risk disclosures. In the main CAR[0,+5] specification, firms with matched AI safety risk experience negative abnormal returns around regulatory-strengthening events, with a matched-risk coefficient of -0.064, and positive abnormal returns around regulatory-reversal events, with a

coefficient of 0.131. Soft-governance events also generate a positive matched-risk coefficient of 0.138, consistent with the interpretation that governance frameworks and voluntary standards may reduce uncertainty or reward firms positioned to adopt credible governance tools. Discourse-signal, enforcement-precedent, and capability-placebo events do not generate significant matched-risk effects in the main CAR[0,+5] window. The absence of comparable effects around capability-placebo events reduces the concern that these results simply reflect generic AI news.

The governance evidence is more nuanced. The matched AI safety risk × matched AI governance interactions generally have the predicted attenuation signs in the main shock categories, but they are not statistically significant. When firms disclose AI safety risk in their own filings, the filing-window market response is smaller if the same filing also discloses a specific AI governance action in Item 1 or Item 7. The interaction between AI safety risk and specific Item 1/Item 7 AI governance action is negative and statistically significant in the full sample and the post-2020 sample. This suggests that investors distinguish concrete governance actions, such as AI policies, AI governance committees, audits, monitoring, or oversight mechanisms, from more generic mitigation language.

For the financial materiality channel, the strongest evidence comes from AI harms that plausibly connect safety concerns to private firm-level costs. We aggregate AI risk disclosures into direct and indirect financial-materiality tiers. Direct financial-materiality exposure, based on regulatory, cyber/privacy, accuracy/reliability, and third-party-dependence risk, is positively priced around regulatory-reversal and soft-governance events but is not significantly penalized around regulatory-strengthening events. Indirect financial-materiality exposure, based on bias/ethics risk, is negatively priced around regulatory-strengthening events, with a coefficient of -0.32, and positively priced around regulatory-reversal events, with a coefficient of 0.437. These results suggest that investors view bias, ethics, and discrimination-related AI harms as financially material when regulation changes the likelihood that those harms become private firm-level costs.

We next test the social preference channel using investor clientele measures. Besides PRI signatory ownership, we use public pension fund ownership to proxy for the social

preference of firms' owners, as public pension funds are norm-constrained and tend to be more ESG-conscious than other types of institutional investors (Bolton et al., 2020). The interaction between AI safety risk and PRI ownership is negative and statistically significant across control sets. Public pension ownership produces a similar pattern. We find that firms with high AI safety risk and public pension ownership have more negative and significant return reactions to AI safety events, ranging from -0.205 to -0.144 across control sets. These pooled-event results indicate that high-AI-safety-risk firms have less favorable event-window responses when socially oriented investors are more important shareholders. These results suggest that socially oriented investor clienteles attenuate favorable repricing, or intensify adverse repricing for firms with high AI safety risk.

Taken together, the findings show that AI safety risk is an emerging technological externality whose public market pricing is salience-dependent. We identify financial materiality and investor social preferences as two complementary channels through which AI safety risk affects asset prices. Capital markets appear able to discipline AI safety externalities, but primarily when harms become salient, financially material, or visible to socially oriented investor clienteles. We also show that investors distinguish generic language from substantive AI governance and respond more to concrete AI governance actions than generic governance disclosures. The paper contributes by showing when public markets can discipline AI safety externalities and by developing scalable firm-level measures of AI safety risk and AI governance.

Our paper contributes to the literature on AI and economics. Existing studies show the effects of AI on innovation, productivity, employment, wages, industrial concentration, inequality, discrimination, and decision quality (Babina et al., 2024; Baily, Brynjolfsson, and Korinek, 2023; Cao et al., 2024; Carvalho, 2025; Fuster et al., 2022; Hui, Reshef, and Zhou, 2024; Jia et al., 2024; Noy and Zhang, 2023). An emerging literature studies AI safety risk and the regulation of transformative AI. Acemoglu (2021) argues that unregulated AI may generate social, economic, and political harms. Jones (2024) shows the tradeoff between AI-driven growth and existential risk, and Jones (2025) studies the scale of socially desirable investment in reducing catastrophic AI risk. Andrews and Farboodi (2025) examine bond-yield responses

to major AI model releases and interpret them through growth, uncertainty, and extreme-outcome channels. Gans (2025a) theorizes that socially minded investors cannot prevent unsafe artificial general intelligence (AGI) development due to the high cost of competing with for-profit firms. We contribute to this literature by providing firm-level evidence on whether public equity investors price AI safety risk and whether market prices create incentives for firms to internalize these risks.

Our study also relates to the pricing and externalities of technological innovation. Innovation, technology shocks, and adoption frictions can affect firm valuation and expected returns (Gârleanu, Panageas, and Yu, 2012; Kogan and Papanikolaou, 2019; Lin, Palazzo, and Yang, 2020). Technological innovation can generate a wedge between private and social payoffs through knowledge spillovers, product-market externalities, and socially costly directions of technical change (Acemoglu, Aghion, Bursztyn, and Hémous, 2012; Bloom, Schankerman, and Van Reenen, 2013; Brown, Martinsson, and Thomann, 2022; Jones and Williams, 1998). AI safety risk is a meaningful setting in which private incentives to develop and deploy frontier technologies may diverge from social welfare, potentially requiring stricter regulation, slower deployment, and mitigation investment (Acemoglu and Lensman, 2024; Gans, 2025b, 2025c; Guerreiro, Rebelo, and Teles, 2023; Jones, 2024, 2025). We examine whether financial markets price these external risks and whether such pricing affects firm-level incentives.

Our study contributes to the literature that uses textual methods to measure risk and various firm characteristics (eg, Bodnaruk, Loughran, and McDonald, 2015; Cong, Liang, Zhang, and Zhu, 2025; Florackis et al., 2023; Hassan et al., 2019; Sautner, van Lent, Vilkov, and Zhang, 2023). Media and disclosure language contain price-relevant information, and financial text requires domain-specific measurement (Loughran and McDonald, 2011; Manela and Moreira, 2017; Tetlock, 2007). Mandatory risk-factor disclosures contain economically meaningful information about firm risk and investors' risk perceptions (Campbell et al., 2014), and firm-identified risk factors can capture exposures not spanned by standard multifactor models (Lopez-Lira, 2023). We extend this literature by developing and validating textual

measures of AI safety risk and AI governance, separating downside safety exposure and governance capacity from general AI exposure and AI opportunity.

The remainder of the paper proceeds as follows. Section 2 presents the background and hypotheses. Section 3 describes the data and empirical strategy. Section 4 reports the main test results and discussions. Section 5 reports the economic channel test results and discussions. Section 6 presents the robustness checks. Section 7 concludes.

2. Background and hypothesis development

The AI safety literature emphasizes that AI-related harms can arise from both human misuse and poorly controlled or misaligned systems (Bucknall and Dori-Hacohen, 2022; Carlsmith, 2022; Hendrycks, Mazeika, and Woodside, 2023; Sandbrink, 2023; Urbina et al., 2022). These risks are not limited to frontier model developers. Downstream firms can also create or amplify AI-related harms through procurement choices, human misuse in high-stakes deployment, poor data governance, weak monitoring, insufficient access controls, and inadequate incident response. We define AI safety risk as a firm's ex ante exposure to financial losses, operational disruptions, or reputational damage from the development, provision, or deployment of AI systems. Given the potential scale and diffusion of these harms, recent work emphasizes the need for greater investment in AI safety and governance (Acemoglu and Lensman, 2024; Bullock et al., 2024; Growiec and Prettner, 2025; Jones, 2025; Future of Life Institute, 2023).

