Intention Inference, Greenwashing, and Asset Prices

Mahamadi Ouoba*
University of Calgary

April 2026

Abstract

Investors evaluating corporate climate actions face a fundamental intention-inference problem: do observed commitments reflect genuine engagement or strategic image management? We develop a Bayesian framework in which investors form credibility priors from prior climate-related communication and update those beliefs upon observing subsequent actions. We show that realized climate actions are more diagnostic of managerial type than prior announcements, implying that otherwise identical actions generate different market reactions depending on inferred credibility, and that misclassifying strategic firms as credible can produce temporary mispricing. We further show that a fixed-precision Bayesian benchmark, in which investors update from disclosures and actions but do not reason about the manager's cost of mimicry, cannot rationalize the joint pattern of cost-asymmetry, cross-venue consistency, and salience-by-cost amplification that we document; an inverse-planning (Bayesian Theory of Mind) interpretation provides a parsimonious account.

To test these predictions, we construct a firm-level measure of inferred managerial credibility based on the specificity, cost acknowledgment, and forward-looking orientation of climate disclosures. Using 4,812 climate-related corporate actions by U.S. firms from 2005 to 2023, we find that credibility amplifies announcement returns: a one-standard-deviation increase in credibility amplifies the three-day abnormal return to a high-strength action by approximately 0.23 percentage points relative to a zero-strength baseline. The benchmark is rejected across three measures of action cost and two fixed-effect structures, with all six triple-interaction estimates positive and significant at the 1% level. The effect persists within firms, attenuates following disclosure standardization shocks, and predicts cross-sectional returns that partially reverse.

These findings suggest that investor beliefs about managerial credibility shape the pricing of climate-related information.

Keywords: Greenwashing, climate disclosures, investor beliefs, intention inference, behavioral asset pricing, climate finance

JEL Codes: G12, G14, G41, Q54, D83

*I thank seminar participants at the University of Calgary, for many helpful comments and suggestions. All errors and omissions are solely mine. University of Calgary. Email: mahamadi.ouoba1@ucalgary.ca.

1 Introduction

Climate-related corporate actions increasingly move capital markets. Yet market reactions to seemingly similar actions vary widely across firms: a net-zero pledge, emissions target, or green investment announcement by one firm can generate a large valuation response, while an economically comparable action by another elicits little reaction. This paper asks why investors price comparable climate actions differently.

We argue that investors face an *intention-inference problem*. They observe climate-action announcements but cannot directly determine whether these actions reflect genuine managerial commitment or strategic image management. Instead, they form beliefs about managerial credibility from prior climate-related communication and update those beliefs when new actions are announced. Similar climate actions can therefore generate different market reactions because investors interpret them through different credibility priors. Since communication style is an imperfect signal of managerial type, some strategic firms may be misclassified as credible. Bayesian updating can then propagate this classification error into prices, generating temporary mispricing that corrects as additional information arrives. A pattern consistent with this mechanism would be a prospective-misclassification gradient: credibility-based pricing rises in the months preceding information revelation events and declines thereafter.

This mechanism is complementary to existing frameworks for pricing climate-related information, but it generates distinct cross-sectional predictions. Preference-based models (Pástor et al., 2021) predict that green assets command premia or discounts because of non-pecuniary investor demand; that channel operates primarily on the level of greenness, not on cross-sectional differences in how a given action is interpreted. Risk-based channels (Bolton and Kacperczyk, 2021) compensate investors for climate-transition exposure; such channels can explain variation in expected returns, but do not by themselves naturally predict that the pricing of similar actions should depend on prior communication credibility, or that this credibility interaction should intensify before greenwashing revelations. The climate-information literature (e.g., Pástor et al., 2022; Engle et al., 2020; Sautner et al., 2023) studies how climate risk information is incorporated into prices once observed. We study an earlier step: how investors assess the credibility of climate-related information before pricing it. The channel therefore operates on cross-sectional variation among firms with similar realized actions, similar observable climate risk, and similar ESG ratings, but different credibility histories.

The core idea is that investors do not treat climate actions as standalone signals. A firm that has produced specific, implementation-oriented climate disclosures—with measurable targets, disclosed costs, and verifiable timelines—carries a stronger reputation for genuine commitment into subsequent actions. A firm relying on vague aspirational language, symbolic pledges, or disclosures lacking operational detail carries a weaker credibility prior. Conditional on action type, economic scope, and announcement format, the same action can therefore generate a different price response because investors attach different probabilities to the action being sincere rather than strategic.

We develop a Bayesian model of investor belief formation to formalize this mechanism. In our framework, investors use prior communication to form beliefs about managerial credibility and update those beliefs upon observing a subsequent climate action. We derive a pricing identity in which the announcement return decomposes as

$$R(D, A) \;=\; \underbrace{[\pi_2(D, A) - \pi_1(D)]}_{\text{belief revision}} \times \underbrace{\Delta(D)}_{\text{quality spread}} \quad ,$$

where $\pi_1(D)$ is the investor's credibility prior formed from prior disclosures D, $\pi_2(D, A)$ is the posterior after

observing action A, and $\Delta(D) = m_S(D) - m_G(D)$ is the investor-perceived valuation difference between a sincere and a strategic firm conditional on disclosure history. The economic intuition is simple: the announcement return equals how much investors revise beliefs about managerial type, scaled by how much those beliefs matter for valuation. This decomposition implies that identical actions can generate different price responses when prior disclosures induce different valuation spreads across firms.

This distinction is central to our mechanism. We do not require high-credibility firms to experience larger belief revisions in every event—a stronger prior can leave less room for updating. Rather, we show that high prior credibility increases the valuation consequences of a given belief revision because investors perceive a larger payoff difference between sincere implementation and strategic mimicry. Over the empirically relevant range of prior credibility, the quality-spread channel dominates, producing a positive interaction between action strength and prior credibility. We map this prediction to an interaction between $ActionStrength_{i,e}$ and the lagged credibility proxy $\hat{\pi}_{i,e-1}$ in the event-study regressions.

Our framework also clarifies the role of misclassification. For genuinely sincere firms, credibility histories help investors identify actions that are more likely to be implemented. For strategic firms mistakenly classified as credible, the same inference process can generate temporary mispricing and subsequent correction at firms whose priors are later falsified. We formalize this as rational belief formation under noisy signals: priors that turn out to be wrong are corrected when new information arrives, without requiring investors to be systematically optimistic about climate claims. We thus generate testable mispricing-and-correction patterns through Bayesian updating, not through behavioral biases in beliefs.

Beyond the core pricing identity, we exploit the inverse-planning structure of our framework to derive additional predictions that distinguish intention inference from reduced-form credibility or fixed-precision signal-extraction models. Standard signal-extraction frameworks treat an action's informativeness as a property of the signal itself. Inverse planning instead asks what the action reveals about the manager's cost-benefit calculation. We show that informativeness depends on the action's cost structure, the consistency of prior communication across venues, and the salience environment in which the action is chosen: costly actions are stronger evidence of sincerity because strategic managers are less willing to incur the cost; consistency across independent disclosure venues is harder to mimic; and when climate attention is high, investors place greater diagnostic weight on costly actions than on cheap ones. We interpret these predictions through a Bayesian Theory of Mind framework in which investors behave as if they infer managerial intent by inverting the manager's choice problem. Our interpretation is observationally close to a class of heterogeneous-precision Bayesian models, but it organizes a set of cost, consistency, and salience predictions not jointly implied by fixed-precision alternatives.

The empirical setting is especially useful because the verification gap between climate claims and observable outcomes is unusually wide. Most climate commitments are voluntary in content even when disclosure is mandatory; firms retain discretion over how ambitiously to frame targets, which baselines to use, and which commitments to emphasize. The verification horizon is also long: net-zero pledges target mid-century outcomes, intermediate emissions targets take years, and operational transition plans unfold over multi-year periods. Finally, the enforcement apparatus for ESG claim accuracy is still developing relative to the mature enforcement regime for financial disclosure. Greenwashing revelations in our sample often arrive through NGO accusations, investigative media, or ESG-rating downgrades rather than through formal regulatory enforcement. This is a difference of degree, not kind. Financial disclosure also contains discretion and delayed verification, but ESG is a setting in which this gap is especially large, making credibility inference particularly important. This combination makes the ESG setting particularly well suited to studying how

credibility inference affects asset prices.

For the remainder of the paper, "inferred credibility" refers to investors' posterior assessment that a firm's climate actions reflect genuine commitment rather than strategic signaling. We treat this as a reduced-form belief about managerial intent and expected implementation discipline. This interpretation matches what investors observe and what the model prices. To assess whether the channel merely captures better executors writing better disclosures, we report a residualized version of the proxy that projects $\hat{\pi}$ onto the full vector of observable environmental quality available before the announcement. The interaction remains large and statistically significant, indicating that the credibility channel is not solely an observable-quality channel.

Empirically, we construct an ex-ante credibility signal from climate-related communication in the 12 months preceding each announcement. The proxy captures specificity, cost acknowledgment, and forward-looking implementation orientation in prior climate communications. Cross-venue consistency enters the paper as a related but distinct construct in two roles: a firm-year, four-venue $Consistency_{i,e-1}$ used in BToM heterogeneity tests, and a quarterly, three-venue cross-venue $Shock_{i,t-1}$ used in the within-firm credibility-shock test. The credibility proxy is constructed using supervised text classification trained on manually labeled climate-related paragraphs and applied to pre-event disclosure windows; a rolling-classifier specification re-estimates the classifier using only pre-t labeled paragraphs as a strict-out-of-time robustness check. *ActionStrength* is coded independently by trained research assistants using a pre-registered blinded rubric, applied without knowledge of $\hat{\pi}$ or stock returns. The coding protocol was completed before any return analysis. Although both measures capture aspects of climate communication, they are constructed from non-overlapping event-time information sets: prior communication history for credibility and announcement-date action characteristics for action strength.

Our sample covers 4,812 climate-related corporate action events by U.S. publicly traded firms from 2005 to 2023, matched to a panel of firm-year ESG communication. We estimate event-study regressions of announcement returns and find that market reactions are larger when prior communication is more implementation-oriented. In our preferred specification, a one-standard-deviation increase in inferred credibility is associated with a 0.23 percentage-point larger announcement return to a high-strength action. In economic terms, the price response to an observationally comparable climate action at a firm in the upper quartile of inferred credibility is larger than the response at a firm in the lower quartile. The credibility proxy also predicts subsequent emissions reductions, green capital expenditures, and adoption of verifiable climate targets out of sample, passes placebo tests using non-climate corporate announcements, and remains predictive after residualizing on observable environmental quality.

We next examine asset-pricing implications. The joint signal of strong climate actions and high inferred credibility predicts the cross-section of stock returns. Portfolios formed on this signal earn economically meaningful positive alphas under the CAPM, attenuating but remaining positive under multi-factor adjustments. These returns subsequently reverse, consistent with belief correction. Reversals are larger among firms later revealed as greenwashers and among firms with ambiguous or inconsistent prior communication, linking the long-horizon pattern to misclassification risk. We treat the reversal evidence as corroborating rather than primary identification (Section 7); the announcement-window interaction and the quasi-experimental designs in Section 8 carry the load-bearing identification.

Additional tests support the mechanism. Within-firm changes in inferred credibility predict announcement returns, while future values of credibility are insignificant in level and interaction placebos. A non-climate-disclosure placebo reinforces specificity: when prior climate communication is replaced with non-ESG corporate disclosures from the same firms, the interaction falls to an economically negligible level, roughly

an order of magnitude below the main estimate. Follow-through implementation events exhibit a larger credibility interaction than initial announcements, despite lower average returns and lower return dispersion, consistent with implementation signals from high-credibility firms receiving higher diagnostic weight. Climate actions are followed by more implementation-oriented disclosures, providing supporting evidence that firms update communication after observable action. The credibility-pricing interaction is smaller after disclosure standardization and larger in settings where intention inference is more consequential, including carbon-intensive firms, firms with low analyst coverage, and firms with high earnings uncertainty.

A novel empirical pattern in our data is a *prospective-misclassification gradient*: the credibility-pricing interaction is larger in the months preceding a greenwashing revelation event and lower thereafter. This prospective pattern complements the firm-side ex-post evidence in Duchin et al. (2025), who document that firms reallocate pollutive assets without reducing aggregate pollution and that ESG ratings detect greenwashing only belatedly. Our evidence shows that investors also overweight credibility priors before the correction arrives.

First, we formalize a fixed-precision Bayesian credibility benchmark and derive testable restrictions that separate it from inverse-planning accounts. We reject the benchmark across three operationalizations of action cost (binary high-cost indicator, continuous capex-intensity, third-party verifiability) and two fixed-effect structures, with all six triple-interaction estimates positive and significant at the 1% level. The rejected restrictions are naturally organized by an inverse-planning interpretation in which investors reason about the cost of mimicry the strategic-type manager faces—a Bayesian Theory of Mind (BToM) mechanism.

Second, we document a *prospective-misclassification gradient*: the credibility-pricing interaction is larger in the months preceding a greenwashing revelation event and lower thereafter. The pattern is more pronounced among firms later revealed as greenwashers, providing investor-side evidence consistent with credibility priors being temporarily misapplied before the correction arrives.

Third, the credibility-pricing interaction varies systematically with the regulatory and informational environment. It is smaller after disclosure standardization, while it is larger during periods of elevated climate salience (measured from financial news and earnings calls). The salience effect is concentrated on costly-to-mimic actions, which is the discriminating prediction: baseline salience amplification is also consistent with rational signal-to-noise adjustment, but salience-by-cost amplification is more consistent with intention inference.

Fourth, we contribute a supervised text classifier for managerial credibility in ESG communication. The classifier achieves an inter-rater agreement of $\kappa = 0.71$ on the high-agreement training sample, with full performance characteristics reported below. The measure is potentially adaptable to other ESG settings, subject to retraining.

Taken together, the results suggest that markets price not only what firms do, but also what investors infer about the credibility behind those actions.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 develops the Bayesian model. Section 4 presents the data. Section 5 describes the construction and validation of the credibility proxy. Section 6 presents the main event-study results, including follow-through and belief-updating tests. Section 7 constructs the Intent Factor and examines cross-sectional return predictability. Section 8 exploits dated greenwashing-revelation events as quasi-experimental timing evidence for belief updating, reports quasi-experimental evidence from disclosure standardization, and presents salience-amplification and heterogeneity mechanism tests. Section 9 presents robustness checks, and Section 10 concludes.

2 Related Literature

This paper connects four strands of research—climate and ESG asset pricing, greenwashing and sustainability disclosures, investor beliefs and behavioral asset pricing, and social cognition and intention inference—to study how investors form firm-specific beliefs about managerial credibility from prior climate communication and how those beliefs shape the pricing of subsequent climate actions.

2.1 Climate and ESG Asset Pricing

Climate risk and ESG characteristics are priced in financial markets. Bolton and Kacperczyk (2021) document a carbon premium; Ilhan et al. (2021) show that climate transition risk appears in option-implied volatility around climate-policy events; Pástor et al. (2021) develop an equilibrium model in which preferences for green assets depress green stock returns in normal times and generate positive shocks when climate concerns intensify; this builds on Heinkel et al. (2001), who model the equilibrium asset-pricing consequences of green investor exclusion. Matsumura et al. (2014) provide early evidence that investors penalize emissions and value voluntary carbon disclosure, and Hartzmark and Sussman (2019) document that Morningstar sustainability rankings generated $12 billion in outflows from low-ranked funds and $24 billion in inflows to high-ranked funds. Edmans (2011) provides earlier evidence that intangible social-responsibility characteristics (employee satisfaction) predict subsequent abnormal returns, suggesting that the market incorporates difficult-to-verify firm attributes only with delay—a pattern consistent with the slow correction we document in Section 7. Krueger et al. (2020) show that institutional investors view climate risks as already materializing and treat engagement, not divestment, as the dominant response channel. Starks (2023) distinguishes "value" from "values" motivations in ESG investing; our framework operates on the "value" margin. Giglio et al. (2025) document substantial heterogeneity in investor beliefs about ESG and climate risks, generating cross-sectional variation in portfolio allocations.

A parallel theoretical literature embeds ESG preferences into rational-expectations frameworks. Goldstein et al. (2022) and Zhou and Kang (2023) show that investors with opposing ESG preferences trade in opposite directions on the same signal, with consequences for price informativeness; Ji et al. (2023) shows that improved ESG disclosure precision can reduce welfare for small investors under asymmetric information.

This literature largely abstracts from how investors form firm-specific beliefs about the credibility of ESG disclosures. We document that observationally similar climate actions generate materially different market reactions depending on what investors infer about the credibility behind them.

2.2 Greenwashing and Sustainability Disclosures

A substantial literature documents that corporate sustainability disclosures frequently diverge from realized climate outcomes (Delmas and Burbano, 2011; Lyon and Montgomery, 2015). A conceptual debate concerns whether deliberate intent is necessary for a disclosure to qualify as greenwashing (Torelli et al., 2020; Seele and Gatti, 2017); our framework adopts an intent-based distinction between strategic and sincere communication as a modeling device. The gap between stated and realized environmental behavior is strategically managed (Teti et al., 2024) and is associated with governance frictions (Ghitti et al., 2024). The same pattern appears at the portfolio-manager level: Raghunandan and Rajgopal (2022) document that self-labeled ESG mutual funds hold stocks with worse compliance records and higher carbon intensity than non-ESG funds offered by the same asset manager. ESG ratings themselves are a noisy aggregate signal: Berg et al. (2022) attribute

56% of cross-rater disagreement to measurement divergence, and Avramov et al. (2022) show that ESG rating uncertainty changes sustainable-investing portfolio and asset-pricing implications.

Duchin et al. (2025) identify a mechanism by which firms improve ESG ratings without reducing aggregate pollution: they divest polluting assets to buyers facing weaker regulatory pressure. Akey and Appel (2021) show that strengthening parent limited-liability protection caused a 5–9% increase in toxic emissions by subsidiaries through reduced abatement investment.

The strategic/sincere distinction maps onto signaling (Spence, 1973) and cheap-talk models. Flammer (2021) documents that corporate green bond issuance generates positive announcement returns and predicts subsequent environmental performance improvements, consistent with costly green investment serving as a credible signal—the same costly-signaling logic our BToM model formalizes at the level of individual climate actions. Bingler et al. (2024) apply a deep-learning classifier to TCFD reports to measure cheap talk and document that institutional ownership reduces it. Our paper differs in measurement object ($\hat{\pi}_{i,e-1}$ measures communication style across the firm's entire prior climate corpus rather than precise-vs-imprecise content within a single document), in pricing specification (we study the *interaction* between credibility prior and action strength), and in identification (CSRD-attenuation, salience amplification, distance-to-revelation gradient, and frozen-prior DDD around greenwashing-revelation events). Related work documents that linguistic opacity in ESG reports is associated with greenwashing risk (Kathan et al., 2025) and that mandatory standardized ESG metrics improve disclosure quality (Christensen et al., 2021), who also emphasize that the welfare effects of mandates depend on whether they shift the credibility of disclosure content as well as the volume of disclosure—a distinction central to our standardization-attenuation result.

Recent empirical work examines capital-market consequences of detected greenwashing. Teti et al. (2024) find no significant average CAR around greenwashing announcements for European firms but document a negative association with prior climate-performance ratings. Gourier and Mathurin (2025) construct a news-based greenwashing salience index and document three distinct attention waves, with the most recent (2018–2022) driven by suspicions toward the financial sector; institutional investors respond by tilting toward certified high-sustainability funds. Grewal et al. (2026) provide causal evidence that mandatory greenhouse-gas reporting reduced carbon greenwashing by 30–44% and noncarbon greenwashing by 23–40% among UK-listed firms. We exploit both the salience and standardization margins in our identification strategy.

The existing literature on greenwashing focuses on firm-side strategies and *ex-post* penalties. We study an ex-ante investor-side channel: whether investors form credibility beliefs from ongoing ESG communication and condition the pricing of subsequent climate actions on those beliefs, before any mismatch is revealed. Our evidence complements the firm-side findings of Duchin et al. (2025) with systematic investor-side evidence on real-time intention inference.

2.3 Investor Beliefs and Behavioral Asset Pricing

A large behavioral literature documents that investor beliefs deviate from the rational benchmark, with consequences for prices: investors underreact to gradual information and overreact to salient events (Barberis et al., 1998), and limited attention generates predictable return patterns around information events. A separate strand studies belief heterogeneity *across investors*: when investors interpret the same signal differently, trading volume and price impact depend on the distribution of beliefs rather than their mean (Harrison and Kreps, 1978; Meeuwis et al., 2022).

Our framework features a different object: a representative investor whose beliefs about managerial credibility vary *across firms* based on each firm's communication history, generating cross-sectional variation in

announcement returns through Bayesian updating rather than through cross-investor disagreement or behavioral biases. The updating itself is rational; the priors may nonetheless embed systematic misclassification, generating the prospective gradient and long-horizon reversal documented in Section 7.

2.4 Social Cognition and Intention Inference

Inferring others' intentions is a fundamental and automatic feature of human cognition. Saxe and Kanwisher (2003) and Saxe (2005) identify a dedicated neural system—centered on the temporoparietal junction—that activates when individuals evaluate the beliefs and intentions behind observed behavior. Behavioral consequences are documented: moral judgments depend on perceived intent (FeldmanHall et al., 2012), and reciprocity in economic experiments conditions on inferred intentions (Charness and Rabin, 2002).

This mechanism has received little formal treatment in asset pricing. Studies of corporate-communication credibility examine disclosure accuracy rather than investor beliefs about the credibility behind disclosures. De Martino et al. (2013) show in experimental bubble markets that ToM reasoning is active in financial decision-making and can systematically bias value computations. We extend this evidence from the laboratory to real ESG markets, introducing intention beliefs as a measurable determinant of market reactions to corporate announcements. The next section formalizes this intention-inference mechanism in a Bayesian model that delivers the empirical predictions tested in the remainder of the paper.

3 Model: Intention Inference in ESG Disclosures

This section develops a tractable Bayesian model of investor belief formation about managerial intentions. We use the model to derive four empirical outputs: a closed-form return decomposition (Eq. (20)) that maps directly into the event-study specification; a positive credibility-pricing interaction over the empirically relevant parameter region (Proposition 2); BToM-specific predictions that discriminate intention inference from reduced-form alternatives (BToM Corollaries); and testable standardization and reversal predictions (Proposition 4 and the empirical reversal tests in Section 7). The micro-foundations establish three properties: first, the level derivative $dR/dD > 0$ holds under Assumption A1 (the quality-spread channel dominates), which is satisfied over the empirically relevant support (Proposition 1); second, the total derivative of the *marginal return impact* $d/dD(\partial R/\partial A)$ remains positive throughout the empirically relevant support because the quality-spread channel dominates even where the underlying cross-partial of the belief-revision function h is negative (that cross-partial is positive analytically whenever $\pi_1 < 1/(1+\ell(A))$, equivalently $\pi_2 < 1/2$) (Proposition 2); and third, disclosure standardization compresses cross-type heterogeneity and thereby weakens the credibility-pricing channel (Proposition 4). All proofs are in Appendix B.

Relation to the rest of the paper. This section develops the paper's unified theoretical framework. We derive four main propositions, two corollaries (Corollaries 1 and 2), and three BToM-specific corollaries. The propositions establish the core pricing mechanism of intention inference, while the predictions deliver sharp empirical implications arising from the Bayesian Theory of Mind (BToM) inverse-planning structure. The results map directly onto the empirical analysis as follows:

- **Propositions 1 and 2** (greenwashing discount and disclosure-dependent amplification of action returns) motivate the baseline event-study regressions in Section 6 (Table 15) and the follow-through tests (Table 16).
- **Proposition 3** (dynamic belief updating) motivates the belief-updating regressions (Table 17).

- **Proposition 4** provides the theoretical motivation for the quasi-experimental design based on disclosure-standardization shocks; the empirical implementation focuses on attenuation of the credibility-pricing interaction following these regulatory shocks (Table 18 and Figure OA-1).
- **Corollary 1** (cross-sectional return dispersion) underpins the heterogeneity analysis (Table 22).
- **Corollary 2** (credibility prior as a pricing state variable) motivates the quasi-experimental and revelation-shock designs in Section 8.
- **BToM Corollaries 1–3** (cost-asymmetry amplification, consistency amplification, and salience-by-diagnosticity interaction) generate the BToM-specific corollaries tested in Section OA-2.3 (Table 27). These three predictions rely on the model's cost-asymmetry and inverse-planning features and go beyond what standard signal-extraction frameworks deliver.

All formal proofs are collected in Appendix B.

3.1 Environment and Microfoundations

Timing. There is a continuum of risk-neutral investors and a continuum of firms. The model is presented for a representative firm; all cross-sectional predictions apply across this continuum. The model has three dates: $t = 0$ (managerial action choice), $t = 1$ (ESG disclosure and investor belief updating), and $t = 2$ (public implementation signal observed and asset payoff realized).

Timing convention. All lagged variables are constructed using calendar-time windows. $\hat{\pi}_{i,e-1}$ uses disclosures in the 12 months prior to the event date, and $\Delta\hat{\pi}$ differences two trailing 12-month disclosure windows; the two windows are non-overlapping when the underlying anchor dates are separated by at least 12 calendar months, which holds for the large majority of firms in the within-firm sample (Section 8.1 reports robustness to restricting to firms with events spaced at least 12 months apart). The index e orders events within a firm but does not define the lag structure; lags are always measured in calendar time.