AI safety risk is a technological externality. Firms capture the private benefits of rapid AI development and deployment, while imposing potential harms on customers, workers, third parties, public institutions, or society. Incomplete regulation and weak governance incentives can lead firms to underinvest in safeguards in developing or using AI relative to the socially efficient level. There is especially likely to be free-riding when AI-related harms are diffuse, delayed, and difficult to attribute, such as harms caused by interactions among developers, vendors, deployers, and users. If investors expect external AI harms to become private financial costs, then capital market pricing can incentivize firms to invest in AI governance.

Recent AI-related stewardship activity suggests that some investors increasingly view AI oversight as relevant to governance and investment decision-making. For example, Apple and Netflix shareholder proposals requesting transparency reports on AI use received 37.5% and 43.3% support, respectively, in 2024, measured as support among votes cast (Harvard Law School Forum on Corporate Governance, 2024, 2025). Investor-specific stewardship materials point in the same direction. Norges Bank Investment Management's responsible-AI expectations emphasize board accountability, transparency, and explainability, and robust risk-management processes covering privacy, security, nondiscrimination, and human oversight and control. These developments suggest that AI safety concerns may affect investor demand.

AI safety risk can matter for firm valuation in two ways. First, if AI safety risk is a priced cash-flow or discount-rate risk, investors should require compensation for holding firms exposed to that risk. This prediction is strongest when AI safety risk is systematic, correlated across firms, or likely to materialize in states of the world that investors particularly dislike. We hypothesize that firms with greater ex ante AI safety risk earn higher subsequent expected returns. Second, even if AI safety risk is difficult to price continuously, markets may reprice it sharply when it becomes salient. AI safety salience events include regulatory announcements, enforcement actions, governance frameworks, deregulation, and public discourse that heighten attention to AI safety. As salient states attract more attention (Bordalo, Gennaioli, and Shleifer, 2013), investors may update their pricing when a salient AI safety event reveals that AI harms are more likely to generate private costs or affect social welfare. Previous studies also suggest that markets may begin pricing a new externality more strongly after policy attention, disclosure salience, litigation risk, or public concern increases (Bolton and Kacperczyk, 2021; Garel et al., 2024; Hartzmark and Sussman, 2019; Hsu, Li, and Tsou, 2023).

AI governance can attenuate the pricing of AI safety risk. We define AI governance as verifiable controls and organizational capacity that direct and control how the firm develops, procures, or deploys AI systems. These controls and capacities are designed to reduce the probability or severity of harmful AI outcomes. In practice, AI governance includes risk identification and documentation, technical safeguards such as model evaluation, red teaming, access controls, data governance, and security engineering, organizational accountability such

as compliance processes and board oversight, and lifecycle monitoring and response, including logging, incident reporting, incident response, and decommissioning procedures. A firm that invests in these capabilities may reduce the likelihood of harmful AI outcomes, limit their severity when they occur, improve detection and attribution, and lower expected financial, operational, legal, or reputational costs from AI safety issues. Thus, AI governance can potentially reduce the extent to which investors price AI safety risk.

As a type of social externality, AI safety risk is conceptually related to broader ESG risk. In the framework of Pedersen, Fitzgibbons, and Pomorski (2021), ESG characteristics can affect prices either because they contain information about firm fundamentals or because investors have direct preferences over environmental or social characteristics. Under the financial materiality channel, ESG-related characteristics or exposures affect expected cash flows, risk, or required returns when environmental or social externalities are expected to become private firm-level costs. Previous studies on environmental risks suggest that markets price externalities most strongly when harms are salient and likely to affect firm value (Bolton and Kacperczyk, 2021; Hsu, Li, and Tsou, 2023; Garel et al., 2024). Under the social-preference channel, ESG characteristics affect investor demand because some investors derive utility or disutility from holding firms with particular environmental or social profiles, even absent immediate cash-flow consequences. Consistent with this framework, AI-enabled harms can be both financially material and socially salient.

Under the financial channel, investors demand compensation for exposure to future private costs associated with AI-enabled harms. Regulation and litigation can convert socially diffused AI safety harms into firms' internal costs. For example, AI-practice violations under the EU AI Act may be subject to fines of up to €35 million or 7% of total worldwide annual revenue. Litigation and enforcement can impose fines, settlements, legal expenses, injunctions, and remediation obligations. Besides regulatory and legal channels, firms may internalize AI safety costs via operational and cybersecurity channels, including model failure, data exfiltration, fraud, and system disruption, as well as market-access channels, including reputational damage, customer distrust, and insurance repricing. We therefore expect AI safety pricing to be stronger for firms whose AI risks are more financially material, such as for firms

with greater exposure to AI regulation and litigation. Adverse regulatory or enforcement shocks should penalize high-risk firms more when these internalization mechanisms are strong. Deregulatory or regulatory-relief events should have the opposite sign if they reduce expected compliance, litigation, or remediation costs.

A second channel is investor demand driven by social responsibility. Under the social-preference channel, some investors dislike holding or financing firms that create, enable, or amplify AI-enabled social harms, even when those harms do not immediately affect cash flows. Hartzmark and Sussman (2019) show that shocks to the salience of sustainability information can shift investor demand. In our setting, salient AI safety events may cause socially responsible investors to underweight high-risk AI firms or demand stronger AI governance.

The two channels are not mutually exclusive, as the same AI-enabled harm may be priced partly because investors expect future private costs and partly because some investors have preferences over social harms. Both channels imply that firms with greater AI safety risk may face sharper negative repricing when AI safety risk becomes more salient. They may also exhibit higher subsequent expected returns if AI safety risk is a compensated exposure or if persistent investor underweighting lowers current prices.

AI governance matters under both pricing channels. Effective AI governance reduces the expected probability or severity of AI failures. Under the financial materiality channel, stronger governance can reduce expected financial losses from potential litigation, regulatory enforcement, and reputational damage that are caused by AI safety issues. Under the social responsibility channel, stronger governance may make AI-intensive firms more acceptable to investors who care about safety, accountability, stewardship, and social harm. Thus, investors may require a smaller risk premium from high-risk firms with stronger governance, react less negatively to them around adverse AI safety events, or positively reprice credible governance disclosure when it becomes public.

Based on the analysis above, we develop four hypotheses. H1 tests whether capital markets price AI safety risk, and H2 tests whether AI governance attenuates that pricing. H3 and H4 distinguish the financial and social-preference channels.

H1a (AI safety risk premium). Firms with higher ex ante AI safety risk earn higher subsequent expected stock returns, controlling for standard return predictors and other characteristics.

H1b (AI safety salience-event repricing). Around AI safety regulation tightening, enforcement, or realized-harm events, firms with higher ex ante AI safety risk experience more negative abnormal returns. Around AI safety deregulation events, the predicted sign reverses.

H2 (AI governance attenuation). Conditional on AI safety risk, firms with stronger AI governance have a smaller AI safety risk premium and a smaller reaction to AI safety salience events.

H3 (Financial materiality channel). The AI safety risk premium and AI safety salience-event repricing effects are stronger when AI-enabled harms are more likely to generate material private costs, such as for firms with greater exposure to AI regulation or litigation.

H4 (Social responsibility channel). The AI safety risk premium and salience-event repricing effects are stronger when socially oriented investors are more important marginal shareholders, especially for AI harms with salient social externalities but weaker immediate financial materiality.

Together, these hypotheses examine whether capital markets can discipline an emerging technological externality before public regulation is complete by pricing in AI safety risk.

3. Data and Empirical Strategy

3.1. Data and sample construction

Our sample consists of U.S. listed firms from 2015 to 2025. We combine SEC EDGAR Form 10-K filings with CRSP daily and monthly stock returns, Compustat financials, Thomson Reuters 13F/s34 institutional ownership data, and PRI-matched institutional ownership. We also hand-collect AI safety, regulation, enforcement, regulatory-reversal, soft-governance, discourse-salience, and capability-placebo events. For PRI ownership, we match institutional managers to the PRI signatory directory and measure each firm's pre-event ownership by PRI-matched 13F managers.