Empirical interpretation of timing. In the empirical implementation, the observed climate-action announcement corresponds to the public revelation of the realized implementation signal A at event time, rather than the private choice of the target intensity α_τ^* made at $t = 0$. Accordingly, $ActionStrength_{i,e}$ is interpreted as a proxy for the observable realization A that updates investor beliefs at the announcement date, consistent with the $t = 2$ information structure of the model. The three-day CAR around the announcement window is the empirical analogue of the model's announcement return $R(D, A) = P_2 - P_1$.

The three-date model is stylized to isolate the intention-inference mechanism around a single climate-action announcement. The liquidating dividend at $t = 2$ is a modeling device that delivers a tractable pricing identity; we interpret it as the present value of future cash flows, not as literal firm termination. Long-horizon return dynamics at six- and 12-month horizons (Section 7) are outside the scope of the one-shot structure. Note that the model's internal ordering ($t = 0$: action choice; $t = 1$: disclosure issued; $t = 2$: action publicly observed) places the action target α_τ^* before the disclosure D — the reverse of the empirical sequence, where $\hat{\pi}_{i,e-1}$ is constructed from disclosures made in the 12 months *before* the announcement. This reflects a deliberate modeling choice: $t = 1$ disclosure is best interpreted as the accumulated communication history that investors use to form prior beliefs π_1, while the $t = 2$ announcement is the new action that triggers updating. The model's causal logic (type → communication style → investor prior → action pricing) maps correctly to the empirical design, even though the calendar ordering within the model does not literally match the real-world timeline. Rigorously, the mapping is as follows: investors at the announcement date

e hold a prior about managerial type formed from disclosures released over the preceding 12 months; the model's $t = 1$ object D stands for this accumulated prior disclosure signal, and the model's $t = 2$ action A is the announcement itself. The model's $t = 0$ action-choice stage is a modeling device used to motivate why sincere and strategic types differ in action scale ($\alpha_S^* > \alpha_G^*$); it does not correspond to an observable calendar date in the empirical design, and no empirical test exploits the $t = 0$ choice directly. Accordingly, we interpret the empirical interaction coefficient β_3 as an analogue of the model's comparative static, not as a structural estimate of a stage-game parameter.

Managerial type. The manager is privately informed about her own type $\tau \in \{S, G\}$, where S denotes *sincere* environmental commitment and G denotes *strategic greenwashing*. The model explicitly treats greenwashing as deliberate strategic behavior: the type-G manager knowingly exploits the reputational benefit v while bearing the higher implementation cost $c_G > c_S$, rather than merely disclosing inaccurately due to measurement error or incompetence. Investors hold a common prior

$$\pi_0 \;=\; P(\tau = S) \;\in\; (0, 1). \tag{1}$$

Environmental quality. The firm has a latent environmental quality $\theta \in \mathbb{R}$ representing its long-run transition-risk profile. Environmental quality depends on managerial type:

$$\theta \mid \tau = S \;\sim\; \mathcal{N}(\mu_S, \sigma_\theta^2), \qquad \theta \mid \tau = G \;\sim\; \mathcal{N}(\mu_G, \sigma_\theta^2), \tag{2}$$

where $\mu_S > \mu_G \geq 0$ and $\sigma_\theta^2 > 0$.

Microfoundation for $\alpha_S > \alpha_G$. At $t = 0$, the manager privately chooses a target action intensity $\alpha \geq 0$ that determines the scale of her subsequent climate program. Her net payoff is

$$B(\alpha, \tau) \;=\; \underbrace{v\,\alpha}_{\text{reputational benefit}} \;-\; \underbrace{c_\tau \frac{\alpha^2}{2}}_{\text{implementation cost}}, \tag{3}$$

where $v > 0$ is the common marginal reputational benefit from a larger climate program and $c_\tau > 0$ is the type-specific marginal cost of implementation. We treat v as an exogenous reduced-form scalar that summarizes the net reputational gain per unit of climate program scale, abstracting from the equilibrium mapping between actions and investor posteriors.

Normalization of the action signal scale. Substituting $\alpha_\tau^* = v/c_\tau$ into the Gaussian likelihood ratio for action A yields

$$\log \frac{f_S(A)}{f_G(A)} \;\propto\; (A - \alpha_G^*)^2 - (A - \alpha_S^*)^2 \;=\; 2A\,v\left(\frac{1}{c_S} - \frac{1}{c_G}\right) - v^2\left(\frac{1}{c_S^2} - \frac{1}{c_G^2}\right),$$

so v does *not* cancel from the inference problem: it scales the diagnosticity of observed actions. We normalize $v = 1$ without loss of generality. All comparative statics and empirical predictions depend only on the cost differential $(c_G - c_S)$, so this normalization fixes the scale of actions while preserving all economic content of the model. The BToM inverse-planning inference remains valid under this normalization: investors infer type from action intensity relative to the cost structure, with the scale fixed by $v = 1$. The key asymmetry

is

$$c_S \;<\; c_G. \tag{4}$$

A sincere manager, who has already built operational expertise in decarbonization and obtains intrinsic utility from environmental improvement, bears a lower marginal cost of real action than a greenwasher, who must divert resources from her core profit-maximization objective to mount a visible environmental program. Maximizing $B(\alpha, \tau)$ over α gives optimal intensities:

$$\alpha_S^* \;=\; \frac{v}{c_S} \;>\; \alpha_G^* \;=\; \frac{v}{c_G}. \tag{5}$$

The ordering $\alpha_S > \alpha_G$ is therefore an endogenous implication of the optimal cost-benefit calculation under $c_S < c_G$, not a direct assumption. Because this is a single-agent optimization with no strategic interaction at this stage, the result is more precisely described as an optimal-choice implication than an equilibrium condition.

At $t = 2$, the market observes a public implementation signal of the manager's climate program. This signal equals the target intensity chosen at $t = 0$ plus idiosyncratic implementation noise that arises from contracting frictions, weather shocks, and supply-chain disruptions:

$$A \;=\; \alpha_\tau^* + \xi, \qquad \xi \;\sim\; \mathcal{N}(0,\, \sigma_A^2), \tag{6}$$

where ξ is independent across firms and of all other randomness in the model. Writing $\alpha_\tau \equiv \alpha_\tau^*$ for brevity, the observable action is therefore distributed as

$$A \mid \tau \;\sim\; \mathcal{N}(\alpha_\tau,\, \sigma_A^2), \qquad \sigma_A^2 > 0, \tag{7}$$

with common variance σ_A^2 across types (maintained for tractability; see Section OA-1.5). Investors observe A but not τ, so A serves as a noisy signal of the unobserved target α_τ^* and hence of type.

ESG disclosure ($t = 1$). At $t = 1$, the manager issues an ESG disclosure

$$D \;=\; \theta + \eta, \qquad \eta \mid \tau \;\sim\; \mathcal{N}(0,\, \sigma_{\eta,\tau}^2), \tag{8}$$

where η is type-dependent communication noise, independent of ξ and ν (the payoff noise defined below in Equation (10)). We additionally maintain that η is independent of the manager's action choice α_τ^* conditional on type τ: while the disclosure at $t = 1$ is issued *after* the action target is chosen at $t = 0$, the communication noise η reflects the manager's disclosure style rather than the specific action quantity—consistent with the interpretation that D signals environmental quality θ with type-dependent fidelity, not the planned action scale. Because $\alpha_\tau^* = v/c_\tau$ is a deterministic constant conditional on type τ, the disclosure noise η is trivially independent of α_τ^* given τ; signal independence between D and A follows from the assumption that the implementation noise ξ is independent of η and θ, ensuring the two channels provide distinct information about managerial type. Greenwashing disclosures are noisier than sincere ones:

$$0 \;<\; \sigma_{\eta,S}^2 \;<\; \sigma_{\eta,G}^2. \tag{9}$$

A sincere manager reports her actual environmental operations with high fidelity; a greenwasher layers strategic framing, selective emphasis, and aspirational language on top of the environmental reality, generating additional noise in the mapping from θ to D. The action scale α_τ^* is governed by the implementation cost parameter c_τ (Equation (5)), while the disclosure noise $\sigma_{\eta,\tau}^2$ is a distinct type-dependent primitive unrelated to c_τ. These two channels—action scale and disclosure style—provide independent signals about managerial type. The action A is the subject of strict inverse-planning inference: because $A = \alpha_\tau^* + \xi$ and $\alpha_\tau^* = v/c_\tau$ is the manager's *optimized* choice, investors invert the cost-benefit problem to infer c_τ from the observed scale. The disclosure D, by contrast, is modeled as a mechanical signal of environmental quality θ with type-dependent noise $\sigma_{\eta,\tau}^2$—a primitive the manager cannot freely manipulate in the baseline model. Inference on D therefore involves standard Bayesian signal extraction rather than inverse planning in the strict sense; BToM provides the structural motivation for why the noise parameters differ by type. While this ordering is treated as a primitive in the baseline model to maintain tractability, Appendix OA-3 provides a simple microfoundation in which disclosure precision is endogenously chosen as a function of type-dependent costs and benefits, delivering $\sigma_{\eta,S}^2 < \sigma_{\eta,G}^2$ as an equilibrium outcome.

Asset payoff. At $t = 2$, the asset pays a liquidating dividend[1]

$$V_2 \;=\; \theta + \nu, \qquad \nu \sim \mathcal{N}(0,\, \sigma_\nu^2), \tag{10}$$

where ν is independent of all other randomness. With risk-neutral investors and a zero discount rate, the stock price at each date equals the conditional expectation of θ (the payoff noise ν is mean-zero and independent of all information, so it drops out of the expectation):

$$P_1 \;=\; \mathbb{E}[\theta \mid D], \qquad P_2 \;=\; \mathbb{E}[\theta \mid D,\, A]. \tag{11}$$

3.2 Intention Inference as Bayesian Inverse Planning

The core mechanism we model is *intention inference*: we assume investors behave as if observing the manager's disclosure D and subsequent action A and solving an *inverse planning* problem to infer the hidden managerial type τ. This formulation maps directly onto the computational framework of *Bayesian Theory of Mind* (BToM) developed in cognitive science (Baker et al., 2009; Jara-Ettinger et al., 2016). In BToM, observers treat an agent's observable outputs as the product of an internal planning process (beliefs $\times$ desires $\rightarrow$ actions) and invert that process via Bayesian inference to recover the latent mental states that generated the observed behavior. In our setting the disclosure D is a mechanical type-dependent signal (Equation (8)); only the target action intensity α_τ^* results from active optimization (Equation (5)). Inverse planning therefore recovers the managerial type τ from two complementary sources: the *style* of D (captured by the noise structure $\sigma_{\eta,\tau}^2$) and the *scale* of the realized action A (driven by c_τ). Both are informative precisely because they reflect the latent type τ and its associated parameters (c_τ, $\sigma_{\eta,\tau}^2$) that the manager cannot freely manipulate, consistent with the naive utility calculus framing of Jara-Ettinger et al. (2016).

Formally, investors maintain a generative model of managerial behavior:

[1]The one-shot three-date structure abstracts from multi-period learning dynamics to deliver a closed-form expression for announcement returns. In a richer dynamic environment, investors would continue to update beliefs about managerial type as additional signals arrive, implying that initial price reactions need not be permanent. The long-horizon return reversals documented empirically are consistent with such gradual belief updating and do not require modeling the full dynamic process explicitly for the event-study predictions.

(i) The manager is of type $\tau \in \{S, G\}$ with prior π_0.

(ii) Given τ, she chooses target intensity α_τ^* (Equation (5)).

(iii) She issues a noisy disclosure D (Equation (8)) and produces a noisy implementation signal A (Equation (6)).

Observing D and A, investors compute the posterior $\pi_2(D, A) = P(\tau = S \mid D, A)$ via Bayes' rule (Section 3.4). This posterior is an intention-inference posterior: it represents the probability that the observed sequence of communication and action was generated by a sincere (low-cost, low-noise) rather than strategic (high-cost, high-noise) planning agent.

This mechanism also connects to the financial neuroscience literature showing that ToM reasoning is active when investors interpret the actions of other market participants (De Martino et al., 2013). Our model formalizes this cognitive operation in the climate domain: the announcement return $R(D, A)$ is the market's equilibrium response to real-time intention inference about the manager behind the climate action. The BToM framing provides the cognitive microfoundation for the cross-sectional predictions in Section 3.3: because investors solve an inverse planning problem, they place differential weight on costly, consistent, and salient actions. The mathematical mechanics of the cost-asymmetry prediction—larger cost differentials increase the separation between type-specific action means, raising diagnosticity—are consistent with standard Bayesian updating and signaling theory (e.g., Spence, 1973). What BToM adds is the structural motivation for why these parameters take the values they do (the manager's cost-benefit optimization), rather than treating signal precision as an exogenous primitive.

Scope of the BToM apparatus. Propositions 1 through 4 and Corollaries 1 and 2 are derived from a Gaussian updating model that is consistent with standard Bayesian inference and does not require the full inverse-planning apparatus of Bayesian Theory of Mind. The additional corollaries in Section 3.3 exploit the cost-asymmetry structure of the model to generate cross-sectional predictions about costly actions, communication consistency, and salience, which standard signal-extraction models do not deliver. These corollaries represent the model's sharpest empirical content and are tested directly in Section 8.

3.3 Inverse-Planning Predictions: Cost Asymmetry, Consistency, and Salience

The following predictions are stated here to motivate the empirical design. The core Gaussian model delivers Propositions 1–4. Three additional predictions follow from the cost asymmetry between sincere and strategic managers ($c_S < c_G$) under inverse-planning interpretation. The formal contrast between these inverse-planning predictions and a fixed-precision Bayesian benchmark—in which moderators affect type-separation but not the residual triple-interaction—is developed in Section OA-1.1; the empirical adjudication is in Tables 27, 28, and 29. Because sincere managers face lower marginal costs of climate investment, *actions* that impose higher differential implementation costs on the strategic (greenwashing) type are more diagnostic of sincerity. In the inverse-planning interpretation, such costly-to-mimic actions receive greater weight in belief updating and valuation. Disclosures, by contrast, enter as a mechanical type-dependent signal (see Equation (8)) with lower noise for sincere managers ($\sigma^2_{\eta,S} < \sigma^2_{\eta,G}$), providing an independent channel for inferring managerial type through differences in precision rather than differential costs.

BToM Corollary 1 (Cost-Asymmetry Amplification)**.** *The coefficient on ActionStrength* $\times \pi_1$ *—capturing how the return amplification from action strength increases with prior credibility beliefs—is larger for actions*

that impose higher differential implementation costs on the strategic (greenwashing) type. In inverse planning, such actions are more diagnostic of low c_τ, leading observers to revise beliefs about managerial intent more strongly.

BToM Corollary 2 (Consistency Amplification)**.** *The credibility-pricing interaction is stronger when prior ESG communication is consistent across venues (sustainability reports, 10-K climate sections, earnings calls, and press releases), holding the level of $\hat{\pi}_{i,e-1}$ fixed.*

Consistency enters as a signal-precision moderator: for two firms with the same observed $\hat{\pi}_{i,e-1}$, investors treat the text-based credibility score as a more diagnostic summary of underlying type when prior communication is cross-venue consistent. In the Bayesian framework, higher consistency corresponds to lower effective disclosure noise $\sigma^2_{\eta,\tau}$ at the firm level, raising the precision κ_τ with which prior disclosure informs type. Under Proposition 2, a more informative disclosure history amplifies the return response to a given action through both the belief-revision and quality-spread channels. The contrast with the fixed-precision benchmark is developed in Section OA-1.1.

BToM Corollary 3 (Salience-by-Diagnosticity Interaction)**.** *The amplification of the credibility channel during periods of high climate salience (Attention$_t$) is larger for highly diagnostic (costly-to-mimic) actions. When the domain is salient, investors engage in more effortful inverse planning and place greater weight on actions that strongly reveal type.*

These predictions are generated by inverse-planning interpretation of investor reasoning. Section OA-1.1 formalizes a fixed-precision Bayesian competitor and shows that, conditional on the appropriate two-way terms, it predicts the triple-interaction coefficients should be zero or negative. Inverse planning predicts strictly positive coefficients. We test the cost-asymmetry, consistency, and salience-by-diagnosticity predictions in Table 27 (Section OA-2.3); Tables 28 and 29 report robustness of the cost-asymmetry result to a continuous capex-intensity gradient and an alternative cost operationalization based on third-party verifiability. Table 22 reports the baseline heterogeneity results (dirty industries, analyst coverage, earnings uncertainty) that motivate these tests. The consistency prediction uses the cross-venue cosine similarity measure of Reimers and Gurevych (2019).

3.4 Belief Updating

Posterior after disclosure. Let $\phi(\cdot\,; \mu, \sigma^2)$ denote the Gaussian density. Marginalizing over θ, the type-conditional densities of D are

$$f_\tau(D) \;=\; \phi\big(D;\, \mu_\tau,\, \sigma_\theta^2 + \sigma_{\eta,\tau}^2\big)\,, \qquad \tau \in \{S, G\}. \tag{12}$$

Bayes' rule gives

$$\pi_1(D) \;\equiv\; P(\tau = S \mid D) \;=\; \frac{\pi_0\, f_S(D)}{\pi_0\, f_S(D) + (1-\pi_0)\, f_G(D)}. \tag{13}$$

This posterior after disclosure is operationalized empirically by $\hat{\pi}_{i,t}$, constructed in Section 5 via supervised text classification of ESG communication. The proxy is a noisy, theory-motivated measure of the belief investors hold after observing prior disclosure history, rather than its exact structural counterpart. In the second-stage updating problem after the action is observed, $\pi_1(D)$ serves as the effective prior. The posterior $\pi_2(D, A)$ is not directly observed but is reflected in market reactions.

Posterior after action. The posterior belief after observing the action is obtained by a second application of Bayes' rule:

$$\pi_2(D, A) \equiv P(\tau = S \mid D, A) = \frac{\pi_1(D)\,\ell(A)}{\pi_1(D)\,\ell(A) + (1 - \pi_1(D))}, \tag{14}$$

where $\ell(A) = g_S(A)/g_G(A)$ is the likelihood ratio of the observed action under the sincere versus strategic type, and $g_\tau(A)$ denotes the conditional density of A given type τ, as implied by Equation (7). Because sincere managers face lower marginal costs of environmental investment, $\ell(A)$ is strictly increasing in action strength. The structural analog in the asset-pricing literature is the Bayesian Updating Speed (BUS) of Han et al. (2023), defined as the ratio of signal precision to total precision $\tau_q/(\tau_c + \tau_q)$. Conceptually, our setting parallels this framework—prior credibility plays a role analogous to a prior on managerial type, and the action's diagnosticity plays a role analogous to signal informativeness—but the analogy is interpretive, not a literal mathematical mapping: $\hat{\pi}_{i,e-1}$ is a probability rather than a precision parameter, and $\ell(A)$ is a likelihood ratio rather than an inverse variance. The claim is only that disclosure standardization compresses the diagnosticity of the prior credibility signal in a way that has the same qualitative effect on belief updating as a reduction in signal precision in Han et al. (2023)'s Gaussian-precision framework. Under this interpretive mapping, **H4**—that disclosure standardization attenuates the credibility-pricing interaction—can be read as the qualitative analog of a reduction in BUS, providing a conceptual complement to the measurement-attenuation argument in Appendix OA-4.1. The belief $\pi_1(D)$ acts as the effective prior: the same realization of A shifts posteriors by different amounts depending on π_1. Figure 1 provides a visual illustration of this updating process for a representative parameterization.

3.5 Valuation and the Announcement Return

Type-conditional price components. By standard Gaussian updating, the posterior mean of θ given type τ and disclosure D is

$$m_\tau(D) \equiv \mathbb{E}[\theta \mid \tau, D] = \mu_\tau + \kappa_\tau(D - \mu_\tau), \qquad \kappa_\tau \equiv \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\eta,\tau}^2} \in (0, 1). \tag{15}$$

Since $\sigma_{\eta,S}^2 < \sigma_{\eta,G}^2$, we have $\kappa_S > \kappa_G$: sincere firms' disclosures are more informative about underlying environmental quality.

Conditional independence and information structure. A key maintained assumption is that the implementation signal A is conditionally independent of latent environmental quality θ given managerial type τ:

$$A \perp \theta \mid \tau.$$

This holds by construction: $A = \alpha_\tau^* + \xi$ with ξ independent of θ (Equation (6)), so A is informative about θ only through its role in updating beliefs about managerial type via the cost channel, and does not provide an additional direct signal about θ beyond type. This assumption implies that the type-conditional posterior mean of θ depends only on disclosure D:

$$\mathbb{E}[\theta \mid \tau, D, A] = \mathbb{E}[\theta \mid \tau, D] \equiv m_\tau(D).$$

This restriction is what allows the announcement return to be written as the product of belief revision and the disclosure-dependent quality spread.

Stock prices. Both prices reduce to weighted averages of type-conditional means:

$$P_1(D) \;=\; \pi_1(D)\, m_S(D) + (1-\pi_1(D))\, m_G(D), \tag{16}$$

$$P_2(D,A) \;=\; \pi_2(D,A)\, m_S(D) + (1-\pi_2(D,A))\, m_G(D). \tag{17}$$

Announcement return. Define the *quality spread*

$$\Delta(D) \;\equiv\; m_S(D) - m_G(D) \;=\; \underbrace{\mu_S(1-\kappa_S) - \mu_G(1-\kappa_G)}_{\text{unconditional gap}} \;+\; \underbrace{(\kappa_S - \kappa_G)}_{>0}\, D. \tag{18}$$

The derivative of Δ with respect to D is

$$\Delta'(D) \;\equiv\; \frac{d\Delta}{dD} \;=\; \kappa_S - \kappa_G \;>\; 0, \tag{19}$$

which is positive because $\sigma^2_{\eta,S} < \sigma^2_{\eta,G}$ implies $\kappa_S > \kappa_G$ (Equation (9)). This constant is used throughout the proofs. Note that because $\kappa_S > \kappa_G$ implies $(1-\kappa_S) < (1-\kappa_G)$, the unconditional gap $\mu_S(1-\kappa_S) - \mu_G(1-\kappa_G)$ is not automatically positive from $\mu_S > \mu_G$ alone. We therefore impose the parameter restriction $\mu_S/\mu_G > (1-\kappa_G)/(1-\kappa_S)$, which ensures $\Delta(0) > 0$. Since the D-sensitive term $(\kappa_S - \kappa_G)D > 0$ for $D > 0$, this restriction guarantees $\Delta(D) > 0$ for all $D \geq 0$.

Remark on the support of D. The disclosure $D = \theta + \eta_\tau$ is a sum of Gaussian random variables and therefore has full support on $\mathbb{R}$, so $\Delta(D) < 0$ is possible for sufficiently negative D. The propositions below restrict attention to $D \geq 0$, which is without material loss of generality under a standard calibration condition: ESG disclosure scores are constructed so that $\mu_G \gg \sqrt{\sigma^2_\theta + \sigma^2_{\eta,G}}$—i.e., even the strategic type's expected disclosure is large relative to the total disclosure noise—which implies $P(D < 0) \approx \Phi(-\mu_G/\sqrt{\sigma^2_\theta + \sigma^2_{\eta,G}}) \approx 0$. This is a calibration condition on the theoretical model parameters: μ_G must be sufficiently large relative to total disclosure noise, and it does not follow mechanically from $\mu_G > 0$ alone. In the empirical implementation, D is proxied by text-based ESG scores that are strictly non-negative by construction—satisfying the support restriction $\mathcal{D} \subset \mathbb{R}_+$—but this empirical non-negativity is a separate object from the Gaussian calibration condition on $P(D < 0)$. All comparative statics involving $\Delta(D)$ are stated over the empirically relevant support $\mathcal{D} \subset \mathbb{R}_+$, consistent with the non-negative disclosure measures used in the data. We define $\mathcal{D}$ operationally as the set of D values between the empirical 1st and 99th percentiles of the text-based disclosure proxy $\hat{D}$ in the panel (equivalently, the interquantile range where $\pi_1(D)$ admits a well-defined monotone relationship to D under the model's noise assumptions). Comparative-static claims in the text should be read as applying within this empirical interior rather than at the theoretical Gaussian tails.

The announcement return follows directly from Equations (16) and (17):

$$\begin{aligned} P_2(D,A) - P_1(D) &= \big[\pi_2 m_S + (1-\pi_2) m_G\big] - \big[\pi_1 m_S + (1-\pi_1) m_G\big] \\ &= (\pi_2 - \pi_1)\big(m_S - m_G\big) = (\pi_2 - \pi_1)\Delta(D), \end{aligned}$$

yielding the compact identity

$$\boxed{R(D,A) \;\equiv\; P_2(D,A) - P_1(D) \;=\; \underbrace{\big[\pi_2(D,A) - \pi_1(D)\big]}_{\text{belief revision}} \times \underbrace{\Delta(D)}_{\text{quality spread} > 0}.} \tag{20}$$

Equation (20) is the paper's core pricing identity. The return is the product of the *belief revision*, measuring how much the action updates the investor's type assessment, and the *quality spread*, measuring how much more valuable a sincere firm is than a greenwashing firm conditional on disclosure D. Both belief updating and the disclosure-dependent quality spread $\Delta(D)$ contribute to the interaction effect; in the empirically relevant region, variation in $\Delta(D)$ dominates.

Local nature of the comparative static. The positive interaction between disclosure and action strength holds over the empirically relevant parameter region—the interior of the prior distribution ($\pi_1 \in (0,1)$) at moderate $\ell > 1$. The cross-partial-positive condition $\partial^2 h/\partial p\,\partial \ell > 0$, which requires $\pi_1 < 1/(1+\ell)$ (equivalently $\pi_2 < 1/2$), holds for the majority but not all of the sample (event-sample mean $\hat{\pi} = 0.47$, SD $= 0.20$; sample median 0.46, 75th percentile 0.64); the proposition's positive comparative-static result obtains nonetheless because the quality-spread channel dominates the potentially negative belief-revision channel by Assumption 1. Globally, because the belief-revision function $h(p,\ell)$ is non-monotonic in p, the sign of $\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right)$ is not guaranteed outside this region. Propositions 1 and 2 therefore establish that this derivative is positive over an empirically relevant support $\mathcal{D} \subset \mathbb{R}_+$, rather than as a global property of the model.

As disclosure D rises across firms, the total derivative $\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right)$ is positive because the quality-spread component $\Delta'(D)\cdot \partial h/\partial \ell$ dominates the potentially negative cross-partial component $\Delta(D)\cdot(\partial^2 h/\partial p\,\partial \ell)\cdot \pi_1'(D)$ over the empirically relevant support $\mathcal{D}$ (where $\pi_1'(D) \equiv d\pi_1/dD$ denotes the derivative of the credibility prior with respect to prior disclosure). We obtain a directional prediction that the marginal effect of action strength on returns is increasing in disclosure D, which we take to the data in the event-study regressions below. In particular, this positive total derivative (Proposition 2) provides a microfoundation for the interaction term between action strength and inferred intent in the empirical specification. The formal derivation is in Appendix B.1.

Belief revision in closed form. Let $\ell(A) \equiv g_S(A)/g_G(A)$ denote the action likelihood ratio. Substituting (14) into the belief revision:

$$\pi_2(D,A) - \pi_1(D) \;=\; h(\pi_1,\,\ell(A))\,, \qquad h(p,\ell) \;\equiv\; \frac{p(1-p)(\ell-1)}{p\,\ell + (1-p)}. \tag{21}$$

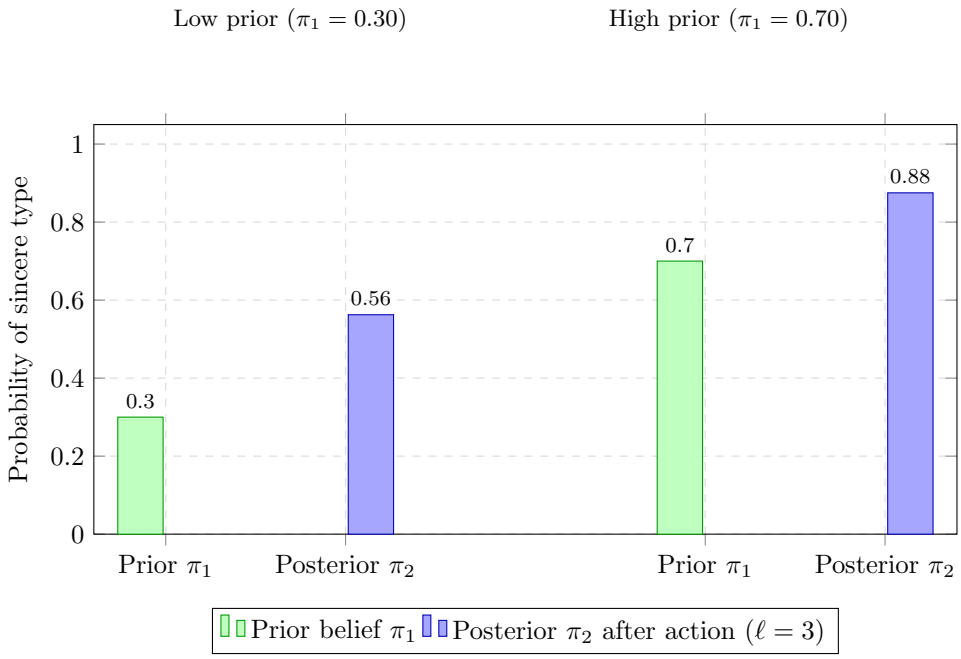
Three properties of h are immediate and sufficient for the propositions that follow:

(P1) $h(p,\ell) > 0$ for all $p \in (0,1)$ and $\ell > 1$, i.e., the belief revision is positive whenever the action is more consistent with sincerity.

(P2) $h(p,\ell)$ is globally non-monotonic (hump-shaped) in p: belief revisions are largest at intermediate priors and smaller when priors are close to 0 or 1. Over the empirically relevant support $p < 1/(1+\sqrt{\ell})$, h is locally *increasing* in p; this is the sub-property that matters for the comparative statics, and is proved formally in Appendix B as eq. (A.4). Analytically, $\partial h/\partial p > 0$ holds whenever $p < 1/(1+\sqrt{\ell})$, equivalently $\pi_2 < \sqrt{\ell}/(1+\sqrt{\ell})$ (established in Appendix B, eq. (A.4)). Note that the P2 threshold $p < 1/(1+\sqrt{\ell})$ is *strictly higher* than the cross-partial threshold $p < 1/(1+\ell)$ used in Proposition 2 (since $\sqrt{\ell} < \ell$ for $\ell > 1$, so $1/(1+\sqrt{\ell}) > 1/(1+\ell)$): the cross-partial $\partial^2 h/\partial p\,\partial \ell > 0$ holds only for $p < 1/(1+\ell)$, equivalently $\pi_2 < 1/2$ (eq. (A.6)).

(P3) $h(p,\ell)$ is strictly increasing in ℓ for all $p \in (0,1)$.

Figure 1: Bayesian belief updating: prior and posterior credibility beliefs under an illustrative action with likelihood ratio $\ell = 3$.



Notes. The figure shows how investors update beliefs about managerial credibility after observing a climate action A with likelihood ratio $\ell(A) = 3$. This parameterization is illustrative: $\ell = 3$ produces large, legible belief revisions but is substantially larger than the typical $\ell \approx 1$ implied by the action distributions in the sample (see Appendix B). The figure is for visual intuition only; the empirical identification relies on the quality-spread channel and does not require any specific ℓ value. The cross-partial condition $\partial^2 h/\partial p\,\partial \ell > 0$ requires $p < 1/(1+\ell)$, which at $\ell = 3$ holds only for $p < 0.25$—well below the sample median. Starting from a low prior ($\pi_1 = 0.30$), the posterior rises to $\pi_2 = 0.5625$ (revision +0.2625); starting from a high prior ($\pi_1 = 0.70$), the posterior rises to $\pi_2 = 0.875$ (revision +0.175). The raw belief revision is larger for the low-prior firm, reflecting the non-monotonic dependence of $h(p, \ell)$ on p. This figure illustrates the belief-updating channel only; the announcement return $R = [\pi_2 - \pi_1] \times \Delta(D)$ need not be larger for the low-prior firm because the quality spread $\Delta(D)$ is increasing in disclosure informativeness, so the high-prior firm's larger $\Delta(D)$ can dominate the smaller belief revision. The return amplification result is established formally in Proposition 2.

Since the log-likelihood ratio $\log \ell(A) = [(\alpha_S - \alpha_G)/\sigma_A^2]\, A + \text{const.}$ is strictly increasing in A (the monotone likelihood ratio property of the Gaussian family with common variance), $\ell(A) > 1$ whenever $A > \bar{A} \equiv (\alpha_S + \alpha_G)/2$, and (P3) implies h is increasing in A for $A > \bar{A}$. Combined with $\Delta(D) > 0$, properties (P1)–(P3) provide the updating mechanics used in the propositions below; the signed return comparative statics additionally require Assumptions 1 and 2 where stated. Figure 1 illustrates the belief-updating mechanics for a representative parameterization.

3.6 Main Results

Assumption 1 (Relevant-support monotonicity and dominance)**.** *Under unequal cross-type disclosure variances, the monotone likelihood ratio property can fail in the tails. We therefore do not impose global monotonicity of $\pi_1(D)$. Instead, we assume there exists an empirically relevant support region $\mathcal{D}$ such that $\pi_1'(D) > 0$ for all $D \in \mathcal{D}$, and that the quality-spread term dominates the belief-revision term over this support:*

$$\Delta'(D)\, h(p, \ell) \;>\; \left| \Delta(D)\, \frac{\partial h}{\partial p}\, \pi_1'(D) \right|$$

for all (p, ℓ) in the empirically relevant support. This condition requires that the direct effect of disclosure on the valuation spread strictly dominates the indirect effect operating through belief updating over $\mathcal{D}$.

Sufficient condition in terms of primitives. *A sufficient condition is that $\sigma^2_{\eta,G}/\sigma^2_{\eta,S}$ is sufficiently large relative to the signal-to-noise ratio $\sigma^2_\theta/(\sigma^2_\theta+\sigma^2_{\eta,S})$. Specifically, when $\kappa_S-\kappa_G \equiv (\sigma^2_{\eta,G}-\sigma^2_{\eta,S})/[(\sigma^2_\theta+\sigma^2_{\eta,S})(\sigma^2_\theta+\sigma^2_{\eta,G})]\cdot\sigma^2_\theta$ is large relative to $|\partial h/\partial p|\cdot|\pi_1'(D)|/h(p,\ell)$, the quality-spread channel dominates. In the empirically calibrated range, where $\hat{\pi}$ is concentrated in $[0.2, 0.7]$ and ℓ is close to unity, the second (belief-revision) term is small relative to the first, so Assumption 1 is satisfied for a wide range of parameter values consistent with the data; Appendix B discusses the relative magnitudes of the quality-spread and belief-revision terms over the empirically relevant region.*

Assumption 2 (Relevant-support dominance for disclosure-dependent amplification)**.** *There exists an empirically relevant support region $\mathcal{D} \subset \mathbb{R}_+$ such that $\pi_1'(D) > 0$ for all $D \in \mathcal{D}$, where $p \equiv \pi_1(D)$, and*

$$\Delta'(D)\,\frac{\partial h}{\partial \ell}(p,\ell) \;>\; \left|\Delta(D)\,\frac{\partial^2 h}{\partial p\,\partial \ell}(p,\ell)\,\pi_1'(D)\right|$$

for all (p, ℓ) induced by (D, A) on that support.

This condition requires that, over the empirically relevant region, the positive quality-spread channel in the marginal return response to action strength strictly dominates the indirect belief-revision channel operating through the cross-partial of h. Because $\partial^2 h/\partial p\,\partial \ell$ changes sign at the stricter threshold $p < 1/(1+\ell)$, this restriction is analytically distinct from Assumption 1, which governs the level derivative dR/dD rather than the cross-partial derivative $d(\partial R/\partial A)/dD$. In the empirically calibrated range where $\hat{\pi} \in [0.2, 0.7]$ and ℓ is close to unity, the cross-partial term is small in magnitude relative to the quality-spread term, so Assumption 2 is satisfied for a wide range of parameter values consistent with the data.

Proposition 1 (Disclosure and Announcement Returns)**.** *Let $R(D, A) = \Delta(D)\,h(\pi_1(D), \ell(A))$ with $\ell(A) > 1$ for $A > \bar{A}$ and $\Delta(D) > 0$. (Here $\bar{A} \equiv (\alpha_S + \alpha_G)/2$ is the threshold above which the action is more consistent with sincerity.) For fixed $A > \bar{A}$,*

$$\frac{dR}{dD} = \Delta'(D)\,h(\pi_1(D),\ell(A)) + \Delta(D)\,\frac{\partial h}{\partial p}(\pi_1(D),\ell(A))\,\pi_1'(D). \tag{22}$$

Under Assumption 1, $\frac{dR}{dD} > 0$ on the empirically relevant support $\mathcal{D}$. Firms whose prior disclosures $D \in \mathcal{D}$ induce a lower credibility prior $\pi_1(D)$ generate a uniformly smaller positive market reaction to any climate action of the same strength.

Proof. Write $R(D, A) = \Delta(D) \cdot h(\pi_1(D), \ell(A))$ and expand the total derivative using the product rule, as in (22).

The first term $\Delta'(D) \cdot h(\pi_1, \ell)$ is strictly positive for all $D \in \mathcal{D}$ and $A > \bar{A}$: $h(\pi_1, \ell) > 0$ by Property (P1) (since $\ell > 1$ when $A > \bar{A}$ and $\pi_1 \in (0, 1)$); and $\Delta'(D) = \kappa_S - \kappa_G > 0$ because $\sigma^2_{\eta,S} < \sigma^2_{\eta,G}$ implies $\kappa_S > \kappa_G > 0$ (Equation (19)).

The second term $\Delta(D) \cdot (\partial h/\partial p) \cdot \pi_1'(D)$ can be negative when $\partial h/\partial p < 0$, i.e., when $\pi_1(D) > 1/(1+\sqrt{\ell})$ (Property (P2)).

Under Assumption 1, the strictly positive quality-spread term $\Delta'(D) \cdot h$ strictly dominates the (potentially negative) second term over the empirically relevant support $\mathcal{D}$, implying $\frac{dR}{dD} > 0$.

Remark on $\pi_1(D)$ monotonicity. When $\sigma^2_{\eta,G} > \sigma^2_{\eta,S}$, the log-likelihood ratio $\Lambda(D) = \log[f_S(D)/f_G(D)]$ is a concave function of D (its second derivative equals $-1/v_S + 1/v_G < 0$, where $v_\tau = \sigma^2_\theta + \sigma^2_{\eta,\tau}$), so $\pi_1(D)$ is

not globally monotone and may decrease for extreme values of D outside $\mathcal{D}$. The result is therefore explicitly local to $\mathcal{D}$ and does not require global monotonicity.

Interpretation. The first term $\Delta'(D) \cdot h$ captures the quality-spread channel: higher disclosure increases the informativeness gap between types, amplifying announcement returns regardless of how π_1 moves. The second term captures the effect of disclosure through belief updating and may attenuate the response when priors are sufficiently high. Assumption 1 ensures the quality-spread channel dominates over the empirically relevant support. □

Interpretation of Proposition 1. The proposition does not require the full sample to lie within the region where $\partial h/\partial p > 0$. As discussed under *Local nature of the comparative static* above, the positive comparative static obtains throughout the empirically relevant support because the quality-spread term dominates the potentially negative belief-revision term by Assumption 1.

Proposition 2 (Local Disclosure-Dependent Amplification of Action Returns)**.** *Under Assumption 2, there exists an empirically relevant support $\mathcal{D} \subset \mathbb{R}_+$ such that the total derivative of the marginal return impact of action strength with respect to disclosure satisfies*

$$\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right) = \ell'(A)\left[\underbrace{\Delta'(D)\,\frac{\partial h}{\partial \ell}}_{>0} + \Delta(D)\,\frac{\partial^2 h}{\partial p\,\partial \ell}\,\pi_1'(D)\right], \tag{23}$$

where $\ell'(A) \equiv d\ell/dA$ denotes the derivative of the likelihood ratio with respect to action strength. The first term is strictly positive for all $D \in \mathcal{D}$ and $A > \bar{A}$. The second term may be negative when $\pi_1(D) \geq 1/(1+\ell(A))$ (equivalently $\pi_2 \geq 1/2$), reflecting the sign change in the cross-partial $\partial^2 h/\partial p\,\partial \ell$. Under Assumption 2, the positive quality-spread component dominates on $\mathcal{D}$, ensuring

$$\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right) > 0.$$

Hence the marginal impact of action strength on announcement returns is increasing in prior disclosure quality on the empirically relevant support.

Remark 1. *Proposition 2 is a local result: the sign of the total derivative is not guaranteed globally due to the non-monotonicity of the belief-revision function $h(p, \ell)$ in p and the potential non-monotonicity of $\pi_1(D)$ outside $\mathcal{D}$ when $\sigma^2_{\eta,G} \neq \sigma^2_{\eta,S}$. The empirical interaction coefficient should therefore be interpreted as evidence for the local comparative static over the empirically relevant distribution of $(\hat{\pi}, A)$, not as a global structural property of the model.*

Proof. Differentiate $R(D, A) = \Delta(D)\, h(\pi_1(D), \ell(A))$ with respect to A:

$$\frac{\partial R}{\partial A} = \Delta(D)\,\frac{\partial h}{\partial \ell}(\pi_1(D), \ell(A))\,\ell'(A).$$

Now differentiate with respect to D. Since $\ell'(A)$ does not depend on D,

$$\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right) = \ell'(A)\,\frac{d}{dD}\left[\Delta(D)\,\frac{\partial h}{\partial \ell}(\pi_1(D), \ell(A))\right].$$

Applying the product rule:

$$\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right) = \ell'(A)\left[\Delta'(D)\,\frac{\partial h}{\partial \ell} + \Delta(D)\,\frac{d}{dD}\left(\frac{\partial h}{\partial \ell}\right)\right].$$

Since D enters h only through $p = \pi_1(D)$,

$$\frac{d}{dD}\left(\frac{\partial h}{\partial \ell}\right) = \frac{\partial^2 h}{\partial p\,\partial \ell}\,\pi_1'(D).$$

Substituting yields (23).

First term. With $Q \equiv p\ell + (1-p)$ (as defined in the appendix proof), $\partial h/\partial \ell = p(1-p)/Q^2 > 0$ (Property (P3)) and $\Delta'(D) = \kappa_S - \kappa_G > 0$ (Equation (19)), so the first term is strictly positive.

Second term. From equation (A.6) (derived in the appendix proof below), $\partial^2 h/\partial p\,\partial \ell = [(1-p) - p\ell]/Q^3$, which is positive when $\pi_2 < 1/2$ and may be negative otherwise. Under Assumption 2, the positive quality-spread component strictly dominates the (potentially negative) second term over $\mathcal{D}$, implying the result. The result is local rather than global: it does not require $\partial^2 h/\partial p\,\partial \ell > 0$ everywhere, only that the total derivative remain positive on the support relevant for the empirical application.

Remark. The empirical interaction coefficient β_3 corresponds to this total derivative rather than the cross-partial of h alone. □

Relation to the empirical interaction. Proposition 2 characterizes the economically relevant total derivative: how the marginal pricing of action strength changes as disclosure D changes. This is the appropriate model analogue of the empirical interaction between *ActionStrength* and $\hat{\pi}$, because π_1 is itself a function of D. A positive regression interaction therefore need not be driven by the local cross-partial $\partial^2 h/\partial p\,\partial \ell$ alone; it may also arise through the strictly positive quality-spread term proportional to $\Delta'(D)\,\partial h/\partial \ell$. The positive interaction need not arise from belief updating alone; it can also reflect the disclosure-dependent quality spread $\Delta(D)$, which mechanically amplifies returns for firms with more credible prior disclosures.

Distinction between Assumptions 1 and 2. Assumption 1 governs the sign of the level derivative dR/dD in Proposition 1, whereas Assumption 2 governs the sign of the derivative of the marginal action effect, $d(\partial R/\partial A)/dD$, in Proposition 2. The two restrictions are analytically distinct because they involve different derivatives of the belief-revision function h: Assumption 1 bounds $\Delta'(D)\,h$ against $|\Delta(D)\,(\partial h/\partial p)\,\pi_1'(D)|$, while Assumption 2 bounds $\Delta'(D)\,(\partial h/\partial \ell)$ against $|\Delta(D)\,(\partial^2 h/\partial p\,\partial \ell)\,\pi_1'(D)|$. Assumption 1 is not sufficient for Proposition 2 because the cross-partial $\partial^2 h/\partial p\,\partial \ell$ can be negative over a wider region of the parameter space than $\partial h/\partial p$ alone. $\partial h/\partial p > 0$ whenever $p < 1/(1+\sqrt{\ell})$, while the cross-partial relevant for return amplification holds only when $p < 1/(1+\ell)$ (equivalently $\pi_2 < 1/2$). They therefore govern different comparative statics and should not be interpreted as alternative characterizations of the same effect.

Remark 2. *The cross-partial of h changes sign at $\pi_1 = 1/(1+\ell(A))$ (equivalently $\pi_2 = 1/2$). Note that $\pi_2 = 1/2$ corresponds to* maximum *posterior variance $\pi_2(1-\pi_2)$, not resolution of uncertainty; the cross-partial turns negative once the action is strong enough to push the posterior above the maximum-uncertainty threshold. As established in the proof, the positive empirical interaction $\beta_3 > 0$ is explained by the total derivative of R with respect to (A, D)—specifically the rising quality spread $\Delta(D)$ for high-disclosure firms—rather than by the cross-partial of h alone. Because $\pi_1(D)$ and $\Delta(D)$ are both functions of D, the empirical proxy $\hat{\pi}$ captures a composite of belief-updating and disclosure-dependent valuation effects.*

Proposition 3 (Dynamic Belief Updating)**.** *A stronger climate action raises the posterior belief π_2 for $A > \bar{A}$:*

$$\frac{\partial \pi_2}{\partial A} > 0. \tag{24}$$

This first result holds globally for all $A > \bar{A}$, $\pi_1 \in (0,1)$, and finite $\ell > 1$; no additional parameter restriction is required. The cross-partial result is local: the marginal effect of the action on the posterior is increasing in the prior belief π_1 over the parameter region relevant for estimation:

$$\frac{\partial^2 \pi_2}{\partial A \, \partial \pi_1} > 0. \tag{25}$$

This cross-partial holds analytically whenever $\pi_1 < 1/(1+\ell)$.

Proof. From $\pi_2 = \pi_1\ell/[\pi_1\ell + (1-\pi_1)]$, we have $\partial\pi_2/\partial A = [\partial\pi_2/\partial\ell] \cdot \ell'(A)$, where $\partial\pi_2/\partial\ell = \pi_1(1-\pi_1)/Q^2 > 0$ and $\ell'(A) > 0$. Hence $\partial\pi_2/\partial A > 0$ for $A > \bar{A}$ and $\pi_1 \in (0,1)$, establishing (24) globally.

For the cross-partial, $\partial^2\pi_2/\partial A\,\partial\pi_1 = [\partial^2\pi_2/\partial\ell\,\partial\pi_1] \cdot \ell'(A)$, where

$$\frac{\partial^2 \pi_2}{\partial \ell \, \partial \pi_1} = \frac{(1-p) - p\ell}{Q^3}$$

is positive if and only if $p < 1/(1+\ell)$ (equivalently $\pi_2 < 1/2$), establishing (25) over this region. □

Over the observed distribution of $(\hat{\pi}, A)$, the positive cross-partial prediction holds locally in the interior region where posterior uncertainty remains material; the sample median $\hat{\pi} = 0.46$ sits near the boundary of the region $\pi_2 < 1/2$, so we interpret the prediction as a local comparative static rather than a global property. Within this region, higher prior credibility implies stronger belief updating.

Remark 3. *Proposition 3 generates two implications for the empirical analysis, which differ in testability. First, climate actions raise next-period credibility scores (H3): this is tested in the belief-updating regressions (Table 17) by regressing $\hat{\pi}_{i,t+1}$ on a binary action indicator. Second, we show that the* price reaction *to a given action is amplified for firms with higher prior credibility beliefs, because the quality spread $\Delta(D)$ scales the announcement return (Equation 20). The empirical manifestation of this amplification is the positive interaction coefficient $\beta_3 > 0$ tested in H1 (Table 15). Testing the underlying nonlinear cross-partial of the belief-revision function $\partial^2 h/\partial p\,\partial\ell$ directly would require observing investor posteriors π_2 at event frequency, which is not available in our data; we therefore treat this as a theoretical implication rather than a separately identified empirical test.*

Remark 4. *Proposition 3 is a property of the posterior updating map $\pi_2(p,\ell)$, treating p and ℓ as arguments of the Bayesian updating rule. By itself, it does not characterize the full disclosure-action interaction in returns, which is governed by the total derivative in Proposition 2.*

Proposition 4 (Disclosure Standardization Weakens the Credibility Channel)**.** *A regulatory shock that reduces cross-type disclosure heterogeneity—formally, $\sigma^2_{\eta,G} \downarrow \bar{\sigma}^2 \geq \sigma^2_{\eta,S}$ —has three consequences:*

(i) Compressed type-discriminability. *As $\sigma^2_{\eta,G} \downarrow \bar{\sigma}^2$, the cross-sectional variance of the log-likelihood ratio $\Lambda(D) = \log[f_S(D)/f_G(D)]$ decreases, attenuating the sensitivity of posterior beliefs $\pi_1(D)$ to variation in disclosure content. In the special case where $\mu_S = \mu_G$ and $\bar{\sigma}^2 = \sigma^2_{\eta,S}$ (complete variance equalization), the likelihood ratio converges to unity uniformly in D and $\pi_1(D) \to \pi_0$; if only the means are equal but residual variance differences remain ($\bar{\sigma}^2 > \sigma^2_{\eta,S}$), the log-likelihood ratio retains a quadratic*

term in D and does not collapse fully. In general the result is a comparative-static attenuation of posterior dispersion rather than a full collapse. Cross-sectional dispersion of π_1 across firms falls.