The SEC sample contains 78,015 10-K filings from 2015 to 2025. Of these, 58,536 have a parseable Item 1A section, 73,454 have extracted Item 1 text, 76,046 have extracted Item 7 text, and 55,932 are matched to EDGAR filing dates. The matched shock-category event-study sample contains 23 AI-related events organized into six shock categories, yielding 81,561 firm-event observations and 4,087 unique firms.

3.2. Variables

We construct date-stamped firm-level measures of AI safety risk, AI governance, AI opportunity, and AI exposure from firms' Form 10-K disclosures. AI exposure captures the extent to which a firm develops, supplies, adopts, or depends on AI. AI safety risk captures exposure to potential AI-enabled harms. AI governance captures verifiable controls and organizational capacity that may reduce the probability or severity of AI safety issues. AI opportunity captures value-creating AI uses, including AI product development, commercialization, productivity gains, and customer adoption. We use 10-K Item 1A (Risk Factors) text to measure disclosed AI safety risk and AI governance, and we use Item 1 (Business) and Item 7 (MD&A) text to measure AI business opportunities and specific AI governance actions.

Rather than scoring filings directly, we first use a language model to extract short AI-related sentences and classify them into categories, and then compute the final indicators from the classified text. Specifically, we count AI-relevant sentences that also contain risk or governance terms and scale them by section length, following the classification-then-count approach in Kim, Muhn, and Nikolaev (2023) and Hassan et al. (2019). We define continuous text-intensity variables as the share of sentences in the relevant filing item that are classified into the relevant AI category, and we define dummy variables that equal one if at least one sentence in the relevant item is classified into the category and zero otherwise. Our textual measures are comparable across firms with different disclosure lengths. We then map these filing-level variables into the event-study, filing-window, and monthly return panels using only information available before the relevant return window.

3.2.1. AI safety risk

Our baseline text-based measure of AI safety risk is AI risk-factor text. For firm i and filing d, we define

$$AIRiskText_{id} = \frac{\text{Number of Item 1A sentences classified as AI-related risk}}{\text{Number of sentences in Item 1A}}.$$

This measure captures the intensity of AI-related risk disclosure in the firm's risk-factor section. We also define AI risk disclosure as an indicator that equals one if the firm includes at least one AI-related risk sentence in Item 1A, and zero otherwise.

We refer to the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) and the EU AI Act for the AI safety risk taxonomy. The taxonomy covers AI model unreliability, unsafe automation, AI misuse risk, privacy and security failures from AI, discriminatory risk from AI, dependence on third-party AI models or infrastructure, weak monitoring, inadequate control or incident response, catastrophic misuse, loss of control, and broader systemic or existential safety concerns. We classify Item 1A AI risk-factor language into the following sub-categories: regulatory exposure, cyber/privacy exposure, bias/ethics, accuracy/reliability, competition, IP/legal risk, third-party dependence, operational risk, catastrophic misuse, loss-of-control, and existential safety.

We also match each category of AI safety risk to each event shock category. For each shock category c, we retain only the Item 1A AI risk sub-categories that are conceptually linked to the economic content of that category. Formally, we define:

$$MatchedAI_SafetyRisk_{idc} = \frac{\text{Number of Item 1A AI-risk sentences in categories matched to } c}{\text{Number of sentences in Item 1A}}.$$

3.2.2. AI governance

Empirically, we define credible AI governance as disclosure of concrete, verifiable controls and organizational capacities that address AI safety risk, including model evaluation, red teaming, access controls, data governance, privacy and security engineering, monitoring, incident response, accountability, compliance, vendor oversight, and board or executive oversight.

We measure AI governance disclosure in two ways. First, in Item 1A, we define AI governance/mitigation as an indicator that is equal to one if the firm's risk-factor section discusses an AI governance, oversight, or mitigation action, such as board oversight, a dedicated risk committee, internal AI policies, monitoring, controls, or other risk-mitigation procedures. This variable captures whether the firm pairs AI risk disclosure with disclosed governance or mitigation capacity. In the matched event-study design, Matched AI governance restricts this governance or mitigation language to the categories linked to the relevant shock family. Second, in Item 1 and Item 7, we define Specific Item 1/7 AI governance action as an indicator equal to one if the firm describes a concrete AI governance action in its business or MD&A disclosure, such as an AI committee, policy, audit process, monitoring system, or formal governance procedure.

3.2.3. AI exposure and AI opportunity

To separate AI risk from general AI exposure or AI-related profit potential, we construct an AI exposure and AI opportunities measure. We extract text from Item 1 and Item 7 and classify AI-related passages that discuss AI product development, customer adoption of AI products or services, revenue growth from AI, productivity gains from AI, and AI commercialization benefits.

We define AI upside/opportunity as an indicator equal to one if Item 1 or Item 7 describes AI as a business opportunity, growth driver, product capability, productivity tool, commercialization strategy, or core technological asset. This variable separates value-creating AI uses from downside AI safety risk. For AI exposure, we define AI share of Business/MD&A as the share of Item 1 and Item 7 sentences that discuss AI, which captures the overall intensity of AI-related business disclosure outside the risk-factor section. In addition, we define High AI exposure as an indicator equal to one if the firm is in the top quartile of combined AI risk and AI upside text intensity.

The detailed definitions of all variables are reported in Appendix Table A1.

3.2.4. Control Variables

We control for systematic risk exposures and firm characteristics commonly used in empirical asset pricing, including market beta, size measured as log market capitalization, book-to-market, momentum, operating profitability, investment, and liquidity. The firm-level characteristics are associated with the Fama–French five-factor framework and Carhart momentum (Fama and French, 1992, 1996, 2004, 2015; Carhart, 1997; Fama, 2014). We also control for log price, idiosyncratic volatility, log turnover, and log Amihud illiquidity.

In our economic channel tests, we additionally include relevant controls such as total 13F institutional ownership and AI exposure where appropriate.

3.3. Empirical tests

This section tests whether capital markets price AI safety risk and whether AI governance attenuates that pricing. H1 has two implications. First, if AI safety risk is a compensated exposure, firms with greater ex ante AI safety risk should earn higher subsequent expected returns. Second, if AI safety risk is priced when it becomes salient, firms with greater ex ante AI safety risk should experience more negative abnormal returns around adverse AI safety, regulatory, enforcement, or realized-harm shocks, and more positive abnormal returns around deregulatory or regulatory-relief shocks. H2 predicts that stronger AI governance weakens both effects, so that high-risk firms with stronger governance should have a smaller AI safety-risk premium and a smaller event-window repricing response.

We measure AI safety risk and AI governance before the return, event, or filing window outcome. We use $S_{i,t}$ to denote firm i's ex ante AI safety-risk exposure, measured primarily from Item 1A AI risk-factor text. We use $G_{i,t}$ to denote AI governance and $A_{i,t}$ to denote non-safety AI controls, including general AI exposure, AI opportunity, and generic AI-related business risk when they are included in a specification. We use $X_{i,t}$ to denote standard return and firm controls.

3.3.1 AI risk premium

We first test H1a using monthly portfolio sorts and Fama–MacBeth regressions. Because AI risk-factor disclosure is sparse and many firms have no AI risk text, the portfolio tests separate firms with zero AI risk disclosure from firms with positive AI risk disclosure. In each month, we form a zero-disclosure portfolio and quartile portfolios among positive AI risk disclosers. We report equal-weighted mean monthly returns for each portfolio and the value-weighted return spread between the highest positive-disclosure quartile and the zero-disclosure portfolio.