(ii) Compressed quality spread. *The D-sensitive component of $\Delta(D)$—the term $(\kappa_S - \kappa_G)D$—shrinks (and vanishes under complete variance equalization $\sigma^2_{\eta,G} \downarrow \sigma^2_{\eta,S}$) as $\kappa_G \to \kappa_S$. Under partial equalization ($\bar{\sigma}^2 > \sigma^2_{\eta,S}$) the component compresses but does not fully disappear; in either case the disclosure-dependent amplification weakens.*

(iii) Weakened credibility pricing. *As dispersion in π_1 falls (i) and Δ loses its disclosure-sensitive component (ii), the return sensitivity captured by the interaction $ActionStrength \times \hat{\pi}_{i,e-1}$ weakens in the post-regulation period for treated firms. Formally, the quality-spread channel $\Delta'(D) = \kappa_S - \kappa_G \to 0$ is rigorously established from Part (ii). The additional channel through reduced sensitivity $|\pi_1'(D)|$ is supported by comparative-static reasoning (Part (i) shows compression of noise-driven posterior dispersion, which is consistent with reduced sensitivity) but is not formally established as a monotone reduction without additional conditions. Appendix B provides the full derivation and makes explicit the scope of the comparative-static argument for the $\pi_1'(D)$ channel; a formal monotonicity result for this channel alone remains an open question.*

The channel is not that ESG disclosure becomes uninformative overall. Investors continue to learn about θ from D via the Gaussian update (15). Rather, standardization removes the differential precision *across types that allows investors to infer type from disclosure style—the specific feature that makes π_1 cross-sectionally variable and empirically relevant.*

Proof. Part (i): Compressed type-discriminability. The log-likelihood ratio $\Lambda(D)$ depends on the difference in disclosure variances $\sigma^2_{\eta,G} - \sigma^2_{\eta,S}$. As $\sigma^2_{\eta,G} \downarrow \bar{\sigma}^2$, the dispersion of $\Lambda(D)$ shrinks, reducing the sensitivity of $\pi_1(D)$ to disclosure content. Full details are in Appendix B.

Part (ii): Compressed quality spread. The disclosure-sensitive component of $\Delta(D)$ is $(\kappa_S - \kappa_G)D$. As $\sigma^2_{\eta,G} \downarrow \sigma^2_{\eta,S}$, we have $\kappa_G \uparrow \kappa_S$, so $\kappa_S - \kappa_G \to 0$, weakening the quality-spread channel. This is rigorously established from the formula for κ_τ in Equation (15).

Part (iii): Weakened credibility pricing. The amplification term in Proposition 2 depends on $\Delta'(D)$ and $\pi_1'(D)$, both of which are attenuated under disclosure standardization via Parts (i) and (ii). Consequently, the interaction $ActionStrength \times \hat{\pi}_{i,e-1}$ weakens for treated firms. The $\Delta'(D) = \kappa_S - \kappa_G \to 0$ channel is formally established; the $\pi_1'(D)$ attenuation channel follows from comparative-static reasoning in Part (i) rather than a formal monotonicity result. □

Testable hypotheses. We obtain six signed, testable hypotheses (H1, H2, H3, H4, H5a, H5b) that map directly onto the estimating equations in Section 8. We additionally derive that the price reaction to a given action is amplified for high-credibility firms through the quality-spread channel; this implication is embedded in H1 and discussed in the remark following Proposition 3.

H1 (Credibility premium). The interaction coefficient on $ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$ is positive: identical climate actions generate larger market reactions at firms with higher prior credibility (Propositions 1 and 2).

H2 (Follow-through amplification). The interaction between follow-through actions and prior credibility is positive: credible implementation of earlier commitments is valued more highly at firms with high prior credibility (Proposition 2).

H3 (Follow-on disclosure response). Climate actions are associated with higher next-period credibility scores in the firm's own subsequent disclosures, conditional on firm fixed effects. This tests firm communication tone, not investor beliefs directly: it asks whether actions are followed by changes in firm communication style consistent with the model's dynamic mechanism, not whether investor beliefs themselves are updated (which is unobserved). It is mapped to Proposition 3 as an indirect, observable implication of the dynamic process, with the explicit caveat that it cannot separately identify investor belief revision from endogenous changes in firm communication; the test uses any climate action (binary) as the regressor, acknowledging that it does not isolate costly or verifiable actions specifically. This is tested via firm-year panel regressions (Table 17).

H4 (Regulatory attenuation). The four-way interaction capturing disclosure standardization is predicted to be negative: the credibility-pricing channel is predicted to weaken when cross-type heterogeneity in disclosure style is compressed (Proposition 4).

H5a (Salience amplification, baseline). The triple interaction $ActionStrength \times \hat{\pi}_{i,e-1} \times Attention_t$ is positive: the credibility-pricing channel strengthens during periods of elevated investor attention to climate risk, reflecting greater cognitive weight placed on intent inference when the domain is salient (Section OA-2.2). The *primary* test of H5a is Table 24 (coefficient 0.54–0.59), which isolates the standalone salience amplification. Table 27 Column (3) also reports the analogous triple interaction (0.51) but conditions jointly on *HighCost*; that specification tests H5a and H5b simultaneously and should be read as a joint test rather than an alternative primary estimate of H5a.

H5b (Salience-by-diagnosticity, BToM-specific). The four-way interaction $ActionStrength \times \hat{\pi}_{i,e-1} \times Attention_t \times HighCost_{i,e}$ is positive: salience amplification is larger for costly-to-mimic actions, because heightened attention induces more effortful inverse planning that disproportionately weights actions that strongly reveal the manager's latent cost structure. This prediction is specific to the BToM framework and goes beyond what the baseline salience amplification in H5a requires (BToM Corollary 3; Table 27, Column (3)).

Corollary 1 (Intention Inference Amplifies Cross-Sectional Return Dispersion)**.** *For any fixed action strength $A > \bar{A}$, the cross-sectional variance of announcement returns $\mathrm{Var}_i[R(D_i, A)]$ is increasing in the dispersion of credibility priors $\mathrm{Var}_i[\pi_1(D_i)]$ under a first-order (local) approximation. Disclosure standardization that compresses $\mathrm{Var}(\pi_1)$ therefore reduces return dispersion even if the distribution of environmental quality θ is unchanged.*

Proof. From $R = \Delta(D) \cdot h(\pi_1(D), \ell)$ and Property (P2) ($\partial h/\partial \pi_1 > 0$ whenever $\pi_1 < 1/(1+\sqrt{\ell})$, see Equation (A.4)), the announcement return R is strictly increasing in the prior credibility π_1 for any fixed action strength $A > \bar{A}$ in the empirically relevant region. Because $\pi_1 = \pi_1(D)$ is itself a function of disclosure D, cross-sectional variation in D induces variation in π_1. The cross-sectional variance of returns is

$$\mathrm{Var}_i[R(D_i, A)] = \mathrm{Var}_i\big[\Delta(D_i) \cdot h(\pi_1(D_i), \ell)\big].$$

Under a first-order (delta-method) approximation around sample means,

$$\mathrm{Var}(R) \approx \left(\frac{dR}{d\pi_1}\right)^2 \mathrm{Var}(\pi_1),$$

where $dR/d\pi_1$ is the total derivative treating D as a function of π_1, obtained via the chain rule:

$$\frac{dR}{d\pi_1} = \Delta(D)\,\frac{\partial h}{\partial \pi_1} + h(\pi_1, \ell)\,\frac{\partial \Delta}{\partial D}\cdot\frac{\partial D}{\partial \pi_1}.$$

The first term is strictly positive by Property (P2) over the region $\pi_1 < 1/(1+\sqrt{\ell})$. For the second term, Equation (19) implies $\partial\Delta/\partial D = \kappa_S - \kappa_G > 0$, and Assumption 1 implies $\partial\pi_1/\partial D > 0$ on $\mathcal{D}$, so that $\pi_1(D)$ is strictly increasing on $\mathcal{D}$ and, by the implicit function theorem, locally invertible with $\partial D/\partial\pi_1 = 1/(\partial\pi_1/\partial D) > 0$ on that region. Hence the second term is also non-negative. It follows that $dR/d\pi_1 > 0$, and therefore $\mathrm{Var}(R)$ is increasing in $\mathrm{Var}(\pi_1)$ under the first-order approximation. This corollary is best interpreted as a reduced-form local implication of the first-order approximation rather than as a global structural result; higher-order terms may affect the magnitude but the directional prediction is robust. Specifically, the monotonicity of R in π_1 is established locally over the region $\pi_1 < 1/(1+\sqrt{\ell})$, where Property (P2) ensures $\partial h/\partial\pi_1 > 0$ and both terms of the chain-rule decomposition are non-negative. Outside this region the sign of $dR/d\pi_1$ depends on the relative magnitudes of the two components and is not guaranteed to be positive; accordingly, the variance result should be interpreted as holding over the empirically relevant support where P2 is satisfied, not as a global property of the model. Empirical tests should therefore be interpreted as local validation of the model's comparative statics over the observed distribution of $(\hat{\pi}_{i,t}, A_{i,e})$, not as tests of global monotonicity. □

Remark 5. *Corollary 1 predicts that cross-sectional* return dispersion *is larger in subsamples where the dispersion of π_1 is wider—consistent with the heterogeneity patterns in Table 22 (dirty industries, low coverage, high uncertainty). Note that Corollary 1 is a variance prediction: it implies that higher dispersion in π_1 directly reduces measurement-error attenuation of $\hat{\beta}_3$ (because the signal-to-noise ratio of the proxy improves when the true signal has greater variance), rather than raising the structural slope itself. The higher cross-sectional variance of returns in those subsamples is a* consequence *of the wider dispersion in π_1, not the mechanism through which attenuation is reduced. Any true structural amplification of β_3 in those subsamples would require a separate mechanism, such as greater information frictions compounding the pricing of narrative intent. Accordingly, we do not interpret differences in coefficient magnitude across subsamples as pure structural differences in β_3. The additional BToM-specific tests in Table 27 (cost asymmetry, consistency, salience-by-diagnosticity) rely on the distinct inverse-planning mechanisms of BToM Corollaries 1–3 rather than dispersion of π_1 alone.*

Corollary 2 (Credibility Prior as a Pricing State Variable)**.** *The investor's credibility prior π_1 enters the return-generating process as a time-varying state variable. Shocks to π_1—through new disclosures, regulation, or revelation events—generate predictable changes in the return sensitivity to future climate actions, independent of changes in cash-flow news about θ.*

Proof. From the mean-centered first-order-plus-cross-partial approximation of $R(D, A)$ around average values $(\bar{A}, \bar{\pi}_1)$ in Equation (OA-1) (introduced formally in Section OA-1.3), the marginal effect of action strength on returns is $\beta_1 + \beta_3(\pi_1 - \bar{\pi}_1)$. Any shock to π_1 therefore shifts this slope, holding fixed the distribution of θ and hence the cash-flow content of *ActionStrength*. In the empirical implementation, $\hat{\pi}_{i,t}$ is scaled to $[0, 1]$ (see Table 13); the interaction coefficient β_3 correctly identifies the cross-partial derivative regardless of centering, and β_1 represents the marginal effect of *ActionStrength* evaluated at $\hat{\pi}_{i,t} = 0$ (not the unconditional mean effect, since the variable is uncentered). □

Remark 6. *Corollary 2 formalizes the language used throughout the paper: beliefs about managerial credibility are a distinct pricing input, not a proxy for current observable environmental quality. It also motivates the quasi-experimental design in Section 8: regulation that compresses cross-type heterogeneity in disclosure (i.e., reduces $\sigma^2_{\eta,G}$) structurally changes the return function itself, lowering the estimated interaction coefficient β_3 independently of any shift in the current level of π_1. This attenuation is a direct prediction of Proposition 4 and is distinct from a mere movement along the existing bilinear surface in Equation* (OA-1) *(defined above).*

The BToM-specific testable predictions that go beyond these two corollaries—cost-asymmetry amplification, consistency amplification, and salience-by-diagnosticity—are developed formally in Section 3.3 and tested in Section OA-2.3.

Empirical mapping and measurement error. The model maps to the regression specification of Section 6 (Market Reactions) via *ActionStrength* (action intensity), $\hat{\pi}$ (the credibility measure as a noisy proxy for prior precision), and their interaction. Classical measurement error in $\hat{\pi}$ implies an attenuation bias toward zero in the prior-by-action interaction β_3, so the estimated coefficient is a lower bound on the true effect. Detailed mapping, the Taylor approximation around the prior mean, and the formal treatment of attenuation are in Online Appendix OA-1 (Section OA-1.3, Equation OA-1, Table OA-1, and Section OA-4.1).

4 Data and Sample Construction

This section describes the data sources, sample construction, and variables used in the empirical analysis. We combine firm-level financial data, climate-related corporate actions, climate performance measures, and textual ESG disclosures into a firm-event dataset.

4.1 Sample Construction

Our sample consists of publicly traded U.S. firms from 2005 to 2023. We begin with firms in the CRSP–Compustat merged database, require non-missing stock returns, accounting variables, and identifiers, and match observations across sources using GVKEY, supplemented by CUSIP and ticker where necessary.

A firm enters the final sample if it has at least one identifiable climate-related corporate action and sufficient ESG communication in the 12 months before the event to construct the inferred credibility measure in Section 5. This requirement conditions on disclosure availability rather than disclosure tone; IPW tests in Section 9 address selection on observables within this support. Event-study estimation additionally requires at least 120 non-missing return observations in the $[-250, -11]$ estimation window.

The final dataset contains 4,812 climate-related corporate actions across 1,856 firms and 19,804 firm-years of ESG communication. The reduction from the 5,200 firms in the full NLP corpus reflects the requirement that each event-sample firm have at least one identifiable climate action.

Follow-through implementation events. We also construct a supplementary sample of 10,808 follow-through implementation events, defined as disclosures that reference prior climate commitments and report measurable progress through emissions reductions, capital allocation, or operational changes tied to earlier targets. We identify candidates using keyword and phrase searches across filings, sustainability reports, press releases, and earnings calls, then manually verify that each candidate reflects concrete implementation rather

than a new or independent commitment. A single initial commitment may generate multiple follow-through updates over time.[2]

Panel B of Table 13 reports summary statistics for the follow-through sample. Relative to primary announcements, follow-through events have lower average returns and dispersion, consistent with incremental implementation rather than new information, and primarily cover emissions reductions, green capital expenditures, and progress on climate targets. They are not highly concentrated: the median firm contributes 2 follow-through events, the mean is 3.4, and the 90th percentile is 8. Primary and follow-through events are mutually exclusive, so each firm-event date appears at most once in the pooled dataset used for Table 16. These events are used to test follow-through amplification (H2) in Section 6.

Support condition. The inclusion requirements define the support over which the empirical design identifies effects: firms with observable climate actions and sufficient prior ESG disclosure. Results should therefore be interpreted as applying to publicly traded U.S. firms actively engaged in climate signaling; generalization to firms with no identifiable climate actions requires assuming the same pricing mechanism operates in lower-activity settings. IPW robustness tests address selection on observables within this support but not selection on latent communication propensity or investor attention.

The full text panel contains 19,804 firm-years. Table 14 uses 18,912 with complete data on emissions reductions, green capital expenditures, and verifiable target adoption; Table 23 uses 18,200 after additionally requiring emissions intensity and green patent data; Table 17 uses 17,940 because the $t+1$ lead credibility measure drops each firm's final observation and 8 additional observations with panel gaps. The primary event-study sample also requires prior climate disclosure within 12 months before the event, producing approximately 4% attrition relative to the raw matched sample. All samples use the same 2005–2023 window and fixed-effects structure unless otherwise noted.

4.2 Climate-Related Corporate Actions

We construct a dataset of firm-level climate-related events, including emissions targets, green capital expenditure programs, renewable energy and clean-technology investments, divestitures or closures of carbon-intensive assets, science-based climate targets, and sustainability-linked financing instruments.

Event dates come from corporate press releases, earnings call transcripts, and SEC filings. We assign each event to the first public disclosure date and exclude events that coincide within the event window with major confounding announcements, including earnings releases, dividend declarations, and merger or acquisition disclosures.

For each event, trained research assistants record the announcement date, action type, and action-strength score. The score $ActionStrength_{i,e}$ is assigned using a pre-registered blinded rubric based on specificity, quantifiability, and scope. It is coded from the focal announcement, while $\hat{\pi}_{i,e-1}$ is constructed only from pre-event disclosures; this temporal separation prevents mechanical overlap. Action-type indicators absorb mean return differences across categories, so the interaction is identified from within-type variation in action strength and prior credibility.

[2] We scan disclosures in the 12–60 months after each primary event for referential language (e.g., "as we committed," "building on," "progress toward," or "in line with our [year] target") combined with climate terms such as "emissions," "net-zero," and "decarbonization." The search yields approximately 14,200 candidate passages. Two trained research assistants apply a blinded rubric requiring explicit linkage to a prior commitment and evidence of concrete implementation. Inter-rater agreement is 87% (Cohen's $\kappa = 0.79$), with disagreements resolved by adjudication. Approximately 76.1% of candidates are retained, yielding 10,808 events. Additional details are provided in the Online Data Appendix.

Abnormal return construction. Market reactions are measured using cumulative abnormal returns (CARs) from a standard market model (Brown and Warner, 1985):

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t}, \tag{26}$$

estimated over trading days $[-250, -11]$ using CRSP value-weighted returns as $R_{m,t}$ and requiring at least 120 non-missing return observations.[3] Abnormal returns are $AR_{i,t} = R_{i,t} - \hat{\alpha}_i - \hat{\beta}_i R_{m,t}$ and are cumulated over the event window. The baseline window is $[-1, +1]$; the interaction coefficient remains positive and significant under $[-2, +2]$ and $[-3, +3]$ windows (Table OA-5). Standard errors for CARs follow Brown and Warner (1985).

4.3 Financial Data

Financial and accounting data come from CRSP and Compustat. CRSP daily returns are used to compute abnormal returns; all return variables are expressed in percentage points. Compustat provides total assets, book equity, leverage, profitability, and capital expenditures.

We construct standard controls, including log total assets, log market capitalization, book-to-market, return volatility, and momentum measured over months $t - 12$ to $t - 2$. Event-study and panel regressions use log total assets; cross-sectional asset-pricing tests use log market capitalization. Control definitions follow Table 13. Accounting and return-based controls are winsorized at the 1st and 99th percentiles, while CARs are not winsorized.

4.4 Environmental Performance Data

Firm-level greenhouse gas emissions and emissions intensity come from Trucost and CDP. These data track realized climate outcomes after climate commitments and serve as external validation outcomes for the credibility measure in Section 5. We supplement them with MSCI ESG Ratings and green innovation measures from USPTO environmental patent filings. These variables enter as controls for underlying environmental exposure and as benchmarks in the validation tests in Section 5.5.

4.5 ESG Communication Data

Corporate ESG communication is collected from sustainability and CSR reports, climate-related 10-K disclosures, earnings call transcripts, and press releases. Text is processed using standard NLP procedures and filtered to climate-related content using a dictionary based on Sautner et al. (2023), augmented with terms for net-zero and science-based targets.

Only pre-event disclosures are used to construct $\hat{\pi}_{i,e-1}$, and text associated with the focal announcement is excluded. The same pre-event texts are also used to construct $Consistency_{i,e-1}$, as described in Section 5.

4.6 Construction of BToM-Specific Variables

To test BToM Corollaries 1–3, we construct two additional variables.

[3]The estimation window partially overlaps with the 12-month disclosure window used to construct $\hat{\pi}_{i,e-1}$. Because prior ESG information may already be incorporated into prices, CARs capture the incremental response to the climate action relative to the pre-event benchmark. This overlap is standard in event-study designs and does not affect cross-sectional identification of the interaction because the estimation window is common across events.

High-cost action indicator ($HighCost_{i,e}$). $HighCost_{i,e} = 1$ for actions involving large verifiable capital commitments, science-based targets, operational restructuring, or measurable emissions reductions with explicit timelines. Purely aspirational actions, such as vague net-zero pledges without operational detail, receive zero. The indicator is coded using the same blinded protocol as $ActionStrength_{i,e}$ and proxies for implementation costs across actions: high-cost actions are more costly for greenwashers to mimic and therefore more diagnostic of sincerity (BToM Corollary 1).

Because *HighCost* and *ActionStrength* both capture commitment scope and verifiability, they are moderately correlated ($r = 0.507$; Online Appendix Table OA-6). The residualized specification in Online Appendix Table OA-8 confirms that the cost-asymmetry result remains after removing this overlap.

Communication consistency ($Consistency_{i,e-1}$). $Consistency_{i,e-1}$ is the average pairwise cosine similarity of Sentence-BERT embeddings (Reimers and Gurevych, 2019) for climate-related paragraphs across sustainability reports, 10-K climate sections, earnings calls, and press releases in the 12 months before the event. If a firm does not issue all four document types, we average over available document pairs; firms with only one document type are assigned missing consistency and excluded from consistency-amplification tests.

Higher values indicate greater cross-venue consistency, which BToM Corollary 2 predicts should amplify the credibility-pricing interaction by lowering effective disclosure noise. $Consistency_{i,e-1}$ is predetermined with respect to both the action and returns, while $HighCost_{i,e}$ is based only on event-day information observable to investors. Neither variable uses post-announcement information.

With the data assembled, we next construct the central empirical object of the paper: the firm-year credibility proxy $\hat{\pi}_{i,t}$ derived from the linguistic content of climate-related disclosures.

5 Measuring the Ex-Ante Credibility Signal from ESG Communication

The model's central empirical object is the investor's posterior belief $\pi_1(D_{i,t})$ that firm i's manager is sincerely committed to climate improvement. Because this belief is not directly observable, we proxy for it using the linguistic structure of corporate sustainability disclosures. The construction exploits the model's implication that implementation-oriented managers issue more specific, verifiable, and forward-looking disclosures, while managers engaged in strategic signaling rely more on vague, aspirational language. We translate this distinction into a firm-year measure $\hat{\pi}_{i,t}$ in three steps: corpus assembly, supervised classification, and firm-level aggregation. The reduced-form interpretation of $\hat{\pi}_{i,t}$ as a noisy proxy for investor credibility beliefs (which may also incorporate correlated implementation capacity) is established in Section OA-1.3; the residualized measure $\hat{\pi}^{res}$ (defined below) is the corresponding robustness check.

5.1 Corpus and Filtering

The corpus draws on four document types from 2005 to 2023: (i) sustainability and CSR reports (company websites, GRI database); (ii) climate-related sections of 10-K filings (SEC EDGAR); (iii) earnings call transcripts (S&P Global); and (iv) corporate press releases (Factiva, IR pages). The corpus contains approximately 280,000 documents across 5,200 U.S. public firms. We retain paragraphs containing at least two tokens from a curated climate dictionary covering emissions, decarbonization, transition risk, renewable

energy, and sustainability governance, removing boilerplate and incidental mentions. The filtered corpus contains roughly 3.8 million paragraph-level observations.

All documents are time-stamped at the public release date. For each firm-year, $\hat{\pi}_{i,t}$ uses only documents released during the 12 months ending in fiscal year t, ruling out look-ahead bias in the firm-level measure.

5.2 The Sincere–Strategic Communication Distinction

The signaling-theoretic premise is that communication is more credible when it is costly to mimic or creates future accountability. Implementation-oriented language commits the firm to measurable targets, operational detail, and verifiable timelines—features harder for a greenwasher to sustain. We operationalize this through three linguistic signatures used by the human raters constructing the training labels: **specificity and verifiability** (quantified targets, implementation timelines, operational detail; vague aspirational language is a negative indicator); **cost acknowledgment** (capital-allocation discussion, restructuring costs, explicit cost-benefit framing); and **forward-looking orientation** (share of climate sentences referring to future actions and targets rather than past achievements; we follow Bozanic et al. 2018). The classifier is fine-tuned on raw paragraph text to predict these labels rather than receiving the engineered features directly, following Tetlock (2007) and Tetlock et al. (2008).

A fourth feature—**cross-venue consistency**—is computed at the firm-year level (Sentence-BERT cosine similarity across document types; Reimers and Gurevych 2019) and used in BToM-specific robustness tests, not in paragraph-level annotation.

5.3 Supervised Classification

Training data. We sample 3,614 paragraphs stratified by document type, firm size quartile, and year. Four trained research assistants independently label each paragraph as *implementation-oriented* or *strategic/image-oriented* using a coding protocol operationalizing the three signatures above. We exclude paragraphs with inter-rater agreement below 75% (approximately 500 paragraphs) and reserve a further 500 as a held-out test set, leaving 2,614 paragraphs for training. Inter-rater agreement on the retained training sample is 83% ($\kappa = 0.71$); on the full 3,614-paragraph sample, $\kappa = 0.61$. The two classes are approximately balanced (52% implementation-oriented).

Classifier. We fine-tune a FinBERT model to produce, for each paragraph d,

$$p_d \equiv P(\text{implementation-oriented} \mid \text{text}_d). \tag{27}$$

On the held-out test set (filtered to high-agreement paragraphs), the classifier achieves F1 = 0.81 and AUC = 0.87. These metrics characterize accuracy on unambiguous paragraphs; performance on borderline cases is lower. Main results are robust to alternative classification thresholds and to a dictionary-based classifier (Table 25, Column (2)).

5.4 Firm-Level Aggregation

Event-study measure. For each climate action event e,

$$\hat{\pi}_{i,e-1} = \frac{1}{|\mathcal{P}^{12}_{i,e-1}|} \sum_{d \in \mathcal{P}^{12}_{i,e-1}} p_d, \tag{28}$$

where $\mathcal{P}^{12}_{i,e-1}$ is the set of climate-related paragraphs in documents released by firm i in the 12 months preceding e. Results are robust to excluding events preceded by any climate-related announcement in the prior 6 months (Online Appendix Table OA-7).