We then estimate monthly cross-sectional return regressions:

$$R_{i,m+1} = \alpha_m + \mu_{s(i),m} + \beta_{1,m}S_{i,m} + \beta_{2,m}G_{i,m} + \beta_{3,m}(S_{i,m} \times G_{i,m}) + \Lambda'_m A_{i,m} + \Gamma'_m X_{i,m} + \varepsilon_{i,m+1} \quad (1)$$

In Equation (1), $R_{i,m+1}$ is firm i's one-month-ahead monthly stock return. $\mu_{s(i),m}$ denotes sector controls within month m. We estimate Equation (1) using monthly Fama–MacBeth regressions and report the time-series average of the monthly slope coefficients, with Newey–West standard errors using six lags.

H1a predicts $\bar{\beta}_1 > 0$, or that firms with greater ex ante AI safety risk earn higher subsequent returns if AI safety risk is a compensated exposure. H2 predicts $\bar{\beta}_3 < 0$, or that stronger AI governance conditional on AI safety risk should reduce the AI safety-risk premium. The marginal expected-return premium associated with AI safety risk is $\frac{\partial R^e_{i,m+1}}{\partial S_{i,m}} = \beta_{1,m} + \beta_{3,m}G_{i,m}$.

3.3.2 AI safety salience event repricing

We next test H1b by examining whether AI safety risk is repriced when AI safety, regulation, enforcement, governance, regulatory reversal, or public discourse becomes salient. We classify the 23 events into six AI safety shock categories: regulatory strengthening, regulatory reversal, soft governance, discourse signals, enforcement precedent, and capability placebo. Regulatory-strengthening events, such as the EU AI Act milestones, Colorado AI Act, CA SB1047 passage, and Executive Order 14110, are adverse shocks because they increase expected compliance, litigation, remediation, reputational, or operational costs for firms

exposed to the matched AI harm. Regulatory-reversal events are relief shocks because they reduce expected regulatory or compliance burdens. Enforcement-precedent events are adverse when they reveal litigation or enforcement costs for specific AI harms. Soft-governance events provide governance tools or standards without imposing direct penalties. Discourse-signal events raise attention to AI safety without an actionable compliance obligation. Capability-placebo events capture broad AI capability or commercialization news rather than safety or regulatory cost shocks.

For example, Executive Order 14110 is a regulatory-strengthening shock. It was signed on October 30, 2023, and later rescinded on January 20, 2025, which we classify as a regulatory-reversal event. The EU AI Act enters the event set through the December 8, 2023 political agreement and later staged implementation milestones, including prohibited practices from February 2, 2025, the EU GPAI Code on July 10, 2025, and GPAI obligations on August 4, 2025. Another common shock is the first global AI Safety Summit in November 2023. The UK hosted the first global AI Safety Summit at Bletchley Park in November 2023, which resulted in the Bletchley Declaration, an agreement among 28 countries to collaborate on AI safety. The Bletchley Park AI Safety Summit is a discourse-signal event because it increased public attention to AI safety without directly imposing a binding compliance cost.

Table 2 summarizes the shock-category taxonomy that underlies the matched event-study design. We expect regulatory-strengthening events to generate negative matched-risk coefficients because they increase the likelihood that AI-related harms become legally or financially internalized. Regulatory-reversal events are predicted to have the opposite sign because they reduce expected compliance or regulatory burdens for firms with high levels of relevant AI safety risk. We expect soft governance events to have weakly nonnegative effects because they provide governance tools or standards that firms can adopt, rather than imposing direct penalties. Discourse signals have no clear directional prediction because they increase attention to AI safety without creating an actionable compliance obligation. We expect enforcement precedent events to have negative effects when they reveal litigation or enforcement costs for specific AI harms, while we use the ChatGPT release as an AI capability placebo because it is not a safety or regulatory shock.

[Insert Table 2 here]

Appendix Table A2 reports the full event list, including event dates, shock-category assignments, predicted signs, firm counts, matched-risk treatment rates, and matched-governance rates. The table shows that matched AI safety-risk disclosure rises sharply over time.

For each event e, we compute firm-level cumulative abnormal returns over short trading-day windows, including $[-1,+5]$, $[0,+5]$, and $[0,+10]$. The baseline abnormal return is market-adjusted. We match each event family to the AI safety-risk categories most directly implicated by the event.

We estimate the following event-stacked cross-sectional regression separately by shock category:

$$CAR_{i,e}^{[a,b]} = \alpha_e + \mu_{s(i)} + \pi_1 S_{i,e}^{M(e)} + \pi_2 G_{i,e}^{M(e)} + \pi_3 \left(S_{i,e}^{M(e)} \times G_{i,e}^{M(e)}\right) + \Lambda' A_{i,e} + \Gamma' X_{i,e} + \varepsilon_{i,e} \quad (2)$$

In Equation (2), $CAR_{i,e}^{[a,b]}$ is firm i's cumulative market-adjusted abnormal return around event e over event window [a, b]. $S_{i,e}^{M(e)}$ denotes firm i's AI safety risk in the categories matched to event e, measured before the event. $G_{i,e}^{M(e)}$ denotes the corresponding matched AI governance. Event fixed effects α_e absorb market-wide event effects, and $\mu_{s(i)}$ denotes GICS sector fixed effects, which absorb persistent industry differences. The vector $X_{i,e}$ contains pre-event market controls, including size, price, momentum, beta, idiosyncratic volatility, turnover, and Amihud illiquidity. We cluster standard errors by firm.

For adverse events, especially regulatory strengthening and enforcement precedent, H1b predicts $\pi_1 < 0$. For regulatory-reversal events, H1b predicts $\pi_1 > 0$ because firms with greater matched AI safety risk should benefit when expected regulatory or compliance burdens decline. H2 predicts that governance attenuates these effects,so π_3 should have the opposite sign from π_1 within a given shock category.

3.3.3 AI Governance Mitigation

To test H2 using event studies, we include matched AI governance and the interaction between matched AI safety risk and matched AI governance in the shock-category regressions. Matched AI governance captures governance or mitigation language that addresses the same AI harm categories implicated by the event. If governance reduces the expected probability or severity of AI safety failures, the interaction should attenuate the matched-risk level effect. For regulatory-strengthening and enforcement-precedent events, this implies a positive interaction coefficient. For regulatory-reversal events, this implies a negative interaction coefficient, because better-governed firms should benefit less from relief if governance has already reduced their effective risk exposure.

We also examine whether investors react when firms disclose AI safety risk and AI governance in their own 10-K filings. These tests provide a disclosure-timing complement to H2. If governance credibly reduces AI safety risk, then the market response to AI risk disclosure should be smaller when the same filing also contains credible governance disclosure or specific governance actions.

For each 10-K filing date f, we compute the firm's filing-window abnormal return $CAR_{i,f}^{[0,+5]}$ and estimate:

$$\begin{aligned} CAR_{i,f}^{[0,+5]} &= \alpha_{t(f)} + \mu_{s(i)} + \rho_1 S_{i,f} + \rho_2 G_{i,f}^{1A} + \rho_3 \left(S_{i,f} \times G_{i,f}^{1A}\right) \\ &+ \rho_4 G_{i,f}^{Act} + \rho_5 \left(S_{i,f} \times G_{i,f}^{Act}\right) + \Lambda' A_{i,f} + \Gamma' X_{i,f} + \varepsilon_{i,f} \end{aligned} \quad (3)$$

In Equation (3), $CAR_{i,f}^{[0,+5]}$ is firm i's cumulative market-adjusted abnormal return over the five trading days after filing f. $S_{i,f}$ is the AI safety-risk text disclosed in the filing, $G_{i,f}^{1A}$ is Item 1A governance or mitigation disclosure, and $G_{i,f}^{Act}$ is a specific AI governance action disclosed in Item 1 or Item 7, such as an AI policy, committee, audit, monitoring process, or oversight mechanism. The fixed effect $\alpha_{t(f)}$ denotes filing-month fixed effects, and $\mu_{s(i)}$ denotes GICS sector fixed effects. The control vector $A_{i,f}$ contains non-safety AI controls, and $X_{i,f}$ uses monthly pre-filing analogues of the market controls in our previous tests. We double cluster standard errors by firm and event month. The key coefficients

are those of the interaction terms, or ρ_3 and ρ_5. If AI safety-risk disclosure is priced as a positive risk premium in filing-window returns, H2 predicts negative interaction term coefficients. If AI safety-risk disclosure is priced as bad news in a given specification, H2 predicts positive interaction term coefficients. In either case, the governance interaction should reduce the magnitude of the market response to AI safety risk.