Firm-year panel measure. For panel regressions, we construct $\hat{\pi}_{i,t}$ analogously over the 12 months ending in fiscal year t:

$$\hat{\pi}_{i,t} \;=\; \frac{1}{|\mathcal{P}_{i,t}|} \sum_{d \,\in\, \mathcal{P}_{i,t}} p_d. \tag{29}$$

Both measures lie in $[0, 1]$, with higher values indicating a larger share of implementation-oriented communication. The simple average implicitly weights paragraphs equally regardless of document type or audience reach—a deliberate simplification prioritizing replicability over a structural model of investor attention.

Residualized measure. $\hat{\pi}_{i,t}$ has a moderate correlation with MSCI ESG scores (≈ 0.31), confirming partial overlap but substantial orthogonal variation. To isolate the component orthogonal to observable quality, we construct

$$\hat{\pi}^{res}_{i,e-1} \;=\; \hat{\pi}_{i,e-1} - \widehat{f}_{i,e-1}, \tag{30}$$

where $\widehat{f}_{i,e-1}$ is the fitted value from regressing $\hat{\pi}_{i,e-1}$ on MSCI ESG scores, emissions intensity, green capital expenditures, log total assets, profitability, leverage, and firm and year fixed effects.

5.5 Validation

We establish validity along three dimensions.

Prediction of real climate outcomes. Table 14 reports panel regressions of next-year emissions reductions, green capital expenditure, and adoption of verifiable climate targets on $\hat{\pi}_{i,t}$, controlling for lagged emissions intensity, firm size, leverage, book-to-market, and firm and year fixed effects. The coefficients on $\hat{\pi}_{i,t}$ are positive and significant at the 1% level: 0.082 (SE 0.019), 0.114 (SE 0.024), and 0.091 (SE 0.021), respectively. A one-SD increase in $\hat{\pi}_{i,t}$ (SD = 0.21) is associated with a 1.7 pp higher probability of subsequent emissions reduction and a 2.4 pp higher probability of verifiable green investment.

These outcome-prediction results are consistent with both a belief channel and a firm-quality channel: a proxy that predicts emissions reductions could capture managerial intent or persistent execution capacity. We do not separate these here; the residualized analysis below addresses the distinction.

Incremental predictive power for returns. Section OA-2.4 reports a horse-race specification (Equation (OA-10)) in which $\hat{\pi}_{i,e-1}$ enters jointly with action-strength interactions on contemporaneous emissions intensity, MSCI ESG, and green capex. The coefficient on $\mathit{ActionStrength} \times \hat{\pi}_{i,e-1}$ remains stable across specifications (1.21–1.28; coefficients on quality interactions are small and insignificant), and the interaction with the residualized measure $\hat{\pi}^{res}$ remains 1.23*** (Table OA-2, Column (5)). This indicates that markets price a credibility component orthogonal to observable execution quality.

The tests do not claim to isolate pure intent; they test whether the market prices a credibility-related communication component beyond observable quality.

Distinctness from existing text measures. Regressing $\hat{\pi}_{i,t}$ on Loughran–McDonald positive and negative word counts, general sentiment, and ESG-tone proxies yields R^2 between 0.09 and 0.14. Including these

as additional controls in the main event-study regressions does not attenuate the interaction, confirming that $\hat{\pi}$ captures a dimension of communication distinct from generic ESG tone.

Stability across document sources. The measure is computed separately for each document type; the average within-firm-year correlation across document pairs is 0.62, indicating that sincere or strategic communication is a property of the firm rather than of a particular document format.

5.6 Measurement Error and Attenuation

$\hat{\pi}_{i,t}$ is a noisy proxy for $\pi_1(D_{i,t})$ because the classifier makes paragraph-level errors and because even a perfectly classified disclosure record captures linguistic output rather than the investor belief it induces. As a heuristic benchmark, under classical measurement error,

$$\text{plim } \hat{\beta}_3 \;\approx\; \rho\,\beta_3, \qquad \rho \in (0,1), \tag{31}$$

where ρ is the partial-variance reliability ratio after projecting out main effects (full derivation in Appendix OA-4.1). Because both $\pi_{i,e-1}$ and $\hat{\pi}_{i,e-1}$ are bounded in $[0,1]$, the error is non-classical and the bias direction is not guaranteed to be purely attenuating. Nonetheless, two features support attenuation as a reasonable first-order approximation: the distribution of $\hat{\pi}_{i,e-1}$ is concentrated away from the boundaries, and classification noise primarily reduces signal precision, which dampens estimated interactions. We therefore interpret estimates as a noisy reduced-form approximation to true mechanism strength, not a strict lower bound. Heterogeneity coefficients in Table 22 are largest in subsamples with greatest cross-sectional dispersion in priors—precisely the settings where ρ is largest—so reported subsample differences partly reflect differential attenuation rather than purely economic heterogeneity.

Distinguishing intent from execution capability. A horse-race specification confirms that the credibility measure $\hat{\pi}$ captures *intent*, not execution capability or firm fundamentals: residualizing $\hat{\pi}$ on observable quality proxies leaves the credibility-pricing interaction essentially unchanged. Within-firm specifications and a placebo using future realized outcomes further support an intent-based interpretation. Full development is in Online Appendix OA-2 (Section OA-2.4, Tables OA-2 and OA-3).

6 Empirical Results

This section presents the event-study results testing **H1**–**H3**, together with a within-section cross-venue credibility-shock test. Cross-sectional asset-pricing tests are reported in Section 7; the main identification designs appear in Section 8.

The key event-study regressors $\hat{\pi}_{i,e-1}$ and $\mathit{ActionStrength}_{i,e}$ are scaled to lie in $[0,1]$; $CAR_{i,e}$ is expressed in percentage points. Standard firm-level controls (log assets, book-to-market, leverage, momentum) enter in their raw units; Fama-MacBeth regressors in Section 7 are standardized to zero mean and unit standard deviation. Coefficients on interaction terms are interpreted as the effect of a one-unit change in the underlying variable (the full $[0,1]$ range); economic magnitudes are also discussed using standard-deviation changes. We read the estimated interaction $\hat{\beta}_3$ as a reduced-form projection using noisy proxies for both belief ($\hat{\pi}_{i,e-1}$ for $\pi_1(D)$) and action (*ActionStrength* for A), not a direct estimate of the structural cross-partial $\partial^2 R/\partial A\,\partial\pi_1$.

Standard errors in event-study regressions are double-clustered at the firm and year levels (Petersen, 2009); panel regressions outside the event window cluster at the firm level only. Portfolio partition definitions are introduced where they are used in Section 7.

6.1 Summary Statistics

Table 13 reports summary statistics for the main variables used in the analysis. The table includes measures of climate actions, the intent signal, firm characteristics, and climate outcomes.

6.2 Market Reactions to Climate Actions

We first document the baseline interaction between action strength and credibility priors, then examine evidence that distinguishes the BToM mechanism from reduced-form alternatives. Table 15 presents the main event-study results. The dependent variable is the cumulative abnormal return over the $[-1, +1]$ window around each climate-related corporate action. Consistent with **H1**, stock market reactions are larger for firms whose prior ESG communication is more implementation-oriented. The interaction coefficient $\hat{\beta}_3$ on $ActionStrength \times \hat{\pi}_{i,e-1}$ is 1.17 (SE 0.30, significant at the 1% level) in the preferred specification with controls (Column (3)), and 1.28 (SE 0.34) without controls (Column (1)). Column (1) reports the raw specification; Column (3) presents our preferred baseline with firm controls, industry-by-year fixed effects, and controls for overall climate performance and action type. Coefficient stability across columns (the estimate ranges from 1.13 to 1.28) suggests that omitted firm characteristics are unlikely to be the primary driver. Following Oster (2019), we implement the bounding approach that exploits joint movement of the coefficient of interest and the R^2 as observable controls are added. From Column (1) of Table 15 (no controls, no fixed effects), $\tilde{\beta} = 1.28$ and $\tilde{R}^2 = 0.08$; from Column (3) (preferred specification with industry×year fixed effects and controls), $\hat{\beta} = 1.17$ and within $R^2 = 0.19$. Setting $R^2_{\max} = 1.3 \times R^2 \approx 0.247$ (the Oster (2019) benchmark), the implied selection parameter is

$$\hat{\delta}^{\text{Oster}} \approx \frac{\hat{\beta}(R^2 - \tilde{R}^2)}{(\tilde{\beta} - \hat{\beta})(R^2_{\max} - R^2)} = \frac{1.17 \times (0.19 - 0.08)}{(1.28 - 1.17) \times (0.247 - 0.19)} \approx 20.5.$$

A value of $\hat{\delta}^{\text{Oster}} = 1$ corresponds to selection on unobservables equal in magnitude to selection on observables. Our estimate of $\hat{\delta}^{\text{Oster}} \approx 20.5$ implies that unobserved confounding would need to be more than twenty times as strongly selected as the included controls to fully attenuate the estimated effect. Because coefficient movement is small relative to the increase in R^2, and because Column (1)'s total R^2 is not strictly comparable to Column (3)'s within R^2, we treat this calculation as a heuristic complement to the coefficient-stability evidence rather than as a structural estimate of selection on unobservables. The main effect of $\hat{\pi}_{i,e-1}$ is positive across all columns but loses statistical significance in the preferred specification (0.29, SE 0.19, Column (3) vs. 0.41*, SE 0.22, Column (1)). This attenuation is consistent with industry-by-year fixed effects absorbing much of the cross-sectional variation in $\hat{\pi}_{i,e-1}$ across industries and time periods. The interaction term survives because it is identified from the combination of cross-industry-year variation in credibility priors (which remains after the main effect is absorbed) and within-industry-year variation in *ActionStrength*; the industry-by-year fixed effects do not include *ActionStrength* × industry-year interactions, so this cross-variation remains available for identification. The interaction coefficient, which requires both a credible action *and* a high credibility prior, is more tightly identified from within-type, within-period comparisons and is therefore less sensitive to this absorption. Residual concerns about observable and

unobservable confounding are addressed in Sections OA-2.4 (residualized $\hat{\pi}$) and 8 (within-firm and quasi-experimental designs); the interaction survives both. Additional robustness is provided by Column (4) of Table 25, which replaces industry-by-year fixed effects with firm fixed effects and restricts identification to within-firm variation over time; the interaction coefficient remains similar in magnitude (1.08, SE 0.36).

Placebo test: future communication. Column (4) of Table 15 replaces the pre-event interaction $ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$ with its future counterpart, $ActionStrength_{i,e} \times \hat{\pi}_{i,e+1}$, where $\hat{\pi}_{i,e+1}$ is constructed from disclosures in the 12 months *following* the event. The estimated coefficient is small and statistically insignificant (0.08, SE 0.29), in contrast to the pre-event interaction in Column (3) (1.17, SE 0.30, significant at the 1% level). This placebo rules out the reverse-causality interpretation that firms planning stronger climate actions systematically shift their communication style in advance, generating a mechanical correlation between future disclosure tone and current announcement returns. Markets price the credibility signal embedded in *prior* communication, not anticipated *future* disclosure content. A complementary change-based placebo replaces the level $\hat{\pi}_{i,e+1}$ with the change $\Delta\hat{\pi}_{i,e+1}$; the interaction coefficient is similarly insignificant (0.06, SE 0.33; Table 26 Column (7)), confirming the result holds under both placebo constructions.

Nonlinearity check: credibility quintiles. Column (6) examines potential nonlinearity by interacting *ActionStrength* with quintile dummies of inferred credibility ($\hat{\pi}_{i,e-1}$; Q1 omitted). The coefficients increase monotonically from 0.51** (Q2) to 0.94*** (Q3) to 1.38*** (Q4) to 1.91*** (Q5), indicating that the pricing effect strengthens progressively across the credibility distribution. This pattern shows that the baseline interaction is not driven by extreme observations and is consistent with a smooth, increasing relation between inferred credibility and announcement returns.

Interpretation of the interaction coefficient. We interpret the estimated coefficient on $ActionStrength \times \hat{\pi}_{i,e-1}$ as capturing how market reactions to climate actions vary with prior credibility beliefs, not as a structural estimate of the model's comparative static. In particular, the coefficient reflects a reduced-form mapping between the observed proxies for belief and action, rather than a direct estimate of the underlying belief-updating mechanism $d/dD(\partial R/\partial A)$. More precisely, the coefficient on $ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$ is the marginal effect of $\hat{\pi}_{i,e-1}$ when $ActionStrength_{i,e} = 1$, or equivalently the marginal effect of action strength when $\hat{\pi}_{i,e-1} = 1$; for other values of these variables, the total marginal effect of either component also includes its main effect.

Economic magnitude in context. The estimated interaction implies that a one-unit increase in $\hat{\pi}_{i,e-1}$, when $ActionStrength_{i,e} = 1$, increases the three-day CAR by approximately 1.17 percentage points. Given that the average absolute announcement return in the sample is approximately 2.5%,[4] this effect represents a substantial share of typical announcement responses. The economic magnitude is as follows: a one-SD increase in inferred credibility corresponds to approximately 0.23 pp higher CAR at $ActionStrength = 1$ and approximately 0.12 pp at the sample mean of *ActionStrength* ($= 0.53$), both measured relative to a sample average absolute return of approximately 2.5%.

[4]Computed as $\mathbb{E}[|\text{CAR}|] \approx \sigma_{\text{CAR}} \times \sqrt{2/\pi} \approx 3.18 \times 0.798 \approx 2.54\%$, where $\sigma_{\text{CAR}} = 3.18\%$ is the cross-sectional standard deviation of CARs from Table 13. This Gaussian zero-mean approximation is used as a heuristic; CAR distributions are typically leptokurtic, so the approximation is appropriate for rough magnitude comparisons but not for tail-risk inference.

Action-type heterogeneity. The baseline pools diverse climate action types (net-zero pledges, green capex, renewable energy investments, divestitures, science-based targets, sustainability-linked financing). To rule out that the interaction reflects compositional differences across action types rather than within-type pricing, Column (5) of Table 15 adds a full set of action-type × *ActionStrength* interactions; the credibility interaction attenuates only to 1.13*** (SE 0.29). Within homogeneous subsets (net-zero pledges with explicit timelines; verifiable green capex), the interaction range is 1.18–1.21 (Online Appendix Table OA-10). The credibility-pricing channel operates within action types, not merely across them.

Beliefs versus fundamentals. Section OA-2.4 addresses the alternative interpretation that the interaction reflects time-varying execution capability rather than belief formation; the residualized specification confirms the credibility channel survives partialling out observable quality.

6.3 Follow-Through Tests

Table 16 examines the 10,808 follow-through implementation events—disclosures linked to prior climate pledges that report measurable progress (summary statistics in Table 13, Panel B). The estimating equation mirrors the baseline but replaces $ActionStrength_{i,e}$ with $FollowThrough_{i,e}$, a binary indicator for a concrete implementation step:

$$CAR_{i,e} = \beta_0 + \beta_1 FollowThrough_{i,e} + \beta_2 \hat{\pi}_{i,e-1} + \beta_3 (FollowThrough_{i,e} \times \hat{\pi}_{i,e-1}) + \Gamma X_{i,e-1} + \mu_i + \lambda_t + \varepsilon_{i,e}. \quad (32)$$

Consistent with **H2**, the credibility-pricing interaction is larger for follow-through events than for primary announcements at the upper end of the action-strength support: the binary follow-through interaction is 1.39*** in the preferred follow-through specification, compared with the continuous primary-event interaction of 1.19 evaluated at $ActionStrength = 1$ under the same fixed-effects structure. We caution against a literal percentage comparison: the 1.39 estimate is the differential credibility slope between binary follow-through events ($FollowThrough \in \{0, 1\}$) and primary events, while the 1.19 estimate is the credibility slope per unit of continuous $ActionStrength \in [0, 1]$. Both regressors span $[0, 1]$, so the coefficients are scale-comparable as effects of a one-unit change in the action-side regressor at $\hat{\pi}_{i,e-1} = 1$, but they parameterize different econometric objects (a discrete-event differential versus a continuous slope). At the sample mean of *ActionStrength* ($= 0.53$), the implied primary-event credibility slope is approximately $1.19 \times 0.53 \approx 0.63$, well below the binary follow-through differential, so the comparison is sensitive to the evaluation point. We do not emphasize the exact 17% numerical comparison; the substantive claim is qualitative—follow-through events exhibit a larger credibility-pricing interaction than primary announcements at the upper end of the action-strength distribution. The amplification reflects a composition effect: firms that both commit *and* subsequently implement are a selected subset for which the credibility proxy is more diagnostic, and the *FollowThrough* indicator absorbs mean-level differences across event types, leaving the interaction to capture within-type credibility variation.

Restricting the follow-through sample to the 1,856 firms that also appear in the primary event-study sample yields 1.37*** in the preferred specification (Table 16, bottom panel), confirming the result is not driven by firms outside the primary identification window. Because $\hat{\pi}_{i,e-1}$ for each follow-through event is constructed from the 12 months preceding that event (not the original commitment), the test asks whether *current* credibility amplifies the market response to implementation—the mechanism we predict.

6.4 Belief Updating

We test whether climate actions are followed by systematic changes in subsequent communication. Table 17 regresses $\hat{\pi}_{i,t+1}$ on a binary climate-action indicator at time t with firm and year fixed effects. Climate actions are followed by an increase of 0.064 (SE 0.015, significant at the 1% level) in subsequent credibility—approximately 0.30 standard deviations of the cross-sectional distribution. This supports **H3**.

Two caveats. First, because $\hat{\pi}_{i,t+1}$ is constructed from the firm's own subsequent disclosures, the test does not separately identify investor belief updating from action-induced changes in disclosure tone; we interpret the result as confirmatory but indirect evidence of the dynamic mechanism. Second, the binary action indicator tests only the directional prediction of Proposition 3, not the structural cross-partial $\partial^2\pi_2/\partial A\,\partial\pi_1 > 0$. Robustness to a fixed 12-month post-event window (rather than calendar-year anchoring) yields qualitatively identical results (Table OA-13).

6.5 Real Climate Outcomes

Table 23 tests whether credibility signals predict actual climate outcomes. Firms with higher credibility scores exhibit larger subsequent reductions in emissions intensity and greater investment in green technologies, confirming that the proxy captures economically meaningful variation in climate-related behavior beyond observed firm characteristics. The outcome variables in Table 23 (emissions intensity and green patents) differ from those in Table 14 (emissions reduction, green capex, and climate target adoption) by design: Table 14 validates the credibility proxy against short-run observable commitments and expenditure decisions, while Table 23 examines longer-run environmental performance. Table 23 reports two specifications with firm and year fixed effects. Column (1) regresses future emissions intensity on $\hat{\pi}_{i,t}$; the negative coefficient (-0.094^{***}, SE 0.021) indicates that higher inferred credibility predicts lower subsequent emissions intensity within the same firm over time. Column (2) uses green patenting activity as a complementary longer-run outcome; the positive coefficient ($+0.072^{***}$, SE 0.018) indicates that higher credibility predicts more green innovation. Cohen et al. (2026) document that traditional energy firms with low ESG scores are among the most prolific producers of high-quality green patents, implying that standard ESG ratings fail to track green-innovation output. Our credibility proxy $\hat{\pi}$, derived from disclosure style rather than ratings, positively predicts green patenting, consistent with the measure capturing credibility-relevant information that ratings miss. The two columns address complementary dimensions of future environmental performance rather than testing the same hypothesis on different samples. As emphasized in Section 5.5, predictive power for real outcomes does not imply that announcement returns reflect rational pricing of those outcomes; rather, $\hat{\pi}_{i,e-1}$ shapes returns through ex-ante belief formation about managerial credibility, prior to the realization of the outcomes documented here.

Transition to asset pricing tests. While the event-study evidence establishes that inferred credibility shapes immediate market reactions to climate announcements, Section 7 answers whether these beliefs are reflected in the broader cross-section of expected returns. Section 7 addresses this question through portfolio sorts, Fama-MacBeth regressions, reversal tests, and factor pricing.

6.6 Identification via Cross-Venue Credibility Shocks

Motivation. Even with controls for observable ESG characteristics, emissions intensity, green capital expenditures, and within-firm variation in $\hat{\pi}$, cross-sectional differences in inferred credibility may still reflect

unobserved firm quality. To address this, we exploit within-firm, short-horizon variation in the coherence of climate-related communication across disclosure venues—variation distinct from cross-rater ESG rating divergence (e.g., Berg et al., 2022; Gibson Brandon et al., 2021), which reflects methodological differences among external providers. Our measure captures firm-controlled communication coherence across sustainability reports, 10-K filings, and earnings calls; short-horizon inconsistencies provide variation in perceived credibility that does not mechanically alter the announced action or contemporaneous environmental outcomes. Because cross-venue consistency is firm-controlled, this design is treated as corroborating evidence; the frozen-prior greenwashing revelation DDD (Table 11) provides the cleaner exogenous variation.

Construction of the cross-venue consistency shock. We measure cross-venue communication consistency using Sentence-BERT embeddings (Reimers and Gurevych, 2019) of climate-related text extracted from sustainability reports, 10-K filings, and earnings call transcripts.[5] For each firm-quarter, we compute the average pairwise cosine similarity across these three channels. We define the cross-venue consistency shock as the negative change in this similarity over a one-quarter window:

$$Shock_{i,t} = -\Delta Consistency_{i,t},$$

so that higher values correspond to a deterioration in communication consistency (e.g., aspirational language in sustainability reports that is not reflected in the more formal 10-K or earnings call). The shock is standardized to have mean zero and unit standard deviation.

Empirical specification. We estimate the following event-study regression, using the frozen-prior credibility measure $\hat{\pi}_i^{\text{pre}}$ fixed at its last pre-shock value:

$$\begin{aligned} CAR_{i,e} = {} & \beta_1 ActionStrength_{i,e} + \beta_3 \big(ActionStrength_{i,e} \times \hat{\pi}_i^{\text{pre}}\big) \\ & + \gamma_1 Shock_{i,t-1} + \gamma_2 \big(Shock_{i,t-1} \times ActionStrength_{i,e}\big) + \gamma_3 (Shock_{i,t-1} \times \hat{\pi}_i^{\text{pre}}) \\ & + \beta_4 \big(Shock_{i,t-1} \times ActionStrength_{i,e} \times \hat{\pi}_i^{\text{pre}}\big) + \Gamma X_{i,e-1} + \mu_i + \lambda_t + \varepsilon_{i,e}. \end{aligned} \tag{33}$$

The main effect $\hat{\pi}_i^{\text{pre}}$ is omitted from the displayed equation because it is a time-invariant firm-level constant and is absorbed by the firm fixed effects μ_i; the interaction terms involving $\hat{\pi}_i^{\text{pre}}$ are identified because they vary within firm through $ActionStrength_{i,e}$ and $Shock_{i,t-1}$. The specification includes the Shock main effect γ_1 (not absorbed by firm or year fixed effects since Shock varies at the firm-quarter level), and the two lower-order interaction terms γ_2 and γ_3 required by the three-way interaction β_4. The coefficient of interest is β_4, which is displayed in Column (2) of Table 1. Under the intention-inference mechanism, a deterioration in perceived credibility attenuates the market response to otherwise identical climate actions, implying $\beta_4 < 0$.

Identification logic and supporting tests. This design isolates short-horizon variation in perceived managerial intent while holding fixed the observable action (*ActionStrength*) and slow-moving fundamentals. Because emissions, green investment, and target adoption evolve gradually, one-quarter declines in cross-venue consistency are more plausibly interpreted as changes in disclosure coherence rather than contempo-

[5] This quarterly three-venue *Consistency* differs from the 12-month four-venue $Consistency_{i,e-1}$ in Section 4.6: press releases are too irregular at quarterly frequency to support a Δ construction. The paper uses three distinct shock-related constructs: $Shock_{i,t-1}$ (continuous quarterly cross-venue drop, this section); $PostShock_e$ (binary post-revelation indicator, Section 8.3); and $CredibilityShock_{i,e}$ (AR(1) innovation in $\hat{\pi}$, Section 9.1).

raneous shifts in underlying climate commitment. All specifications include controls for contemporaneous earnings surprises, return volatility, and firm-level news sentiment, in addition to firm and year fixed effects.

Supporting tests confirm that the shock operates through the belief channel rather than fundamentals. First, a one-standard-deviation increase in $Shock_{i,t-1}$ reduces $\hat{\pi}_{i,t}$ by 0.11 ($p < 0.01$), indicating a strong first stage. Second, falsification regressions show that the shock has no predictive power for subsequent emissions intensity, green capital expenditures, or adoption of verifiable climate targets (all $p > 0.10$). Third, a placebo test shows no significant effect of the shock on returns in periods without climate-action announcements, indicating that the effect operates through the pricing of climate actions rather than a general repricing of the firm.

Tables 1 and 11 report the results. These are two distinct designs with separate estimands. *First*, the continuous cross-venue consistency shock: in the three-way specification, the two-way coefficient γ_2 on $ActionStrength \times Shock_{i,t-1}$ (Table 1, Column (2)) is -0.39 (SE 0.19) and measures the attenuation at $\hat{\pi}^{\text{pre}} = 0$. The marginal attenuation for a firm with prior credibility $\hat{\pi}^{\text{pre}}$ is $\gamma_2 + \beta_4 \hat{\pi}^{\text{pre}}$; evaluated near the sample mean of $\hat{\pi}$, the implied attenuation is about -0.67 percentage points ($-0.39 - 0.59 \times 0.47$). *Second*, the frozen-prior revelation DDD: the four-way interaction β_7^{DDD} (Table 11, Column (1)) is -0.68^{**} (SE 0.27), with β_7^{DDD} identified from within-revealed-firm pre-versus-post comparisons around the dated revelation event; never-revealed firms enter the specification only to estimate lower-order terms and absorb common time effects, not to identify β_7^{DDD} itself. Both designs show negative and significant attenuation after credibility-damaging shocks, though their coefficients are not literal percentage reductions of the baseline $ActionStrength \times \hat{\pi}$ slope.

Table 1: Cross-Venue Consistency Shocks: Two-Way and Three-Way Specifications

Dependent variable: CAR $[-1,+1]$	(1) Two-way	(2) Three-way
ActionStrength $\times\hat{\pi}_i^{\text{pre}}$	1.19***	1.17***
	(0.32)	(0.31)
$Shock_{i,t-1} \times ActionStrength_{i,e}$	−0.41**	−0.39**
	(0.18)	(0.18)
$Shock_{i,t-1} \times ActionStrength_{i,e} \times \hat{\pi}_i^{\text{pre}}$		−0.59**
		(0.29)
Controls	Yes	Yes
Firm FE	Yes	Yes
Year FE	Yes	Yes
Observations	4,812	4,812
Within R^2	0.20	0.21

Notes. This table reports a condensed version of the cross-venue consistency credibility-shock test. $Shock_{i,t-1}$ is the standardized negative change in cross-venue communication consistency. $\hat{\pi}_i^{\text{pre}}$ is fixed at its pre-shock value (frozen prior). Column (2) includes the triple interaction (-0.59^{**}, SE 0.29). The revelation-based frozen-prior four-way interaction is reported in Table 11 (baseline coefficient -0.68^{**}, SE 0.27, Column (1); robustness range -0.65 to -0.79 across Columns (1)–(6)). The revelation and cross-venue consistency designs use distinct treatment operationalizations—a dated RepRisk revelation event versus a continuous text-based consistency measure—and deliver economically similar but not identical effects (-0.68^{**} and -0.59^{**}, respectively). Column (6) of Table 11 additionally reports the revelation estimate after excluding events preceded by a consistency deterioration, which strengthens the coefficient to -0.79^{***}: the revelation design is therefore not simply picking up the consistency-shock variation, and the two approaches capture distinct—though potentially related—sources of credibility deterioration (external revelations versus internal communication inconsistency) rather than a single common mechanism. All lower-order terms required by the saturated model are included in estimation but suppressed for display. Supporting tests show that the shock reduces inferred credibility ($\Delta^{\text{shock}}\hat{\pi} \approx -0.11$, $p < 0.01$). Here $\Delta^{\text{shock}}\hat{\pi}$ denotes the post-minus-pre-shock change in the firm-level credibility proxy $\hat{\pi}_{i,t}$, measured over a 12-month window around the shock date. This is distinct from the 4-quarter first-difference $\Delta\hat{\pi}_{i,e-1} = \hat{\pi}_{i,e-1} - \hat{\pi}_{i,e-5}$ used in the within-firm specification of Section 8.1; we use the superscript $^{\text{shock}}$ here and throughout this subsection to distinguish the shock-based change from the event-aligned first-difference. The shock does not predict subsequent emissions, green capital expenditures, or target adoption, and has no effect on returns in non-action periods. Standard errors are double-clustered at the firm and year levels; Driscoll–Kraay estimates yield similar results. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

Interpretation. These results provide within-firm evidence that deteriorations in communication consistency attenuate the market response to climate actions, holding the action and observable fundamentals fixed. The strong first stage, falsification tests, and frozen-prior variant strengthen the identification case and address the concern that the baseline result reflects stable firm quality.

7 Asset Pricing Tests: The Intent Factor

The event-study evidence shows that investor beliefs about managerial credibility shape announcement returns. This section asks whether the same beliefs appear in the broader cross-section of returns. We construct a monthly Intent Factor based on the composite signal

$$Signal_{i,t} = ActionStrength_{i,t} \times \hat{\pi}_{i,t},$$

where t indexes calendar months. $ActionStrength_{i,t}$ is the score of firm i's most recent climate action within the prior 12 months and equals zero if no climate action occurred in that window; results are robust to an exponential-decay weighting (6-month half-life) in place of the hard 12-month lookback. These asset-pricing tests are reduced-form: they can reflect time-varying risk premia, temporary mispricing, or both. The long-horizon reversal evidence helps distinguish these interpretations, but the reversal is corroborating rather than the paper's primary identification.

Intent factor construction. Because 75% of firm-months have no climate action in the prior 12 months, a full-universe quintile sort would allocate many zero-signal observations arbitrarily. We therefore sort active firms into quintiles based on $Signal_{i,t}$. The long leg is the top active-firm quintile; the short leg is all zero-signal firms:

$$IntentFactor_t = R_t^{HighIntent} - R_t^{ZeroIntent}.$$

This factor prices the composite action-credibility signal. Results are robust to defining the short leg as the bottom quintile of active firms (Online Appendix). Double sorts below isolate credibility while conditioning on action strength.

Time-series alpha. Table 2 reports positive abnormal returns. Monthly alphas are 0.42% under the CAPM ($t = 3.21$), 0.31% under FF4 ($t = 2.48$), and 0.28% under FF5, corresponding to annualized alphas of 5.04%, 3.72%, and 3.36%. The alphas survive controls for size, value, momentum, profitability, and investment. Because the short leg consists of inactive firms, we interpret the factor as a composite contrast between high-intent active firms and inactive firms, not as a pure credibility portfolio.

The benchmark models do not include climate-specific factors. A climate-factor-augmented benchmark may therefore attenuate the alpha. Consistent with this possibility, climate controls reduce the 12-month reversal coefficient in Table 6 from -0.19^{***} to -0.04.

Table 2: The Intent Factor: Long High-Signal Minus Low-Signal Firms

	(1) CAPM	(2) FF4 (MKT, SMB, HML, MOM)	(3) FF5 (MKT, SMB, HML, RMW, CMA)
α (monthly, %)	0.42***	0.31**	0.28**
	(0.13)	(0.13)	(0.13)
MKT	−0.18**	−0.21**	−0.19**
	(0.08)	(0.09)	(0.09)
SMB		0.09	0.08
		(0.07)	(0.07)
HML		0.12	0.11
		(0.08)	(0.08)
MOM		0.24**	
		(0.10)	
RMW			0.18**
			(0.08)
CMA			−0.06
			(0.09)
Annualized α (%)	5.04	3.72	3.36
Observations (months)	228	228	228
R^2	0.08	0.14	0.17

Dependent variable: monthly excess return (%) on the long-short "Intent Factor" portfolio, $R_t^{\text{long}} - R_t^{\text{short}}$. Sample: 228 monthly observations (January 2005–December 2023).

Intent factor construction. The factor is formed using a two-stage procedure to avoid arbitrary tie-breaking among the large mass of zero-signal observations (≈75% of firm-months have no climate action in the prior 12 months). Firms with non-zero $Signal_{i,t} = ActionStrength_{i,t} \times \hat{\pi}_{i,t}$ are sorted into quintiles; the long leg is the top quintile of this active-firm distribution. The short leg consists of *all* firms with zero signal (no climate-action event in the prior year), value-weighted. This construction ensures full reproducibility and preserves the economically meaningful contrast between high-intent actors and inactive firms. Results are robust to defining the short leg as the bottom quintile of active firms (Appendix OA). The sample spans January 2005 to December 2023 (228 months). All regressions are time-series regressions of monthly excess returns on risk factors from Ken French's data library. Column (2) uses the Fama-French four-factor model (MKT, SMB, HML, MOM). Column (3) uses the Fama-French five-factor model (MKT, SMB, HML, RMW, CMA) without momentum. Annualized α is computed by multiplying the monthly alpha by 12. Standard errors are Newey-West (Newey and West, 1987) adjusted with 12 lags. Displayed standard errors are rounded to two decimal places; t-statistics are computed from unrounded standard errors and therefore may differ slightly from dividing the displayed α by the displayed SE (e.g., the CAPM $t = 3.21$ and FF4 $t = 2.48$ reflect unrounded computation). ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Double sorts on action strength and credibility. Table 3 separates the credibility channel from action-strength composition. Within low-action firms, returns rise modestly across credibility quintiles, from 0.14% to 0.22% per month. Within high-action firms, returns rise from 0.28% to 0.65%, and the high-minus-low action spread increases monotonically from 0.14% to 0.43% across credibility quintiles. This gradient is the portfolio analogue of the positive event-study interaction.

Table 3: Double Sorts on ESG Action Strength and Credibility

	Credibility ($\hat{\pi}_{i,t}$ quintiles)				
ActionStrength	Low	Q2	Q3	Q4	High
Low *ActionStrength*	0.14	0.17	0.19	0.20	0.22
High *ActionStrength*	0.28	0.37	0.46	0.54	0.65
High − Low	0.14	0.20*	0.27**	0.34***	0.43***
	(0.11)	(0.12)	(0.13)	(0.13)	(0.14)

Cell entries: value-weighted average monthly excess returns (%), averaged across 228 months (2005m1–2023m12); approximately 850 firms per month, yielding roughly 85 firms per quintile-half cell on average. Because approximately 75% of firm-months have $ActionStrength_{i,t} = 0$ (no climate action in the prior 12 months), the cell-size distribution is asymmetric under the median split used here: zero-signal firms are placed in the low-*ActionStrength* half (since the sample median of $ActionStrength_{i,t}$ is zero), so low-*ActionStrength* cells are substantially larger than high-*ActionStrength* cells. This asymmetry implies that the High−Low spread coefficients in the extreme high-*ActionStrength* quintiles inherit lower precision than the average cell size of 85 would suggest.

Notes. This table reports value-weighted average monthly excess returns from double-sorted portfolios formed on ESG action strength and inferred credibility. At the beginning of each month, firms are sorted into two groups based on $ActionStrength_{i,t}$ (above/below median) and independently into quintiles of inferred credibility $\hat{\pi}_{i,t}$; portfolio returns are then computed for the subsequent month. The sample spans January 2005 to December 2023 (228 months); the average number of firms per month is approximately 850. Reported portfolio returns are monthly percentages. The "High−Low" row reports within-column differences: the return on the high-ActionStrength portfolio minus the low-ActionStrength portfolio *within each credibility quintile*. This captures the action-strength spread conditional on credibility level; the monotone increase in this spread across credibility quintiles reflects the cross-partial prediction of the model. Standard errors for spread portfolios are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Fama-MacBeth return predictability. Table 4 shows that $Signal_{i,t}$ predicts one-month-ahead returns. The coefficient is 0.18 without controls, 0.14 with standard controls, and 0.11 after adding momentum. The low-credibility interaction is negative (-0.09^{**} without momentum and -0.08^{**} with momentum), consistent with weaker pricing when prior credibility is low. With standardized regressors, the final estimate implies about 1.3% higher annualized excess returns per one-standard-deviation increase in the signal. The Fama-MacBeth estimate and the Intent Factor alpha measure different objects, so we interpret them as complementary.

Table 4: Predictability of Returns: Fama-MacBeth Regressions

	(1)	(2)	(3)
$Signal_{i,t}$ ($ActionStrength \times \hat{\pi}$)	0.18***	0.14***	0.11**
	(0.05)	(0.05)	(0.05)
$ActionStrength \times \mathbf{1}[\hat{\pi} \in Q1]$		−0.09**	−0.08**
		(0.04)	(0.04)
Size (log ME)		−0.06**	−0.05*
		(0.03)	(0.03)
Book-to-market		0.09**	0.08**
		(0.04)	(0.04)
Momentum			0.12***
			(0.03)
Months (cross-sections)	228	228	228
Avg. firms/cross-section	≈850	≈850	≈850
Avg. R^2	0.02	0.06	0.09

Notes. Fama-MacBeth cross-sectional regressions of next-month excess returns. The dependent variable is monthly excess return. The primary regressor is $Signal_{i,t} = ActionStrength_{i,t} \times \hat{\pi}_{i,t}$. Where included, the corresponding main effects enter separately but are suppressed from the table for brevity. All continuous regressors are standardized to zero mean and unit standard deviation. Controls include size (log market equity), book-to-market, and momentum. The sample spans January 2005 to December 2023 (228 monthly cross-sections; average of approximately 850 firms per cross-section). Standard errors are Newey–West with 12 lags; Driscoll–Kraay estimates yield similar results. Standard errors are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Placebo and falsification tests. Table 5 shows that predictability is specific to the ESG credibility-action signal. The coefficient disappears when we use a non-ESG signal, a twice-lagged signal, or a within-month permutation of $\hat{\pi}_{i,t}$.

Table 5: Placebo and Falsification Tests: Fama-MacBeth Regressions

	(1) Baseline	(2) Non-ESG signal (placebo)	(3) Lagged signal placebo ($t-2$)	(4) Permutation placebo
$Signal_{i,t}$	0.11**	0.02	0.03	−0.01
	(0.05)	(0.04)	(0.04)	(0.04)
$ActionStrength \times \mathbf{1}[\hat{\pi} \in Q1]$	−0.08**	−0.01	−0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Controls	Yes	Yes	Yes	Yes
Months (cross-sections)	228	228	228	228
Avg. firms/cross-section	≈850	≈850	≈850	≈850
Avg. R^2	0.09	0.05	0.05	0.04

Notes. Fama-MacBeth cross-sectional regressions of next-month excess returns. Column (1) repeats the baseline specification; Column (2) replaces the ESG-based signal with an analogous non-ESG disclosure signal; Column (3) uses the twice-lagged signal $Signal_{i,t-2}$; and Column (4) permutes $\hat{\pi}_{i,t}$ across firms within month while preserving its monthly cross-sectional distribution. Controls include size (log market equity), book-to-market, and momentum. The sample spans January 2005 to December 2023 (228 monthly cross-sections; average of approximately 850 firms per cross-section). Newey–West standard errors with 12 lags are reported in parentheses, following standard practice in monthly return predictability studies to accommodate autocorrelation up to one year. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Long-horizon reversal. Tables 6 and 7 examine post-signal return dynamics. The composite signal predicts positive returns at one month ($+0.12^{***}$) but negative returns at longer horizons: -0.09^{**} at 6 months and -0.19^{***} at 12 months in the panel specification, with similar signs in Fama-MacBeth regressions. The sign change from short to long horizons is the canonical overreaction-and-correction pattern (see, e.g., Hirshleifer and Teoh, 2003). Climate controls attenuate the 12-month coefficient to -0.04, indicating that part of the reversal is correlated with priced climate risk.

A purely persistent-risk interpretation would more naturally predict persistence rather than a sign change, although time-varying climate risk exposures may explain part of the longer-horizon pattern. The reversal is therefore more consistent with belief correction than with a purely persistent-risk interpretation, but we treat it as corroborating, not load-bearing, evidence. The announcement-window interaction, BToM-specific heterogeneity tests, and frozen-prior revelation design provide the paper's primary identification.

Table 6: Long-Horizon Return Reversals: Panel Regressions with Firm and Year Fixed Effects

Horizon	1-month	6-month	12m (no controls)	12m (+climate controls)
Composite signal ($ActionStrength \times \hat{\pi}$)	$+0.12^{***}$	-0.09^{**}	-0.19^{***}	-0.04
	(0.04)	(0.04)	(0.07)	(0.05)
Climate controls	No	No	No	Yes
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	197,000	197,000	197,000	197,000
Within R^2	0.12	0.14	0.15	0.16

The sample is an event-firm panel consisting of 850 firms that issued at least one climate action during the sample period (197,000 firm-months). The dependent variable is the cumulative market-model adjusted excess return over the indicated horizon, expressed in percentage points and constructed from monthly excess returns. The composite signal is the standardized interaction $ActionStrength \times \hat{\pi}$ measured at the end of month $t-1$. Column (4) adds the full set of climate-related controls (carbon intensity, green investment measures, and climate-news shocks). Standard errors are double-clustered by firm and year. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 7: Long-Horizon Return Reversals: Fama–MacBeth Cross-Sectional Regressions

	6-month ahead returns		12-month ahead returns	
	(1) Baseline	(2) + Controls	(3) Baseline	(4) + Controls
Composite signal ($ActionStrength \times \hat{\pi}$)	-0.08^{*}	-0.02	-0.17^{***}	-0.04
	(0.04)	(0.04)	(0.07)	(0.06)
Climate controls	No	Yes	No	Yes
Observations (firm-months)	197,000	197,000	197,000	197,000

The sample is the same event-firm panel as in Table 6 (850 firms; 197,000 firm-months). In each month t, we estimate a cross-sectional regression of future cumulative excess returns (constructed analogously to Table 6) over the subsequent 6- or 12-month horizon on the composite signal and, where indicated, climate-related controls. Reported coefficients are time-series averages of the monthly cross-sectional slopes (Fama–MacBeth). Standard errors are Newey–West adjusted with 6 lags for the 6-month horizon and 12 lags for the 12-month horizon to account for overlapping returns. The composite signal is measured at the end of month $t-1$. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Mechanism: strategic versus sincere firms. Table 8 splits the event sample by prior credibility. The action-strength slope is larger for high-credibility firms than for low-credibility firms: 1.24 versus 0.89. The marginal effects implied by the full interaction model are 1.29 and 0.91, respectively. These results mirror the event-study interaction and confirm the cross-partial prediction of Proposition 2: market sensitivity to action strength is increasing in inferred credibility.

Table 8: Heterogeneity by Inferred Credibility (Mechanism Test)

		Panel A: Unconditional slopes	
Subsample	AS coeff.	Obs.	R^2
Low credibility ($\hat{\pi}_{i,e-1} \leq$ median)	0.89***	2,406	0.18
	(0.24)		
High credibility ($\hat{\pi}_{i,e-1} >$ median)	1.24***	2,406	0.21
	(0.27)		
	Panel B: Marginal effects implied by full interaction model		
	Preferred: $\hat{\beta}_1 = 0.55$, $\hat{\beta}_3 = 1.17^{***}$		
Subsample	Mean $\hat{\pi}$ in bin	Implied marginal effect of ActionStrength	Difference vs. low bin
Low credibility ($\hat{\pi}_{i,e-1} \leq$ median)	0.31	0.91	–
High credibility ($\hat{\pi}_{i,e-1} >$ median)	0.63	1.29	+0.38***

Notes. This table tests the intention-inference mechanism by examining whether the sensitivity of announcement returns to *ActionStrength* varies with prior inferred credibility ($\hat{\pi}_{i,e-1}$).

Panel A vs. Panel B. Panel A reports coefficients from regressions of CAR$[-1,+1]$ on *ActionStrength* estimated separately within low- and high-credibility subsamples. These coefficients are *group-level averages* of the marginal effect and need not equal the marginal effect evaluated at the group mean.

Formally, under the interaction specification

$$CAR_{i,e} = \beta_1 \, ActionStrength_{i,e} + \beta_2 \hat{\pi}_{i,e-1} + \beta_3 \big(ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \big) + \ldots,$$

the marginal effect is

$$\frac{\partial CAR_{i,e}}{\partial ActionStrength_{i,e}} = \beta_1 + \beta_3 \hat{\pi}_{i,e-1}.$$

The split-sample coefficient in bin g can be written as

$$\tilde{\beta}_1^{(g)} = \mathbb{E}[\beta_1 + \beta_3 \hat{\pi}_{i,e-1} \mid g] + \beta_3 \cdot \mathrm{Cov}\big(ActionStrength_{i,e}, \hat{\pi}_{i,e-1} \mid g \big),$$

so it reflects both the average marginal effect and within-bin composition.

Implied marginal effects. Panel B evaluates the marginal effect implied by the preferred interaction specification:

$$\frac{\partial \mathrm{CAR}}{\partial ActionStrength} = 0.55 + 1.17 \times \hat{\pi}_{i,e-1}.$$

Evaluating at the mean $\hat{\pi}$ within each bin yields 0.91 (low) and 1.29 (high). The formula $1.17 \times (0.63 - 0.31) = 0.374$ rounds to 0.37, while the table reports +0.38 because the displayed marginal effects are themselves rounded to two decimal places (round(1.29) − round(0.91) = 0.38); the discrepancy is a rounding artifact, not a numerical inconsistency. These implied effects are strictly increasing in $\hat{\pi}$, consistent with the positive interaction coefficient.

Interpretation. Both panels deliver a consistent conclusion: the sensitivity of announcement returns to *ActionStrength* is higher for firms with more credible prior communication. Panel B shows this directly from the interaction model, while Panel A provides supporting reduced-form evidence.

Specification. All specifications include the full set of controls (log assets, book-to-market, leverage, prior returns, emissions intensity, analyst coverage) and industry×year fixed effects. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Heterogeneity in return predictability. Table 9 shows that predictability is stronger where information frictions are higher. The signal coefficient rises from 0.11 in the full sample to 0.19 among low-analyst-coverage firms, and the direct *Signal* × *LowAnalyst* interaction is +0.10**. Predictability is also stronger among high-controversy firms (0.16 versus 0.08). These patterns are consistent with the credibility channel being most relevant when investor learning is harder and greenwashing risk is more salient.

Table 9: Predictability of Returns: Fama-MacBeth Regressions and Heterogeneity

	(1) Full sample	(2) Low analyst coverage	(3) Full + analyst interaction	(4) High controversy	(5) Low controversy
$Signal_{i,t}$	0.11**	0.19***	0.12**	0.16***	0.08*
	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)
$ActionStrength \times \mathbf{1}[\hat{\pi} \in Q1]$	−0.08**	−0.12**	−0.08**	−0.11**	−0.06
	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)
$\mathbf{1}[LowAnalyst_{i,t}]$			−0.03		
			(0.04)		
$Signal_{i,t} \times \mathbf{1}[LowAnalyst_{i,t}]$			+0.10**		
			(0.05)		
Size (log ME)	−0.05*	−0.07**	−0.05*	−0.06**	−0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Book-to-market	0.08**	0.10**	0.08**	0.09**	0.07*
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Momentum	0.12***	0.13***	0.12***	0.14***	0.11***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Months (cross-sections)	228	228	228	228	228
Avg. R^2	0.09	0.11	0.09	0.10	0.08

Notes. Fama-MacBeth cross-sectional regressions of monthly excess returns with heterogeneity terms added to the baseline signal specification. The dependent variable is next-month excess return. $Signal_{i,t} = ActionStrength_{i,t} \times \hat{\pi}_{i,t}$. $\mathbf{1}[\hat{\pi} \in Q1]$ is an indicator for the bottom credibility quintile. $\mathbf{1}[LowAnalyst_{i,t}]$ is an indicator for the bottom tercile of analyst coverage. Column (2) restricts the sample to low-analyst-coverage firms. Column (3) includes both the main effect of $\mathbf{1}[LowAnalyst_{i,t}]$ and its interaction with $Signal_{i,t}$, so that the interaction coefficient captures differential signal pricing rather than baseline return differences across coverage groups. Columns (4) and (5) split the sample by ESG controversy (above/below the prior-year median). All continuous regressors are standardized to zero mean and unit standard deviation. The sample spans January 2005 to December 2023 (228 monthly cross-sections; average of approximately 850 firms per cross-section in Column (1); subsample columns have correspondingly smaller cross-sections). Controls include the standard asset-pricing characteristics used elsewhere in the paper. Newey–West standard errors with 12 lags are reported in parentheses, following standard practice in monthly return predictability studies to accommodate autocorrelation up to one year. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Caveat: reverse causality and selection. The asset-pricing tests are observational. Strategic timing and sample selection cannot be fully ruled out, so we interpret these results as complementary to the within-firm, regulatory, and revelation-based identification tests that follow.

8 Identification Strategy

Our empirical objective is to determine whether investors price climate actions through beliefs about managerial credibility rather than through correlated firm quality, implementation capacity, or disclosure sophistication. We address this challenge with three complementary layers of evidence: within-firm changes in inferred credibility, greenwashing-revelation shocks with a frozen-prior DDD, and quasi-experimental variation from disclosure standardization and climate salience. Attenuation percentages are benchmarked to the within-specification baseline interaction because samples and fixed-effect structures differ across designs;

they should therefore be compared qualitatively rather than as common-denominator estimates. Throughout this section, index e denotes events (or event-equivalent quarters in within-firm analyses), while t indexes calendar time.

8.1 Within-Firm Variation in Credibility

The estimating sample remains at the event level: each observation is a climate-action announcement. We construct $\Delta\hat{\pi}_{i,e-1} = \hat{\pi}_{i,e-1} - \hat{\pi}_{i,e-5}$, the four-quarter change in perceived credibility ending before the event. For most firms, the two trailing 12-month disclosure windows are non-overlapping; for dense-event firms, results are robust to requiring events spaced at least 12 months apart and to using calendar-year anchoring rather than event-order anchoring. Lagged levels of $\hat{\pi}$ are included in some specifications to absorb slow-moving disclosure trends.

Table 26 reports estimates with firm fixed effects, firm-specific linear trends, and the full set of controls. The interaction $ActionStrength_{i,e} \times \Delta\hat{\pi}_{i,e-1}$ remains large and statistically significant, declining from 1.45 in Column (1) to 1.24 in Column (5) as controls are added and to 1.16 in the residualized specification (Column (6)). Lead terms are economically small and insignificant, indicating that future changes in perceived credibility do not predict current announcement returns. Forward-looking communication adjustments are also unlikely to explain the result: the correlation between $\hat{\pi}_{i,e-1}$ and $ActionStrength_{i,e}$ is only 0.14, and the interaction is stable after residualizing $\hat{\pi}_{i,e-1}$ against observable action-quality proxies (Table OA-2, Column (5)).