These tests map directly to H1 and H2. The portfolio and Fama–MacBeth tests examine whether AI safety risk earns a higher expected return, as predicted by H1a. The matched event-study tests examine whether AI safety risk is repriced when regulation, enforcement, governance, or public discourse makes AI harms salient, as predicted by H1b. The event-window and filing-window governance interactions examine whether governance attenuates the pricing of AI safety risk, as predicted by H2.

4. Main Test Results and Discussions

4.1 Descriptive statistics

Table 1 reports the sample composition and distribution of the AI text measures. The filing sample contains 78,015 10-K filings from 2015 to 2025, of which 58,536 have a valid Item 1A risk-factor section. The Item 1A sample covers 10,921 unique firms. The matched event-study sample contains 23 AI-related events, 81,561 firm-event observations, and 4,087 unique firms. The CAR[0,+5] event-study sample contains 81,495 firm-event observations.

[Insert Table 1 here]

Appendix Table A4 reports text extraction and SEC filing date match quality statistics. The overall Item 1A-to-EDGAR match rate is 99.7%, supporting the validity of both the filing-window and event-window designs and reducing the concern that the results are driven by uneven text extraction or filing-date availability across years.

AI risk-factor disclosure is economically meaningful but still sparse. In the 2018–2025 filing-year sample, the mean share of Item 1A text devoted to AI safety risk is 0.004, while the 25th percentile, median, and 75th percentile are all zero. The AI safety-risk disclosure dummy has a mean of 0.107, indicating that 10.7% of filings contain at least one AI-related risk-factor sentence. The more common sub-categories include operational risk, regulatory exposure,

cyber/privacy risk, third-party dependence, competition, IP/legal risk, accuracy/reliability, and bias/ethics. Catastrophic misuse, loss-of-control, and existential-safety language are almost absent.

This distribution shows that public firms do not generally describe AI safety risk in the frontier-safety language emphasized in the AI safety literature. Instead, they disclose AI risk through channels that are closer to financial materiality, such as regulation, privacy, cybersecurity, model reliability, third-party dependence, operational disruption, and legal or reputational exposure.

AI governance disclosure is rarer than AI risk disclosure. Item 1A AI governance or mitigation language appears in only 1.0% of filings, and specific AI governance actions disclosed in Item 1 or Item 7 appear in only 0.5% of filings. By comparison, general AI disclosure and AI opportunity disclosure appear in 14% and 12.8% of filings, respectively. This gap suggests that firms are increasingly willing to discuss AI as a business opportunity or risk exposure, but only a small subset disclose concrete AI oversight structures, policies, audits, committees, or monitoring processes.

Figures 1 and 2 show that AI risk disclosure rises sharply over time and becomes more intensive after the generative-AI wave. Figure 1 documents the annual increase in the share of filings with any AI safety risk disclosure and decomposes the increase into major risk categories. By 2025, AI safety risk disclosure exceeds 50% of filings, with regulatory exposure and cyber/privacy among the largest categories and accuracy/reliability and bias/ethics also rising.

[Insert Figure 1 here]

Figure 2 shows the corresponding cross-sectional distribution from 2018 through 2025. The mass at zero falls over time, and a right tail of more intensive AI disclosure emerges after the generative-AI wave.

[Insert Figure 2 here]

Figure 3 extends the disclosure-trend evidence to the matched event-study sample. Each point reports the matched-risk treatment rate for one of the 23 AI-related events, defined as the share of firms in the CAR[0,+5] event-window panel whose pre-event Item 1A risk factors

discuss the AI harm category matched to that event's shock category. The figure shows that matched treatment is sparse in the earliest events but rises sharply through 2024 and 2025 as firms increasingly add AI risk-factor language to their 10-K disclosures. For example, the treatment rate rises from low single digits for early 2023 events to more than 30% for several 2025 events, and the later EU GPAI events reach 46.4%. These trends support the empirical design of separating zero-disclosure firms from positive disclosers and using event fixed effects and matched risk categories to avoid conflating AI risk repricing with the mechanical growth of AI disclosure over time.

[Insert Figure 3 here]

4.2. AI risk premium

Table 3 reports the monthly portfolio-sort results. The equal-weighted zero-disclosure portfolio earns an average monthly return of 0.764%. Among positive AI safety risk disclosers, the four equal-weighted quartile portfolios earn 0.57%, 1.412%, 1.23%, and 0.548%, respectively. The equal-weighted returns therefore do not show a monotonic increase from low to high AI safety-risk intensity. The value-weighted spread between the highest positive-disclosure quartile and the zero-disclosure portfolio is 0.799% per month and is significant at the 10% level. This provides some evidence of compensation among larger high-AI-risk disclosers, but the non-monotonic equal-weighted pattern suggests that AI safety risk is not yet a stable unconditional return factor.

[Insert Table 3 here]

We report Fama–MacBeth regressions using one-month-ahead monthly stock returns in Table 4. We also split the expected-return tests before and after July 2023, the month of the first matched AI regulatory event. The coefficient on continuous AI safety risk is -0.021 with a standard error of 0.09 in the full sample, -0.005 with a standard error of 0.136 in the pre-salience sample, and -0.052 with a standard error of 0.043 in the post-salience sample. None of these coefficients is statistically significant. The AI safety-risk disclosure dummy is also insignificant in the full and pre-salience samples, but it becomes positive and significant after

July 2023, with a post-salience coefficient of 0.003. The AI governance dummy is insignificant, and the AI safety risk × AI governance interaction is negative but statistically insignificant.

[Insert Table 4 here]

Overall, the expected-return evidence provides partial support for H1a. AI safety risk is not priced as a stable unconditional cross-sectional return factor. The evidence is stronger after regulatory salience increases, but it is concentrated in the binary disclosure margin rather than in the continuous AI safety-risk intensity measure.

4.3. AI safety salience event repricing

We next examine whether firms with matched ex ante AI risk disclosure are repriced when specific AI safety and regulatory shocks occur. Table 5 reports the main matched shock-category event-window regressions using CAR[0,+5]. For regulatory-strengthening events, the coefficient on matched AI safety risk is -0.064 and statistically significant at the 1% level. When regulation increases the expected cost of AI-related harms, firms whose prior disclosures identify matched AI risk categories experience more negative abnormal returns.

[Insert Table 5 here]

We find the opposite pattern for regulatory-reversal events. For regulatory-reversal events, matched AI safety risk is positive and highly significant, with a coefficient of 0.131. High-matched-risk firms therefore gain value when expected regulatory or compliance burdens are reduced. This sign reversal suggests that high-risk firms lose value when AI safety costs are more likely to be internalized, but gain value when those expected costs are relaxed. The symmetry between regulatory strengthening and regulatory reversal is consistent with changes in expected compliance, litigation, remediation, operational, and reputational costs.