These results indicate that, even within the same firm, increases in perceived credibility amplify the market reaction to climate actions, ruling out explanations based solely on static firm quality, persistent disclosure style, or time-invariant managerial characteristics.

8.2 Cross-Sectional Identification

The baseline cross-sectional design is the event-study specification estimated in Section 6. We restate it here only to fix notation for the identification tests that follow:

$$CAR_{i,e} = \beta_0 + \beta_1 ActionStrength_{i,e} + \beta_2 \hat{\pi}_{i,e-1} + \beta_3 \left(ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}\right) + \Gamma X_{i,e-1} + \delta_{j(i)\times t} + \varepsilon_{i,e}, \quad (34)$$

where $CAR_{i,e}$ is the cumulative abnormal return over the $[-1,+1]$ window and $\delta_{j(i)\times t}$ denotes industry-by-year fixed effects, with $j(i)$ denoting firm i's industry. The credibility proxy is constructed entirely from pre-event communication, while $ActionStrength_{i,e}$ is coded from the announcement using a blinded rubric. Thus, reverse causality from announcement returns to measured credibility is mitigated by timing, but time-varying unobserved firm quality remains the central threat addressed by the within-firm, revelation, and quasi-experimental designs below.

8.3 Greenwashing Revelation Shocks

A direct test of the intention-inference mechanism requires exogenous variation in perceived managerial credibility that is plausibly orthogonal to underlying firm quality and implementation capacity at the time of the climate-action announcement. Greenwashing revelations provide such a setting: external third-party disclosures of previously hidden misalignment between communication and actions update investors' beliefs

about managerial credibility *after* the original announcement, without mechanically changing the original climate action or the fundamentals known at the announcement date.

Identification does not rely on the market reaction to the revelation itself. Rather, we exploit how the *same ex-ante signal* (*ActionStrength* interacted with the pre-event credibility proxy $\hat{\pi}_{i,e-1}$) is priced differently once investors receive credible information contradicting prior high-credibility priors. This isolates belief revision in the pricing identity (20).

We identify revelation events using RepRisk ESG risk incident data, which flags greenwashing as the intersection of "misleading communication" and an environmental issue, along with severity, reach, and novelty scores. We focus on "credible" revelations—those with at least medium severity and substantial media reach—to minimize noise from minor or unsubstantiated claims. For each firm with at least one such event, we split climate-action announcements into pre- and post-revelation subsamples based on the date of the first qualifying revelation.

The main specification is a triple-difference design:

$$\begin{aligned} CAR_{i,e} = \beta_0 &+ \beta_1 ActionStrength_{i,e} + \beta_2 \hat{\pi}_{i,e-1} + \beta_3 \big(ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}\big) \\ &+ \beta_4 PostReveal_{i,e} + \beta_5 \big(ActionStrength_{i,e} \times PostReveal_{i,e}\big) \\ &+ \beta_6 (\hat{\pi}_{i,e-1} \times PostReveal_{i,e}) \\ &+ \beta_7^{\text{rev}} \big(ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times PostReveal_{i,e}\big) + \gamma \mathbf{X}_{i,e} + \mu_i + \delta_{j(i)\times t} + \varepsilon_{i,e}, \end{aligned} \tag{35}$$

where $PostReveal_{i,e}$ equals one for climate-action events occurring after the firm's first credible greenwashing revelation, μ_i are firm fixed effects, and $\delta_{j(i)\times t}$ are industry-by-year fixed effects ($j(i)$ denotes firm i's industry). The coefficient β_7^{rev} is the key object of interest and is predicted to be negative: the credibility-pricing interaction should attenuate once investors update their beliefs about managerial credibility. The superscript distinguishes this three-way revelation coefficient from the four-way frozen-prior DDD coefficient β_7^{DDD} below.

Table 10 reports the results. In the pre-revelation subsample, the interaction between *ActionStrength* and prior inferred credibility $\hat{\pi}_{i,e-1}$ is positive and significant at the 1% level (1.48, SE 0.39). Post-revelation, the same interaction collapses sharply to 0.41 (SE 0.32) and loses statistical significance. In the pooled specification, the triple-difference coefficient β_7^{rev} is negative and significant (-0.97, SE 0.34). Column (4) excludes events within ± 30 days of the revelation to mitigate anticipation effects and yields virtually identical results.

Two identification points deserve clarification. First, $Revealed_i$ is used as a *post-hoc sample partition* rather than as an instrument or causal treatment: we split the sample by whether a firm is eventually revealed to have greenwashed, to examine how pricing changes after the revelation. The endogeneity of RepRisk coverage (larger, more media-exposed firms are more likely to be flagged) does not invalidate the β_7^{rev} estimate, which is identified from within-firm pre-versus-post comparisons. A remaining limitation is that RepRisk revelations may correlate with unobservable firm characteristics—such as heightened political exposure, activist targeting, or latent litigation risk—that independently affect announcement returns; the matched-sample design addresses observable selection but cannot fully rule out unobservable selection on these dimensions. Second, the frozen-prior design in Table 11 is the preferred specification precisely because it does not rely on $Revealed_i$ as exogenous: it uses the revelation event as a dated shock and compares revealed-firm pre-shock and post-shock observations using a frozen $\hat{\pi}^{\text{pre}}$ measured before the shock. Never-revealed firms enter the regression only to estimate lower-order terms and absorb common time effects; identification

of β_7^{DDD} comes from within-revealed-firm pre-versus-post variation and not from a Revealed-versus-not-Revealed cross-sectional contrast. Third, $Revealed_i$ is right-censored: events near the end of the sample have less post-event time in which to observe a revelation, so some late-sample firms classified as never-revealed may in fact be uncensored greenwashers whose revelation falls outside the observation window. The estimate of β_7^{rev} in Equation (35) is largely robust to this concern because β_7^{rev} is identified *within revealed firms* from pre-versus-post variation, with firm fixed effects; never-revealed firms have $PostReveal = 0$ throughout and therefore do not contribute to identifying β_7^{rev} directly. Right-censoring would primarily affect cross-firm contrasts (e.g., comparing the level of the credibility-pricing slope between Revealed and never-Revealed firms), not the within-revealed-firm change at revelation. The same reasoning applies to the frozen-prior DDD in Table 11, which is also identified from within-revealed-firm pre-versus-post variation.

This pattern argues against a pure firm-quality interpretation. If $\hat{\pi}$ primarily captured persistent organizational capability or disclosure sophistication, the pricing of subsequent climate actions should remain stable—or even strengthen—after a revelation, as markets learn about underlying weaknesses. Instead, the post-revelation decline is consistent with earlier high returns reflecting temporary over-attribution of credibility that was subsequently corrected.

Placebo tests support this interpretation. Re-estimating Equation (35) using (i) non-climate corporate announcements (e.g., product launches, non-climate capex) and (ii) ESG controversies unrelated to environmental or climate claims (e.g., labor or governance incidents) yields insignificant triple-difference coefficients. For the non-climate placebo, *ActionStrength* is replaced by an analogous specificity-and-scope index constructed for non-climate announcements; $\hat{\pi}_{i,e-1}$ retains its climate-communication-based construction. The attenuation is thus specific to updates in climate-related credibility.

Table 10: Greenwashing Revelations as Credibility Shocks: Pre- vs. Post-Revelation Pricing

	(1) Pre	(2) Post	(3) Pooled	(4) Pooled (excl. ±30d)
Dependent variable: CAR $[-1,+1]$				
$ActionStrength_{i,e}$	0.62***	0.48**	0.59***	0.57***
	(0.19)	(0.21)	(0.18)	(0.18)
$\hat{\pi}_{i,e-1}$	0.44*	0.12	0.39*	0.37*
	(0.23)	(0.25)	(0.21)	(0.21)
$ActionStrength \times \hat{\pi}_{i,e-1}$	1.48***	0.41	1.39***	1.34***
	(0.39)	(0.32)	(0.37)	(0.36)
$PostReveal_{i,e}$			−0.31	−0.28
			(0.28)	(0.27)
$ActionStrength \times \hat{\pi}_{i,e-1} \times PostReveal$			−0.97***	−0.91***
			(0.34)	(0.33)
Lower-order interactions	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Action-type indicators	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry × Year FE	Yes	Yes	Yes	Yes
Observations	2,841	1,971	4,812	4,506
Within R^2	0.21	0.18	0.23	0.23

Notes. Pre-revelation subsample (Column (1), $N = 2{,}841$) includes two groups: (i) climate actions occurring before a firm's first credible greenwashing revelation, and (ii) all actions by firms that never experience a qualifying revelation over the sample period. Post-revelation subsample (Column (2), $N = 1{,}971$) includes only actions by firms that do experience a revelation, occurring after that firm's first qualifying revelation date. Firm fixed effects absorb singleton firms (those with exactly one post-revelation event), which contribute no identifying variation; the effective identifying sample is therefore smaller than the reported $N = 1{,}971$. Results are robust to excluding singletons and to specifications without firm fixed effects. The pooled specification (Column (3)) uses the full event sample ($N = 4{,}812$) with $PostReveal_{i,e} = 1$ only for post-revelation events of eventually-revealed firms; never-revealed firms contribute to the pre-period baseline and are assigned $PostReveal_{i,e} = 0$ throughout. A credible revelation is defined as a RepRisk incident of medium or high severity intersecting misleading communication and environmental issues with substantial media reach. The full event sample of 4,812 observations is used: firms that never experience a credible revelation have $PostReveal_{i,e} = 0$ for all their events and therefore contribute entirely to the pre-revelation group. The pre-revelation coefficient (1.48***) is accordingly a mixture of always-clean firms and eventually-revealed firms before their first revelation; the triple-difference β_7^{rev} identifies the *within-revealed-firm* change in credibility pricing after the revelation date, holding the never-revealed firms in the baseline. Robustness checks restricting to only firms that eventually experience a revelation (so that the pre/post comparison is purely within-firm) yield qualitatively identical results. Column (3) pools the sample and includes the triple-difference coefficient β_7^{rev} from Equation (35). Column (4) excludes events within ±30 days of the revelation date. Controls include log assets, book-to-market, leverage, emissions intensity, and analyst coverage. Standard errors are double-clustered at the firm and year levels. The sharp attenuation post-revelation indicates that markets initially over-attributed credibility and corrected upon revelation of misalignment, providing direct evidence of the belief-based pricing channel. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Frozen-prior triple-difference (Table 11). The rolling $\hat{\pi}_{i,e-1}$ continues to update after a revelation shock, which could mechanically correlate the shock with the regressor. We therefore fix $\hat{\pi}_i^{\text{pre}}$ at its last pre-shock value and estimate a frozen-prior DDD:

$$\begin{aligned} CAR_{i,e} = {} & \beta_1 \mathit{ActionStrength}_{i,e} + \beta_2 \hat{\pi}_i^{\text{pre}} + \beta_3 \big(\mathit{ActionStrength}_{i,e} \times \hat{\pi}_i^{\text{pre}}\big) \\ & + \beta_4 (\mathit{PostShock}_e \times \mathit{Revealed}_i) + \beta_5 \big(\mathit{PostShock}_e \times \mathit{Revealed}_i \times \mathit{ActionStrength}_{i,e}\big) \\ & + \beta_6 (\mathit{PostShock}_e \times \mathit{Revealed}_i \times \hat{\pi}_i^{\text{pre}}) \\ & + \beta_7^{\text{DDD}} \big(\mathit{PostShock}_e \times \mathit{Revealed}_i \times \mathit{ActionStrength}_{i,e} \times \hat{\pi}_i^{\text{pre}}\big) + \mathbf{X}'_{i,e-1}\Gamma + \mu_i + \lambda_t + \varepsilon_{i,e}, \end{aligned} \tag{36}$$

where $\mathit{PostShock}_e = 1$ for events occurring after the firm's first credible greenwashing revelation date (the same threshold used to define $\mathit{PostReveal}_{i,e}$ in Equation (35); the change in notation reflects the shock-time interpretation of this DDD specification). For never-revealed firms ($\mathit{Revealed}_i = 0$), $\mathit{PostShock}_e$ is set to zero throughout, so all $\mathit{Revealed}_i \times \mathit{PostShock}_e$ interaction terms vanish and these firms contribute only to the lower-order main effects and to the firm fixed-effect partition. Identification of β_7^{DDD} comes from within-revealed-firm pre-versus-post comparisons. The main effect of $\hat{\pi}_i^{\text{pre}}$ is absorbed by firm fixed effects; the object of interest is the four-way interaction β_7^{DDD}. Table 11 reports a baseline estimate of -0.68^{**} (SE 0.27). The coefficient remains tightly clustered across controls, matched samples, alternative revelation definitions, and IPW specifications (-0.65 to -0.71), and strengthens to -0.79^{***} when events preceded by cross-venue consistency deteriorations are excluded. The matched-sample specification balances revealed and never-revealed firms on pre-shock observables (all post-match p-values ≥ 0.64; Online Appendix Table OA-16) and yields -0.65^{**}. This is the paper's sharpest quasi-experimental test because the credibility regressor is held fixed while investor interpretation changes after an external revelation shock.

Table 11: Frozen-Prior Revelation DDD: Robustness to Non-Overlapping Shocks

	(1) Base	(2) Ctrl.	(3) Match	(4) Alt.	(5) IPW	(6) No cons. shock
$AS \times \hat{\pi}^{pre}$ $\times$ *Revealed* $\times$ *Post* (β_7^{DDD})	−0.68** (0.27)	−0.71** (0.28)	−0.65** (0.26)	−0.69** (0.27)	−0.67** (0.27)	−0.79*** (0.30)
Controls	No	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Matched sample	No	No	Yes	No	No	No
IPW	No	No	No	No	Yes	Yes
Excl. consistency shocks	No	No	No	No	No	Yes
Observations	4,812	4,812	4,812	4,812	4,812	3,941

Notes. The main four-way interaction is −0.68** (Column (1)). When events preceded by a consistency deterioration are excluded (Column (6)), the coefficient strengthens to −0.79***. The consistency-shock design (Table 1, Column (2)) yields −0.59** (SE 0.29). The estimates cluster in a tight range of −0.65 to −0.71 in the main specifications (Columns (1)–(5)), with Column (6) at −0.79 reflecting the deliberately restricted sample. This pattern suggests that the revelation and consistency-shock designs capture distinct, though potentially related, sources of credibility deterioration rather than the same underlying variation. Column (3) re-estimates the specification on a 1:1 nearest-neighbor propensity-score matched sample (928 Revealed firms matched to 928 never-Revealed controls); the propensity score is estimated from pre-shock characteristics (size, book-to-market, leverage, emissions intensity, analyst coverage, ESG ratings, pre-shock $\hat{\pi}$, and industry fixed effects). Column (5) reweights observations using inverse propensity weights from the same matching model. Column (6) additionally excludes events at firms that experienced a cross-venue consistency deterioration in the 12 months preceding the climate action, isolating revelation-shock variation from consistency-shock variation. The credibility measure $\hat{\pi}^{\text{pre}}$ is fixed at its last pre-shock value (frozen prior). All lower-order terms are included in estimation. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Analyst forecast revisions and the cash-flow channel. Announcement returns could reflect sentiment, attention, or discount-rate effects rather than genuine updating of cash-flow expectations. To distinguish these channels, we examine whether the same *ActionStrength* $\times \hat{\pi}_{i,e-1}$ signal that predicts higher CARs also predicts upward revisions in analyst EPS forecasts—an independent, non-price measure of expectation updating.

We merge the 4,812-event sample with I/B/E/S and construct the revision as the change in the consensus EPS forecast from the $[-30, -2]$ to $[+2, +30]$ day window around the announcement, scaled by share price at $t-5$ and winsorized at the 1% and 99% levels. We require at least two active analysts in both windows, yielding 3,684 events. We also construct a 90-day post-event window as a robustness check.

Before testing revisions, we confirm that the baseline CAR interaction is stable on this subsample (coefficient 1.13, compared to 1.17 in the full sample). We then estimate

$$\text{EPSRevision}_{i,e} = \beta_1 \mathit{ActionStrength}_{i,e} + \beta_2 \hat{\pi}_{i,e-1} + \beta_3 \big(\mathit{ActionStrength}_{i,e} \times \hat{\pi}_{i,e-1}\big) + \Gamma' X_{i,e-1} + \mu_i + \lambda_t + \varepsilon_{i,e}, \tag{37}$$

where $X_{i,e-1}$ includes the same controls as in the baseline CAR specification, and μ_i and λ_t denote firm and year fixed effects, respectively. The equation as written corresponds to Columns (2)–(4); in Column (1), μ_i and λ_t are replaced by industry×year fixed effects to match the baseline CAR specification in Table 15 Column (3).

Table OA-17 reports the results. The interaction coefficient is positive and statistically significant ($\beta_3 = 0.042$, SE 0.019) in the 30-day window, indicating that analysts revise earnings expectations more favorably when a strong climate action is announced by a firm with higher prior inferred credibility. Evaluated at the

sample median *ActionStrength* (0.53) and a one-standard-deviation increase in $\hat{\pi}_{i,e-1}$ (0.20), the interaction implies an EPS revision of approximately $0.042 \times 0.53 \times 0.20 \approx 0.0045$ of share price (i.e., approximately 0.45% of price). This test provides cash-flow-channel evidence that analysts revise expectations in the same direction as the announcement return. We do not use the analyst revision magnitude to explain the long-horizon reversal, which is a separate return-dynamics result. The estimates are directionally similar, though somewhat smaller, in the 90-day window. As a complementary test, we examine ex-post forecast errors (realized next-quarter EPS minus pre-event consensus, scaled by price) and find a positive interaction (0.028, SE 0.017, $p < 0.10$). The interaction in analyst forecast revisions ($\hat{\beta}_3 = 0.042$, SE 0.019, significant at the 5% level) is independent evidence that the credibility signal is associated with revisions in expected cash flows, not only with short-window price reactions. This makes pure sentiment or attention alternatives less tenable.

8.4 Post-Revelation Return Reversals

If investors initially over-infer managerial credibility from climate communication and later revise these beliefs after credible contradictions, the credibility-pricing interaction should partially reverse over time, especially for firms later revealed as greenwashers. We test this prediction using post-announcement abnormal returns over one-, three-, six-, and 12-month horizons:

$$\begin{aligned} R_{i,e+h} = \alpha &+ \beta_1 \big(\textit{ActionStrength}_{i,e} \times \hat{\pi}_{i,e-1} \big) \\ &+ \beta_2 \big(\textit{ActionStrength}_{i,e} \times \hat{\pi}_{i,e-1} \times \textit{Revealed}_i \big) + \underbrace{\cdots}_{\text{lower-order terms}} + \gamma \mathbf{X}_{i,e} + \mu_i + \lambda_t + \varepsilon_{i,e+h}, \end{aligned} \tag{38}$$

where $R_{i,e+h}$ is the cumulative abnormal return over horizon $h \in \{1, 3, 6, 12\}$ months, and $\textit{Revealed}_i$ equals one for firms that experience a credible greenwashing revelation *after the event date* e. This is a stricter version of the $\textit{Revealed}_i$ partition used in Equation (35) (any-time revelation): in the reversal test we restrict to revelations that postdate the event, since only post-event revelations can falsify pre-event credibility priors. The coefficient β_2 captures whether the credibility-pricing interaction reverses more strongly among subsequently revealed firms; the saturated specification includes all required lower-order terms. Because $\textit{Revealed}_i$ is defined using future information, this is an ex-post classification test rather than an investor-information test. It is valid for detecting latent misclassification if revealed firms were already misaligned at the announcement date and if revelation status is not caused by prior stock performance; pre-announcement abnormal returns and credibility priors do not predict eventual revelation status ($p > 0.20$ in both cases).

Table 12 reports negative post-event returns on the credibility-pricing interaction at all horizons, with stronger reversals among firms later revealed as greenwashers. At six months, the differential reversal is -0.72 (SE 0.24, $p < 0.01$). Restricting the 12-month test to events with a full post-event return window yields similar estimates (-0.37^{***} and -0.68^{***}), so the result is not driven by late-sample truncation. These event-level CAR tests complement, but are not directly comparable to, the monthly return-predictability tests in Table 6: here the clock starts at the climate-action announcement, whereas the asset-pricing panel uses rolling monthly signals.

Table 12: Long-Horizon Reversal of the Composite Signal

Horizon	1 month	3 months	6 months	12 months
Composite signal ($AS \times \hat{\pi}_{i,e-1}$)	−0.12	−0.28*	−0.41**	−0.35***
	(0.08)	(0.15)	(0.18)	(0.11)
Revealed× Composite signal	−0.25**	−0.48**	−0.72***	−0.65***
	(0.12)	(0.21)	(0.24)	(0.19)
Observations	4,812	4,812	4,812	4,812

Notes. This table reports event-level panel regressions of cumulative abnormal returns at different post-announcement horizons on the composite signal ($ActionStrength \times \hat{\pi}_{i,e-1}$) and its interaction with the $Revealed_i$ indicator. The sample is the pooled 4,812 climate-action events; the regressions are cross-sectional across events rather than monthly Fama–MacBeth panels. $Revealed_i$ equals one for firms that experience a credible greenwashing revelation after the event date. For the 12-month horizon, late-sample events (after December 2022) use the maximum available return window rather than being dropped.
As a robustness check, we re-estimate the 12-month specification restricting to events with a full 12-month post-event return window (events through December 2022). The base coefficient becomes -0.37^{***} (SE 0.12) and the interaction coefficient becomes -0.68^{***} (SE 0.20), with $N = 4{,}312$, which are similar in sign and magnitude to the main estimates. This indicates that the reversal pattern is not driven by late-sample truncation.
Standard errors are Newey–West adjusted with 12 lags to account for overlapping post-event return windows across the four horizon columns; this is a deliberate departure from the paper's default of double-clustering by firm and year, which is appropriate for non-overlapping CAR windows but understates dependence in the overlapping long-horizon setting used here. Results are similar (within one standard error) under double-clustering at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Distance to greenwashing revelation. Mispricing intensity declines monotonically with proximity to a firm's eventual greenwashing revelation, consistent with the credibility-pricing channel attenuating as belief revisions become inevitable. Online Appendix OA-2 develops this design in detail (Section OA-2.5, Table OA-4, and Figure OA-2).

9 Robustness

We address potential concerns regarding sample selection, omitted heterogeneity, proxy construction, and the specificity of the credibility signal to climate-related communication.

Selection into the event sample. Firms that issue climate actions and have sufficient prior climate disclosures may differ systematically from the broader population. We estimate a probit model predicting inclusion in the event sample using firm size, book-to-market, leverage, emissions intensity, analyst coverage, and industry fixed effects, and then re-estimate the main specification using inverse probability weights. The IPW interaction coefficient is 1.23 with firm and year fixed effects (Table OA-18) and 1.19*** (SE 0.34) with the preferred industry-by-year fixed effects (Table OA-11), close to the unweighted main estimate of 1.17. IPW addresses selection on the observable firm characteristics included in the probit; selection on unobservables (e.g., latent communication propensity, time-varying investor attention to climate topics) is not addressed by this procedure.

Sensitivity to controls and fixed effects. The raw cross-sectional specification yields 1.28 (Table 15, Column (1)); adding controls and year fixed effects gives 1.21; and the preferred specification with industry-by-year fixed effects gives 1.17. Across the main specifications, the interaction ranges from 1.13 to 1.28.

Specificity to climate disclosures. We run two complementary placebo tests. The *low-disclosure* placebo restricts the sample to firms with limited prior climate disclosure; the interaction collapses to near zero (Table OA-19, Column (2)). Because that restriction mechanically reduces text volume, the cleaner test is the *non-climate-disclosure* placebo, which replaces prior climate communication with non-ESG corporate disclosures from the same firms while preserving text volume. In that specification, the interaction falls to 0.12—an order of magnitude below the 1.17 baseline (Table 25, Column (3))—confirming specificity to climate-credibility inference rather than general communication quality.

Window overlap. The market-model estimation window used to compute abnormal returns partly overlaps with the trailing disclosure window used to construct $\hat{\pi}_{i,e-1}$. Recomputing abnormal returns using reduced-overlap estimation windows ending 63 and 126 trading days before the event yields interaction coefficients of 1.14 and 1.11 (Table OA-12), indicating that mechanical overlap does not materially affect the results.

Additional controls and announcement clustering. The interaction remains significant after adding environmental-performance controls (emissions intensity, ESG scores, green capital expenditure) and governance controls (institutional ownership, board independence, CEO equity incentives) (Table OA-9). Excluding events preceded by another climate announcement from the same firm within the prior six months yields 1.14*** (Table OA-7), ruling out contamination from announcement clustering within disclosure programs.

Summary. The credibility-pricing channel is robust to IPW reweighting on observables (1.19 vs. 1.17), fixed-effect structure (1.13–1.28), reduced-overlap CAR windows (1.11–1.14), governance and environmental controls, rolling-classifier construction (1.12–1.13, below), and placebo signals. The order-of-magnitude collapse in the non-climate-disclosure placebo (0.12 vs. 1.17) is hard to square with generic disclosure quality as the primary driver.