We also find that matched AI safety risk is positive and significant around soft-governance events, with a coefficient of 0.138. Matched AI governance is also positive and significant in the soft-governance specification. Soft-governance events, such as voluntary commitments, governance profiles, and AI safety-institution coordination mechanisms, do not necessarily impose binding compliance costs, but they may reduce uncertainty, create implementation tools, or reward firms positioned to adopt credible governance instruments.

The remaining categories provide weaker evidence in the main CAR[0,+5] window. The matched AI safety-risk coefficient is positive but statistically insignificant for discourse-signal events and enforcement-precedent events. The ChatGPT capability placebo also does not generate a statistically significant matched-risk effect in the main CAR[0,+5] window. Rather than reflecting a generic reaction to AI news, the main results therefore support salience-based internalization: firms with matched AI safety risk are repriced when the event changes the expected private cost or governance environment of the relevant AI harm.

Appendix Table A6 shows that the regulatory-strengthening result is negative across alternative CAR windows, while regulatory-reversal and soft-governance effects are positive in the main windows. Appendix Table A7 reports event-level two-step and randomization-inference diagnostics. Appendix Table A8 shows that the regulatory-strengthening result does not depend on the re-dated EU AI Act event. Appendix Table A9 reports placebo tests, including the ChatGPT capability placebo and random-match placebo exercises. These tests support the broad sign pattern of salience-based repricing, but the random-match placebo indicates that the evidence is weaker for the precise category-to-subcategory matching. We therefore interpret the event results primarily as evidence that AI safety risk is priced when salient events change expected costs or attention, rather than as definitive validation of individual risk-category mapping.

Taken together, the event-study evidence supports H1b more strongly than the expected-return evidence supports H1a. Markets do not appear to price AI safety risk continuously as a stable expected-return factor. They do, however, reprice AI safety risk when regulation, regulatory relief, or governance events change the expected private cost or salience of AI harms. The results suggest that capital markets may discipline AI safety externalities, but primarily when those externalities become financially material through regulation, compliance obligations, litigation exposure, or reputational consequences.

4.4. AI Governance Mitigation

In Table 5, we also include matched AI governance and the matched AI safety risk × matched AI governance interaction in each shock-category regression. Consistent with

governance attenuation, the interaction is positive for regulatory-strengthening events, where governance should reduce negative repricing, and negative for regulatory-reversal and soft-governance events, where governance should reduce the upside from relief or governance reassurance. However, these event-window interaction coefficients are not statistically significant. Thus, the event-window evidence is directionally consistent with H2 but does not establish statistically strong governance attenuation.

Appendix Table A11 provides additional event-window governance tests using alternative governance indicators. Panel A uses specific AI governance disclosed in 10-K Items 1 and 7, while Panel B uses the Item 1A AI governance/mitigation indicator. The regulatory-strengthening interaction is positive and marginally significant when governance is measured using the Item 1A AI governance/mitigation indicator, but the other interaction estimates are generally imprecise. Generic Item 1/Item 7 AI-governance rhetoric is too sparse to identify a separate interaction.

The filing-window regressions in Appendix Table A10 provide clearer support for the governance channel. Around firms' own 10-K filing dates, the interaction between AI safety risk and specific Item 1/Item 7 governance action is negative and statistically significant in the full sample and post-2020 sample. The coefficient is -0.206 in the full sample and -0.22 in the post-2020 sample. The same interaction is negative but statistically insignificant in the shorter GenAI-era subsample. Interactions involving broader or more generic governance measures are weaker.

Overall, the H2 evidence supports a narrower governance interpretation. The event-window interaction tests are directionally consistent with governance attenuation, but they are not statistically strong enough to establish broad event-window mitigation. The filing-window evidence shows that investors respond less strongly to AI safety-risk disclosure when the same filing contains a concrete Item 1/Item 7 AI governance action. Thus, the evidence suggests that investors distinguish concrete governance actions from generic mitigation language, and that substantive AI governance can attenuate the pricing of AI safety risk.

5. Economic channel tests and results

5.1. Financial Materiality Channel: When Do AI Harms Become Private Costs?

The financial materiality channel predicts that investors price AI safety risk more strongly when AI-enabled harms are more likely to be internalized as private firm-level costs. These costs may arise through regulation, enforcement, litigation, remediation, cybersecurity loss, operational disruption, reputational damage, procurement exclusion, counterparty restrictions, or higher compliance costs. We therefore test whether event-window repricing of AI safety risk is stronger for AI risk categories that plausibly connect AI safety concerns to private financial costs.

We operationalize financial materiality by aggregating AI safety-risk disclosures into two tiers. The direct financial materiality tier, Q_DFM, sums Item 1A AI risk text related to regulatory exposure, cyber/privacy, accuracy/reliability, and third-party dependence. These categories have relatively direct cash-cost channels, such as fines, breach-related losses, model failure costs, contractual exposure, and vendor or infrastructure dependence. The indirect financial materiality tier, Q_IFM, captures bias and ethics risk text. This category may initially operate through reputational or social harm channels, but it can become financially material through discrimination claims, civil-rights enforcement, litigation, remediation obligations, and customer or worker backlash.

We estimate the following event-stacked cross-sectional regression:

$$\mathrm{CAR}_{i,e}^{[0,+5]} = \alpha_e + \mu_{s(i)} + \beta_D Q_{i,e}^{\mathrm{DFM}} + \beta_I Q_{i,e}^{\mathrm{IFM}} + \Gamma' X_{i,e} + \varepsilon_{i,e} \qquad (4)$$

In Equation (4), Q_DFM and Q_IFM are entered jointly, α_e denotes event fixed effects, $\mu_{s(i)}$ denotes GICS sector fixed effects, and $X_{i,e}$ denotes pre-event market controls. We cluster standard errors by firm.

Table 6 reports the results of our financial materiality channel tests using CAR[0,+5] around AI safety salience events. The strongest pattern appears for Q_IFM. Q_IFM is negative and significant around regulatory-strengthening events, with a coefficient of -0.32, and positive and significant around regulatory-reversal events, with a coefficient of 0.437. This sign reversal is consistent with the financial internalization interpretation that bias, ethics, and

discrimination-related AI safety risks are penalized when regulation tightens and rewarded when regulatory burden is relieved.

[Insert Table 6 here]

By contrast, Q_DFM is not significantly penalized around regulatory-strengthening events. However, Q_DFM is significantly positive around regulatory-reversal events, with a coefficient of 0.032, and significantly positive around soft-governance events, also with a coefficient of 0.032. This suggests that direct cost-channel disclosures matter, but the clearest adverse-versus-relief symmetry appears in the indirect financial-materiality tier. Investors, therefore, appear to treat bias, ethics, and discrimination-related AI harms as financially material when regulation changes the likelihood that those harms become private firm-level costs.

Overall, the results suggest that markets price AI safety risk most clearly when a salient event makes social or reputational harms more likely to be converted into legal, regulatory, or operational costs. Direct financial materiality disclosures are priced around relief and soft-governance events, but they are not significantly penalized around regulatory-strengthening events in the joint specification.

5.2. Social Responsibility Channel: Investor Clientele and AI Safety Repricing

The social-responsibility channel predicts that the pricing of AI safety risk should depend on the importance of investors who are especially sensitive to social externalities. We use two investor-clientele proxies. The first is PRI-matched ownership, defined as the percentage of a firm's shares outstanding held by 13F managers matched to the UN PRI signatory directory, measured at the latest reportable quarter before the event. PRI-matched investors are identified through normalized-name fuzzy matching, curated overrides, and manual checks of high-impact managers. Higher values indicate greater PRI-linked investor ownership and proxy for socially responsible investor exposure.

The second measure is public pension ownership. Public pension funds are likely to be more socially constrained because they face public mandates, beneficiary scrutiny, and political accountability, and Bolton et al. (2020) show that public pension funds are more likely to

support social and environmental proposals than other institutional investors. We use Thomson Reuters 13F public pension manager set (s34) and measure public pension ownership at the latest reportable quarter before the event.

We conduct the following event study for the social responsibility channel:

$$\mathrm{CAR}_{i,e}^{[0,+5]} = \alpha_e + \mu_{s(i)} + \beta S_{i,e}^{M} + \delta\,\mathrm{SocialPref}_i + \gamma\left(S_{i,e}^{M} \times \mathrm{SocialPref}_i\right) + \Gamma' X_{i,e} + \varepsilon_{i,e} \quad (5)$$

In Equation (5), $S_{i,e}^{M}$ is matched AI safety risk, $\mathrm{SocialPref}_i$ is either PRI-matched ownership or public-pension ownership, α_e denotes event fixed effects, $\mu_{s(i)}$ denotes GICS sector fixed effects, and $X_{i,e}$ contains the standard pre-event market controls. We cluster standard errors by firm. For both ownership variables, the event regressions pool non-placebo AI safety salience events, with standard errors clustered by firm. The coefficient of interest is γ, which captures whether socially oriented investor ownership heightens the event-window pricing of AI safety risk.

Table 7 reports the PRI ownership tests. The sample pools the non-placebo AI safety salience events and includes GICS sector and event fixed effects. The coefficient on AI safety risk is positive and statistically significant across specifications, while the interaction between AI safety risk and PRI ownership is negative and statistically significant across all control sets. The interaction ranges from -0.249 in the baseline specification to -0.204 in the full-control specification. Thus, firms with higher AI safety risk experience less favorable event-window abnormal returns when PRI-matched investors are more important shareholders.

[Insert Table 7 here]

Table 8 reports the corresponding public pension ownership tests. The interaction between AI safety risk and public pension ownership is negative and statistically significant across specifications. The coefficient is -0.173 in the baseline specification, -0.205 with baseline market controls, -0.199 after adding Fama–French and momentum controls, and -0.144 in the full-control specification. These estimates imply that high-AI-safety-risk firms have less favorable event-window stock-price responses when public pensions are more important shareholders.

[Insert Table 8 here]

We also examine whether the pricing results are accompanied by observable ownership reallocation, and report the results in Appendix Table A12. The evidence does not show significant post-event changes in pension, non-pension, or total 13F ownership of high-AI-risk firms across shock categories. Because 13F ownership is quarterly and adjusts slowly, these tests are best interpreted as supporting evidence. The lack of detectable ownership reallocation suggests that the social responsibility channel operates through event-window pricing, marginal demand, engagement expectations, or shorter-horizon trading not captured by quarterly 13F changes, rather than through large observable post-event portfolio shifts.

Overall, the evidence supports the social-responsibility channel. PRI-matched ownership and public-pension ownership both significantly contribute to AI safety-risk repricing in the pooled non-placebo event sample. This pattern is consistent with socially oriented investor clienteles attenuating favorable repricing, or intensifying adverse repricing, for firms whose disclosures identify AI safety risk. The results suggest that AI safety risk is priced not only when harms are likely to become private costs through regulation or litigation, but also when AI safety becomes salient to investors with stronger social responsibility concerns.

6. Measurement Validation and Robustness checks

6.1. Placebo and robustness tests

Appendix Table A5 reports robustness to alternative control sets in the matched AI safety-risk event study. Across specifications, matched AI safety risk is negative around regulatory-strengthening events and positive around regulatory-reversal and soft-governance events. The regulatory-strengthening estimate becomes less precise when the overall AI safety-risk share is added, suggesting that regulatory events may activate broader AI-risk concerns as well as the matched harm categories. Regulation reversal and soft governance remain the most stable positive event categories.

In Appendix Table A6, we report the results using various CAR windows. Regulatory strengthening remains negative across all five windows, including CAR[-1,+3], CAR[-1,+5], CAR[0,+3], CAR[0,+5], and CAR[0,+10]. Regulation reversal is positive in the main windows, and soft governance remains positive across all windows. Discourse signals, enforcement

precedent, and the capability placebo are less stable, consistent with the weaker directional predictions for these categories.

In Appendix Table A7, we address the small number of events in each category using event-level two-step estimates, sign tests, inverse-variance-weighted means, and randomization-inference tests. The pooled firm-event estimates are strongest for regulatory strengthening, regulatory reversal, and soft governance. Event-level summaries are more mixed, especially for regulatory strengthening, where the equal-weighted event-level mean is not significant, even though the inverse-variance-weighted mean is negative and significant. Randomization inference supports the pooled estimates for regulatory strengthening, regulatory reversal, and soft governance, while enforcement precedent is not supported by randomization inference. These diagnostics support the main salience-pricing interpretation.

We remove the EU AI Act event and report the test results in Appendix Table A8. The EU AI Act enters the matched event study at the December 8, 2023 political agreement rather than the later March 2024 procedural vote. Because the single-event estimate is entangled with nearby macro news, Appendix Table A8 re-estimates the regulatory-strengthening family excluding the EU AI Act event. The matched AI safety-risk coefficient remains negative and statistically significant across the audited windows, so the regulatory-strengthening result does not rest on this one event.

Appendix Table A9 reports placebo tests. The ChatGPT release placebo does not generate a statistically significant matched-risk effect in the main CAR[0,+5] window. The random-match placebo tests show that the true matched coefficients are not always statistically distinguishable from random category matches. These results suggest that the strongest inference is about the sign and salience of AI safety repricing.

Appendix Table A13 reports stacked difference-in-differences tests as an alternative dynamic design. These tests are broadly consistent with salience-based repricing for regulatory-strengthening and regulatory-reversal events, while also showing that dynamic responses vary across later event months and shock categories.

6.2. Event-set audit and textual measure validation

In Appendix Table A3, we audit the event set to strengthen the interpretation of the financial materiality and social responsibility channels. The audit classifies the 23 events by event scope, social externality salience, and degree of surprise. Events are coded as socially salient when they make salient harms such as discrimination, civil rights, automated hiring, biometric surveillance, privacy, consumer deception, or worker and public-safety harms. Targeted and socially salient events are more directly tied to specific harm-internalization channels, while broad or anticipated events may generate weaker price reactions because some information may be incorporated before the event date.

We conduct several textual measure validation exercises. First, two research assistants manually and independently code a stratified random sample of 1,500 filing sections using the same 12-category taxonomy as the LLM, providing extraction precision, recall, and classification agreement metrics. Second, we compare keyword-based and LLM-based approaches for the text measures, testing whether LLM measures subsume keyword measures in return regressions, following Siano (2025). These procedures support the validity of our textual measures. For the sake of brevity, we omit the details here.

7. Conclusions

This paper examines whether public equity markets price AI safety risk and whether disclosed AI governance attenuates firms' exposure to that risk. Using U.S. listed firms from 2015 to 2025, we construct measures of AI safety risk, AI governance, AI opportunity, and general AI exposure from 10-K disclosures. We document a sharp increase in AI safety-risk disclosure following the rise of generative AI. These disclosures, however, are concentrated in financially material business channels, including regulation, cyber and privacy risk, operational disruption, legal exposure, and discrimination-related harms. By contrast, disclosure of concrete AI governance and mitigation capacity remains limited, indicating a gap between firms' recognition of AI-related downside exposure and their disclosed ability to manage it.

Our evidence suggests that AI safety risk is not yet priced as a stable, unconditional risk factor across public equities. Instead, market pricing is event-driven and depends on the

salience of expected private costs. Firms with pre-existing AI safety-risk disclosures matched to the event topic experience negative abnormal returns around regulatory-strengthening events and positive abnormal returns around regulatory-reversal events. Soft-governance events also generate positive repricing for these firms, consistent with governance frameworks reducing uncertainty or benefiting firms positioned to adopt credible oversight tools. In filing-window tests, firms that disclose concrete AI governance actions experience attenuated market responses to AI safety-risk disclosure. We find the strongest evidence for the financial materiality channel from bias, ethics, and discrimination-related harms. We also show that firms with higher AI safety risk have less favorable abnormal returns when PRI-matched investors or public pension funds are more important shareholders, consistent with a social responsibility channel.

Taken together, the findings show that AI safety risk is an emerging technological externality whose public-market pricing depends on salience, financial materiality, and investor social preferences. Public markets may contribute to disciplining AI safety risk when regulatory, legal, or governance events clarify the expected private costs of AI-related harms, and when norm-constrained investors are attentive to those harms. At the same time, the absence of a stable unconditional premium suggests that market discipline remains incomplete and event-dependent. The findings also imply that disclosure, corporate AI governance, and regulation are complementary mechanisms, with credible governance most relevant when firms disclose concrete actions rather than generic mitigation language. The paper contributes firm-level evidence on how capital markets respond to AI safety risk and provides scalable measures for future research on corporate AI governance and the role of investors in shaping incentives for safer AI development and deployment.

References

Acemoglu, D. (2021). Harms of AI. NBER Working Paper No. 29247.

Acemoglu, D., & Lensman, T. (2024). Regulating transformative technologies. American Economic Review: Insights, 6(3), 359–376.

Acemoglu, D., Aghion, P., Bursztyn, L., & Hemous, D. (2012). The environment and directed technical change. American Economic Review, 102(1), 131-166.

Andrews, I., & Farboodi, M. (2025). Do Markets Believe in Transformative AI? (No. w34243). National Bureau of Economic Research.

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics, 151, 103745.

Baily, M., Brynjolfsson, E., & Korinek, A. (2023). Machines of mind: The case for an AI-powered productivity boom. The Brookings Institution.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131(4), 1593–1636.

Bloom, N., Schankerman, M., & Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. Econometrica, 81(4), 1347–1393.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. Journal of financial and quantitative analysis, 50(4), 623-646.

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk?. Journal of Financial Economics, 142(2), 517-549.

Bolton, P., Li, T., Ravina, E., & Rosenthal, H. (2020). Investor ideology. Journal of Financial Economics, 137(2), 320-352.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. Journal of Political Economy, 121(5), 803-843.

Brown, J. R., Martinsson, G., & Thomann, C. (2022). Can environmental policy encourage technical change? Emissions taxes and R&D investment in polluting firms. The Review of Financial Studies, 35(10), 4518-4560.

Bucknall, B. S., & Dori-Hacohen, S. (2022). Current and near-term AI as a potential existential risk factor. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 119-129).

Bullock, J. B., Chen, Y. C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. M., & Zhang, B. (Eds.). (2024). The Oxford handbook of AI governance. Oxford University Press.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19(1), 396–455.

Cao, S., Jiang, W., Wang, J., & Yang, B. (2024). From man vs. machine to man+ machine: The art and AI of stock analyses. Journal of Financial Economics, 160, 103910.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of Finance, 52(1), 57–82.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv preprint arXiv:2206.13353.

Carvalho, J. P. (2025). The Political-Economic Risks of AI. Available at SSRN 5137622.

Cong, L. W., Liang, T., Zhang, X., & Zhu, W. (2025). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. Management Science, 71(12), 10727-10739.

Fama, E. F. (2014). Two pillars of asset pricing. American Economic Review, 104(6), 1467–1485.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. The Journal of Finance, 47(2), 427–465.

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. The Journal of Finance, 51(1), 55–84.

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. Journal of Economic Perspectives, 18(3), 25–46.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1), 1–22.

Florackis, C., Louca, C., Michaely, R., & Weber, M. (2023). Cybersecurity risk. The Review of Financial Studies, 36(1), 351-407.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5-47.

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter. Retrieved from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gans, J. S. (2025a). Can Socially Minded Governance Control the Artificial General Intelligence Beast?. Management Science, 71(10), 8188-8199.

Gans, J. S. (2025b). How learning about harms impacts the optimal rate of artificial intelligence adoption. Economic Policy, 40(121), 199–219.

Gans, J. S. (2025c). Regulating the direction of innovation. Journal of Public Economics, 246, 105375. doi:10.1016/j.jpubeco.2025.105375

Garel, A., Romec, A., Sautner, Z., & Wagner, A. F. (2024). Do investors care about biodiversity?. Review of Finance, 28(4), 1151-1186.

Garleanu, N., Panageas, S., & Yu, J. (2012). Technological growth and asset pricing. The Journal of Finance, 67(4), 1265-1292.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.

Growiec, J., & Prettner, K. (2025). The Economics of p (doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI. arXiv preprint arXiv:2503.07341.

Guerreiro, J., Rebelo, S., & Teles, P. (2023). Regulating artificial intelligence. NBER Working Paper No. 31921.

Hartzmark, S. M., & Sussman, A. B. (2019). Do investors value sustainability? A natural experiment examining ranking and fund flows. The Journal of Finance, 74(6), 2789-2837.

Harvard Law School Forum on Corporate Governance (2024). Unveiling Key Trends in AI Shareholder Proposals. Retrieved from https://corpgov.law.harvard.edu/2024/09/29/unveiling-key-trends-in-ai-shareholder-proposals/

Harvard Law School Forum on Corporate Governance (2025). Investor Views on AI Oversight: What Do Proxy Votes Tell Us? Retrieved from https://corpgov.law.harvard.edu/2025/05/30/investor-views-on-ai-oversight-what-do-proxy-votes-tell-us/

Hassan, T. A., Hollander, S., Van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. The quarterly journal of economics, 134(4), 2135-2202.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv preprint arXiv:2306.12001.

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423–1465.

Hsu, P. H., Li, K., & Tsou, C. Y. (2023). The pollution premium. The Journal of Finance, 78(3), 1343-1392.

Hui, X., Reshef, O., & Zhou, L. (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. Organization Science, 35(6), 1977-1989.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. Academy of Management Journal, 67(1), 5-32.

Jones, C. I. (2024). The AI dilemma: Growth versus existential risk. American Economic Review: Insights, 6(4), 575–590.

Jones, C. I. (2025). How Much Should We Spend to Reduce AI's Existential Risk? (No. w33602). National Bureau of Economic Research.

Jones, C. I., & Williams, J. C. (1998). Measuring the social return to R&D. The Quarterly Journal of Economics, 113(4), 1119–1135.

Kim, A., Muhn, M., & Nikolaev, V. (2023). From transcripts to insights: Uncovering corporate risks using generative AI. arXiv preprint arXiv:2310.17721.

Kogan, L., & Papanikolaou, D. (2019). Technological innovation, intangible capital, and asset prices. Annual Review of Financial Economics, 11(1), 221-242.

Lin, X., Palazzo, B., & Yang, F. (2020). The risks of old capital age: Asset pricing implications of technology adoption. Journal of monetary economics, 115, 145-161.

Lopez-Lira, A. (2023). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. Journal of Financial Economics, 123(1), 137-162.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. doi:10.6028/NIST.AI.100-1.

Norges Bank Investment Management. (2023). Responsible artificial intelligence.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187-192.

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible investing: The ESG-efficient frontier. Journal of Financial Economics, 142(2), 572-597.

Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv preprint arXiv:2306.13952.

Sautner, Z., van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. The Journal of Finance, 78(3), 1449–1498.

Siano, F. (2025). The news in earnings announcement disclosures: Capturing word context using LLM methods. Management Science, 71(11), 9831-9855.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature machine intelligence, 4(3), 189-191.

World Economic Forum. (2024). Responsible AI playbook for investors. Retrieved from https://resaro.ai/insights/articles/responsible-ai-playbook-for-investors