9.1 Robustness to Classifier Look-Ahead and Proxy Construction

Because the supervised classifier used to construct $\hat{\pi}_{i,t}$ is trained on the full 2005–2023 corpus, its parameters are not strictly predetermined for earlier firm-years. We address this with a rolling estimation procedure: for each year t, the classifier is re-estimated using only labeled paragraphs from documents dated before t and then applied out-of-sample. We also construct $\mathit{CredibilityShock}_{i,e}$ as the AR(1) innovation in $\hat{\pi}_{i,e-1}$, isolating revisions in perceived credibility and removing persistent components that may reflect slow-moving firm characteristics, following shock-based constructions in textual asset pricing (Hirshleifer et al., 2025b,a).

Table OA-15 shows that the interaction remains close to the 1.17 baseline under both the rolling classifier (1.12) and the CredibilityShock measure (1.13). The remaining concern—selection of labeled training paragraphs—is narrow because the underlying linguistic features (specificity, cost acknowledgment, and forward-looking orientation) are stable over the sample period.

Identification limitations. Residual selection on unobservables, endogenous communication responses, and non-classical measurement error in the bounded proxy $\hat{\pi} \in [0, 1]$ cannot be fully eliminated. These concerns are mitigated by observable-quality controls (Table OA-2), within-firm designs (Table 26), disclosure-standardization attenuation (Table 18), frozen-prior revelation tests (Table 11), and placebo tests in which the interaction collapses when climate-specific communication is removed (Table OA-19). The estimates

should therefore be read as reduced-form evidence that the pricing channel reflects credibility-based inference rather than a structural estimate of managerial type.

Measurement scope: firm-level intent versus value-chain relocation. Our framework measures investor beliefs about firm-level managerial intent using the focal firm's own disclosures and announced actions. Value-chain reallocation, as documented by Duchin et al. (2025) when firms divest pollutive plants to supply-chain partners without reducing aggregate pollution, is a distinct phenomenon studied with different data. In our investor-side measurement, such a divestiture by a firm with implementation-oriented prior communication may register at announcement as a credible strong action and only later be repriced as ex-post information arrives—which is precisely the misclassification-and-correction dynamic the paper documents. The firm-side evidence in Duchin et al. (2025) and our investor-side evidence are complementary.

10 Conclusion

This paper studies how investors price corporate climate actions when managerial credibility is uncertain. The central finding is that market reactions depend not only on the actions firms announce, but also on the credibility investors infer from prior climate-related communication. Observationally similar actions generate larger announcement returns when prior communication is more specific, implementation-oriented, and costly to mimic. The effect survives controls for observable environmental quality and is present within firms, indicating that investors use communication histories as credibility signals when interpreting subsequent climate actions.

We formalize this mechanism in a Bayesian framework in which investors form credibility priors from prior communication and update those beliefs after observing actions. We show that announcement returns reflect belief revision scaled by a credibility-dependent valuation component, and we derive testable predictions based on action cost, cross-venue consistency, and salience. The empirical evidence supports these predictions: credibility matters more for actions that are harder to mimic, for firms with more consistent communication across disclosure venues, and in periods when climate information is more salient. These patterns distinguish the credibility-inference channel from explanations based solely on observable environmental quality or exogenous signal precision.

The evidence comes from several complementary research designs. Within-firm changes in inferred credibility predict changes in announcement returns, and placebo tests weaken alternative interpretations based on reverse causality or general disclosure quality. Around credible greenwashing revelations, the credibility-pricing relation declines once contradictory information becomes public, and is empirically larger in the months preceding revelation events; the prospective-gradient interpretation is supported by matched-sample and inverse-probability-weighted specifications, but the timing of revelations remains correlated with media and NGO attention cycles and so is not fully exogenous. Following disclosure standardization, the channel is also smaller, though we cannot fully separate genuine compression of the credibility-relevant variation in disclosures from mechanical compression of the proxy. Taken together, these patterns are consistent with investor credibility inference as a meaningful component of climate-action pricing.

The return-predictability evidence is observational corroboration. The joint signal of strong actions and high inferred credibility is associated with cross-sectional return predictability that subsequently partially reverses, consistent with belief correction; the long-horizon reversal attenuates under climate-factor controls (Section 7). We treat the asset-pricing tests as corroborating; the announcement-window evidence and the

quasi-experimental designs in Section 8 provide the primary support for the credibility-pricing channel.

The findings also have implications for disclosure regulation. Mandatory standardization is typically justified by comparability, but our results highlight an additional channel: by compressing narrative discretion, standardization reduces investors' reliance on noisy communication-style signals and limits the pricing consequences of mistaken credibility attribution. This highlights a trade-off: voluntary regimes preserve cross-firm communication variation but leave greater scope for strategic image management, whereas standardized regimes improve comparability and reduce the scope for credibility-based misclassification.

More broadly, the paper highlights the belief-formation step that precedes pricing. In climate markets, investors must infer not only the likely cash-flow implications of an action, but also the credibility of the manager undertaking it. An open question is whether a similar credibility-inference channel operates in other settings where managerial intent is difficult to verify in real time, including merger announcements, R&D guidance, and other forms of voluntary corporate disclosure.

References

Akey, P. and Appel, I. (2021). The limits of limited liability: Evidence from industrial pollution. *Journal of Finance*, 76(1):5–55.

Avramov, D., Cheng, S., Lioui, A., and Tarelli, A. (2022). Sustainable investing with ESG rating uncertainty. *Journal of Financial Economics*, 145(2):642–664.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3):329–349.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343.

Berg, F., Kölbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344.

Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N. (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

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

Bozanic, Z., Roulstone, D. T., and Van Buskirk, A. (2018). Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, 65(1):1–20.

Brown, S. J. and Warner, J. B. (1985). Using daily stock returns: The case of event studies. *Journal of Financial Economics*, 14(1):3–31.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869.

Christensen, H. B., Hail, L., and Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies*, 26(3):1176–1248.

Cohen, L. H., Gurun, U. G., and Nguyen, Q. H. (2026). The ESG–innovation disconnect: Evidence from green patenting. *Management Science*. Published online January 5, 2026.

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., and Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6):1222–1231.

Delmas, M. A. and Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, 54(1):64–87.

Duchin, R., Gao, J., and Xu, Q. (2025). Sustainability or greenwashing? Evidence from the asset market for industrial pollution. *Journal of Finance*, 80(2):699–754.

Edmans, A. (2011). Does the stock market fully value intangibles? employee satisfaction and equity prices. *Journal of Financial Economics*, 101(3):621–640.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebel, J. (2020). Hedging climate change news. *Review of Financial Studies*, 33(3):1184–1216.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., and Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, 123(3):434–441.

Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, 142(2):499–516.

Ghitti, M., Gianfrate, G., and Palma, L. (2024). The agency of greenwashing. *Journal of Management & Governance*, 28(3):905–941.

Gibson Brandon, R., Krueger, P., and Schmidt, P. S. (2021). ESG rating disagreement and stock returns. *Financial Analysts Journal*, 77(4):104–127.

Giglio, S., Maggiori, M., Stroebel, J., Tan, Z., Utkus, S., and Xu, X. (2025). Four facts about ESG beliefs and investor portfolios. *Journal of Financial Economics*, 164(1):103984.

Goldstein, I., Kopytov, A., Shen, L., and Xiang, H. (2022). On ESG investing: Heterogeneous preferences, information, and asset prices. NBER Working Paper 29839, National Bureau of Economic Research.

Gourier, E. and Mathurin, H. (2025). A greenwashing index. Working paper, ESSEC Business School; SSRN version last revised August 22, 2025.

Grewal, J., Richardson, G., and Wang, J. (2026). Effects of mandatory carbon reporting on greenwashing. *Accounting Review*. Early access.

Han, Y., Cui, X., Tian, G. Y., and Wang, P. (2023). Bayesian investor belief updating speed and market underreaction to earnings announcements. *Australian Accounting Review*, 33(1):66–85.

Harrison, J. M. and Kreps, D. M. (1978). Speculative investor behavior in a stock market with heterogeneous expectations. *Quarterly Journal of Economics*, 92(2):323–336.

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

Heinkel, R., Kraus, A., and Zechner, J. (2001). The effect of green investment on corporate behavior. *Journal of Financial and Quantitative Analysis*, 36(4):431–449.

Hirshleifer, D., Mai, D., and Pukthuanthong, K. (2025a). War discourse and disaster premium: 160 years of evidence from the stock market. *Review of Financial Studies*, 38(2):457–506.

Hirshleifer, D., Mai, D., and Pukthuanthong, K. (2025b). War discourse and the cross section of expected stock returns. *Journal of Finance*, 80(6):3589–3637.

Hirshleifer, D. and Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1–3):337–386.

Ilhan, E., Sautner, Z., and Vilkov, G. (2021). Carbon tail risk. *Review of Financial Studies*, 34(3):1540–1571.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8):589–604.

Ji, Y., Xu, W., Zhao, Q., and Jia, Z. (2023). ESG disclosure and investor welfare under asymmetric information and imperfect competition. *Pacific-Basin Finance Journal*, 78:101982.

Kathan, M. C., Utz, S., Dorfleitner, G., Eckberg, J., and Chmel, L. (2025). What you see is not what you get: ESG scores and greenwashing risk. *Finance Research Letters*, 74:106710.

Krueger, P., Sautner, Z., and Starks, L. T. (2020). The importance of climate risks for institutional investors. *Review of Financial Studies*, 33(3):1067–1111.

Lyon, T. P. and Montgomery, A. W. (2015). The means and end of greenwash. *Organization & Environment*, 28(2):223–249.

Matsumura, E. M., Prakash, R., and Vera-Muñoz, S. C. (2014). Firm-value effects of carbon emissions and carbon disclosures. *Accounting Review*, 89(2):695–724.

Meeuwis, M., Parker, J. A., Schoar, A., and Simester, D. (2022). Belief disagreement and portfolio choice. *Journal of Finance*, 77(6):3191–3247.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Pástor, Ľ., Stambaugh, R. F., and Taylor, L. A. (2021). Sustainable investing in equilibrium. *Journal of Financial Economics*, 142(2):550–571.

Pástor, Ľ., Stambaugh, R. F., and Taylor, L. A. (2022). Dissecting green returns. *Journal of Financial Economics*, 146(2):403–424.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, 22(1):435–480.

Raghunandan, A. and Rajgopal, S. (2022). Do ESG funds make stakeholder-friendly investments? *Review of Accounting Studies*, 27(3):822–863.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, 9(4):174–179.

Saxe, R. and Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "Theory of Mind". *NeuroImage*, 19(4):1835–1842.

Seele, P. and Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies. *Business Strategy and the Environment*, 26(2):239–252.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Starks, L. T. (2023). Presidential address: Sustainable finance and ESG issues—Value versus values. *Journal of Finance*, 78(4):1837–1872.

Teti, E., Etro, L. L., and Pausini, L. (2024). Does greenwashing affect company's stock price? Evidence from Europe. *International Review of Financial Analysis*, 93:103195.

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

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3):1437–1467.

Torelli, R., Balluchi, F., and Lazzini, A. (2020). Greenwashing and environmental communication: Effects on stakeholders' perceptions. *Business Strategy and the Environment*, 29(2):407–421.

Zhou, X. and Kang, J. (2023). Searching for ESG information: Heterogeneous preferences and information acquisition. *Journal of Economic Dynamics and Control*, 153:104693.

A Tables

Table 13: Summary Statistics

Variable	Mean	SD	P25	Median	P75	Obs.
Panel A: Primary Climate-Action Events ($N = 4{,}812$)						
CAR $[-1, +1]$ (%)	0.42	3.18	−1.25	0.38	2.05	4,812
ActionStrength	0.53	0.29	0.31	0.52	0.74	4,812
Credibility prior $\hat{\pi}_{i,e-1}$ (event sample)	0.47	0.20	0.28	0.46	0.64	4,812
Credibility prior $\hat{\pi}_{i,t}$ (full panel)	0.47	0.21	0.29	0.46	0.64	19,804
Log assets	9.18	1.54	8.07	9.05	10.23	19,804
Book-to-market	0.64	0.42	0.38	0.55	0.80	19,804
Leverage	0.29	0.18	0.14	0.27	0.41	19,804
Emissions intensity	0.82	0.75	0.24	0.58	1.11	19,804
Analyst coverage	8.6	6.1	3.0	7.0	12.0	19,804
Panel B: Follow-Through Implementation Events ($N = 10{,}808$)						
CAR $[-1, +1]$ (%)	0.31	2.87	−1.18	0.29	1.82	10,808
ActionStrength	0.48	0.26	0.27	0.49	0.69	10,808
Credibility prior $\hat{\pi}_{i,e-1}$	0.46	0.21	0.27	0.45	0.65	10,808
Action type shares (fraction of events)						
Emissions reductions	0.38	–	–	–	–	10,808
Green capex / restructuring	0.29	–	–	–	–	10,808
Target progress reports	0.22	–	–	–	–	10,808
Other / mixed	0.11	–	–	–	–	10,808
Follow-through events per firm (mean / median)	3.4 / 2.0		–	–	–	–

Panel A reports summary statistics for the primary sample of climate-related corporate action announcements (event level). Panel B reports statistics for the supplementary follow-through implementation events used in the H2 tests (Section 6). CAR is the three-day cumulative abnormal return over the $[-1, +1]$ window. *Variable units. ActionStrength* is a 0–1 normalized score from a pre-registered RA coding rubric (higher values denote actions with greater scope, cost, and verifiability). The credibility prior $\hat{\pi}$ is the model-implied posterior probability that the firm is the sincere type, taking values in $[0, 1]$. The credibility prior proxy is shown separately for the event-study sample (4,812 observations; $\hat{\pi}_{i,e-1}$, SD = 0.20) and the full firm-year panel (19,804 observations; $\hat{\pi}_{i,t}$, SD = 0.21). Throughout the paper, economic-magnitude statements involving announcement returns use the *event-sample* standard deviation of 0.20 (e.g., $1.17 \times 0.20 \approx 0.23$ pp), since $\hat{\pi}_{i,e-1}$ is the relevant proxy in those regressions; the full-panel SD of 0.21 is used where panel-level regressions (validation, belief-updating) are discussed. The full-panel SD of 0.21 yields approximately 0.25 pp as a reference, a difference of approximately 8%. The P25 of $\hat{\pi}_{i,e-1}$ in the event-study sample is 0.28 (reported above). In auxiliary tabulations (available in the online appendix), the 10th and 90th percentiles of $\hat{\pi}_{i,e-1}$ in the event-study sample are approximately 0.21 and 0.68, respectively; the numerical coincidence between the 10th percentile (0.21) and the full-*panel* SD of 0.21 (a different object from the event-sample SD of 0.20) is a cross-object artifact of the sample rather than a structural property of the bounded support.

Action types in Panel B are mutually exclusive after manual classification. The per-firm distribution is computed over the 3,178 unique firms with at least one follow-through event; the 90th percentile is 8 events per firm, indicating that the sample is not dominated by a small number of firms. The follow-through sample contains more unique firms (3,178) than the primary event sample (1,856) because the follow-through identification procedure scans *all* firms in the broader NLP corpus (approximately 5,200 firms) for referential implementation language, not only the 1,856 firms that had an identifiable primary climate-action event meeting the event-study inclusion criteria. A firm may have a follow-through disclosure referencing a prior commitment (e.g., "delivering on our 2019 net-zero pledge") without the original pledge satisfying the event-study filters (minimum prior disclosure window, stock-return data availability). These firms are excluded from the primary event sample but remain eligible for the follow-through sample. The 1,322 additional firms are similar to the primary sample on observable dimensions (size, leverage, emissions intensity); full summary statistics for Panel B are available in the online appendix. The follow-through regressions (Table 16) pool primary and follow-through events; results are robust to restricting follow-through events to firms that also appear in the primary event sample. Firm characteristics (log assets, book-to-market, leverage, emissions intensity, analyst coverage) are not re-reported in Panel B as they are identical when matched to the same firm. The sample consists of U.S. publicly traded firms from 2005 to 2023. Log assets is the natural logarithm of total book assets (Compustat AT). Emissions intensity is carbon emissions scaled by firm sales. Analyst coverage is the number of analysts issuing earnings forecasts in I/B/E/S.

Table 14: Validation of the Credibility Measure

Dependent variable	(1) Emissions reduction	(2) Green capex	(3) Climate targets
Credibility prior $\hat{\pi}_{i,t}$	0.082***	0.114***	0.091***
	(0.019)	(0.024)	(0.021)
Lagged emissions intensity	−0.047**	−0.031*	−0.039**
	(0.021)	(0.018)	(0.019)
Log assets	0.014	0.036*	0.022
	(0.012)	(0.020)	(0.017)
Leverage	−0.021	−0.008	−0.017
	(0.018)	(0.025)	(0.021)
Book-to-market	−0.011	0.005	−0.009
	(0.010)	(0.014)	(0.013)
Firm FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	18,912	18,912	18,912
R^2	0.29	0.31	0.27

This table provides validation of the credibility prior measure by examining whether higher inferred credibility predicts subsequent real environmental actions at $t+1$. The dependent variables are binary indicators of climate follow-through: emissions reduction, green capital expenditure, and adoption of a verifiable climate target (each coded as one if the firm exhibits the relevant outcome in year $t+1$). Coefficients are therefore percentage-point changes in the probability of the relevant outcome; the body text reports a 1.7 pp probability change associated with a one-SD increase in $\hat{\pi}_{i,t}$. Continuous emissions-intensity outcomes are reported separately in Table 23. All control variables are measured at time t, with emissions intensity entering in lagged form. The key explanatory variable is $\hat{\pi}_{i,t}$, the estimated posterior belief that the firm is sincere. The timing ensures that the explanatory variables precede the outcome realization, mitigating concerns that contemporaneous outcomes drive the credibility measure. The positive and statistically significant coefficients on $\hat{\pi}_{i,t}$ across all outcomes indicate that firms perceived as more credible subsequently undertake greater real environmental actions. As discussed in Section 5.5, predictive power for real outcomes does not distinguish between the intention channel and persistent firm capability or disclosure sophistication; we interpret $\hat{\pi}_{i,t}$ as a composite credibility signal, not as a clean measure of intent orthogonal to firm quality. These results validate the economic relevance of the measure but do not imply that it isolates managerial intent separately from correlated firm characteristics such as implementation capacity. They are consistent with a framework in which investor beliefs about managerial credibility reflect both persistent firm characteristics and updated information from prior communication. The sample of 18,912 firm-years is smaller than the full panel of 19,804 because we restrict to observations with non-missing values for all three outcome variables (emissions reduction, green capital expenditure, and climate target adoption). Standard errors are clustered at the firm level and reported in parentheses (firm-only rather than double-clustered; see inference policy in Section 6). ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Note: standard errors in this table are clustered at the firm level (not double-clustered) because the panel structure is at the firm-year level rather than the event level.

Table 15: Market Reaction to Climate Actions

Dependent variable: CAR $[-1,+1]$	(1) No controls	(2) Controls	(3) Preferred	(4) Future placebo	(5) Type × Strength	(6) Quintile interactions
ActionStrength	0.63***	0.58***	0.55***	0.53***	0.51***	0.55***
	(0.18)	(0.17)	(0.16)	(0.17)	(0.16)	(0.17)
Credibility prior $\hat{\pi}_{i,e-1}$	0.41*	0.36*	0.29	0.27	0.25	
	(0.22)	(0.21)	(0.19)	(0.19)	(0.18)	
ActionStrength $\times\hat{\pi}_{i,e-1}$	1.28***	1.21***	1.17***		1.13***	
	(0.34)	(0.32)	(0.30)		(0.29)	
ActionStrength $\times\hat{\pi}_{i,e+1}$				0.08		
				(0.29)		
ActionStrength × Q2($\hat{\pi}$)						0.51**
						(0.26)
ActionStrength × Q3($\hat{\pi}$)						0.94***
						(0.28)
ActionStrength × Q4($\hat{\pi}$)						1.38***
						(0.30)
ActionStrength × Q5($\hat{\pi}$)						1.91***
						(0.37)
Controls	No	Yes	Yes	Yes	Yes	Yes
Action-type indicators	No	Yes	Yes	Yes	Yes	Yes
Action-type × ActionStrength	No	No	No	No	Yes	Yes
Firm FE	No	No	No	No	No	No
Year FE	No	Yes	No	No	No	No
Industry × Year FE	No	No	Yes	Yes	Yes	Yes
Observations	4,812	4,812	4,812	4,812	4,812	4,812
R^2 / Within R^2	0.08	0.14	0.19	0.19	0.22	0.23

This table reports the baseline event-study regressions. The dependent variable is the cumulative abnormal return over the $[-1,+1]$ event window around a climate-related corporate action. *ActionStrength* measures the economic importance of the event. $\hat{\pi}_{i,e-1}$ is the pre-event credibility prior constructed from ESG communication in the 12 months preceding the event. Control variables include log total assets (firm size), book-to-market, leverage, prior returns, emissions intensity, and analyst coverage. Action-type indicators (included in Columns (2)–(6)) absorb mean return differences across action categories (net-zero pledges, green capex, science-based targets, etc.), ensuring the interaction is identified from within-type variation in credibility priors. Column (1) is a raw cross-sectional specification with no fixed effects or controls. Column (2) adds controls, action-type indicators, and year fixed effects. Column (3) is the preferred specification: it replaces year fixed effects with industry-by-year fixed effects. Column (4) is a placebo specification that replaces $\hat{\pi}_{i,e-1}$ with $\hat{\pi}_{i,e+1}$; the estimated coefficient (0.08, SE 0.29) is insignificant, ruling out reverse causality. Column (5) augments Column (3) with action-type × *ActionStrength* slope interactions; the coefficient on *ActionStrength* $\times\,\hat{\pi}_{i,e-1}$ remains 1.13*** (SE 0.29), ruling out composition effects. Column (6) replaces the continuous interaction with *ActionStrength* interacted with dummies for inferred-credibility quintiles (Q1 omitted as reference). These coefficients are incremental effects relative to the Q1 ActionStrength slope of 0.55; the implied total slopes are 0.55 (Q1), 0.55+0.51 = 1.06 (Q2), 0.55+0.94 = 1.49 (Q3), 0.55 + 1.38 = 1.93 (Q4), and 0.55 + 1.91 = 2.46 (Q5). The total slopes increase monotonically across quintiles, confirming that the pricing of credibility strengthens progressively across the distribution and is not driven by extreme observations. No specification includes firm fixed effects; the within-firm identification is reported separately in Table OA-2 Column (1) and Table 26. The interaction coefficient is stable across specifications (1.28, 1.21, 1.17, 1.13 — a 12% range). R^2 denotes the standard OLS R^2 in Column (1); "Within R^2" in Columns (2)–(6) denotes R^2 after partialling out fixed effects. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 16: Follow-Through Tests

Dependent variable: CAR $[-1,+1]$	(1) No controls	(2) With controls†
FollowThrough	0.72***	0.66***
	(0.21)	(0.20)
Credibility prior $\hat{\pi}_{i,e-1}$	0.38**	0.32*
	(0.19)	(0.18)
FollowThrough $\times\hat{\pi}_{i,e-1}$	1.45***	1.39***
	(0.37)	(0.35)
Firm FE	Yes	Yes
Year FE	Yes	Yes
Observations	15,620	15,620
Within R^2	0.21	0.24
Restricted to primary-sample firms (1,856 firms)		
FollowThrough $\times\hat{\pi}_{i,e-1}$	1.42***	1.37***
	(0.36)	(0.34)
Observations (restricted)	9,874	

Notes. The dependent variable is the cumulative abnormal return (CAR) over the $[-1,+1]$ window around each announcement event, including both initial and follow-through actions. Column (2) is the preferred specification, including firm controls (log total assets, book-to-market, leverage, prior returns, emissions intensity, and analyst coverage); it yields an interaction of 1.39 (SE 0.35). Column (1) presents a specification without additional firm-level controls; both columns include firm and year fixed effects. *FollowThrough* is a binary indicator for a concrete implementation action following an earlier climate commitment, such as a verifiable emissions reduction, green capital expenditure, or operational restructuring. The sample combines 4,812 primary climate-action events with 10,808 follow-through implementation events identified from explicit textual references to prior commitments and manual verification (see Section 4 and Table 13, Panel B for summary statistics). Primary events correspond to the initial announcement of a climate commitment or action, while follow-through events are subsequent disclosures that reference and report measurable progress on previously announced commitments. The two samples are mutually exclusive by construction, so each firm–event date enters the regression at most once. Because a single initial commitment can generate multiple verifiable updates over time, the number of follow-through events exceeds the number of primary announcements. The main specification pools primary and follow-through events so that $FollowThrough_{i,e}$ varies within the sample. Restricting to follow-through events only, the credibility interaction $\hat{\pi}_{i,e-1}$ remains positive and significant; restricting to primary events only, the $ActionStrength \times \hat{\pi}_{i,e-1}$ interaction remains consistent with the baseline. Note that within each subsample $FollowThrough_{i,e}$ is constant (zero or one), so these subsample tests examine the credibility interaction within each event type separately, not a FollowThrough interaction. All specifications include firm and year fixed effects. Standard errors are double-clustered at the firm and year levels, consistent with the primary event-study regressions. The bottom panel restricts the sample to firms that also appear in the primary event-study sample (1,856 firms). This restriction ensures that the firms in both panels are observably comparable; it does not require that each follow-through event's originating commitment be inside the primary event-study identification window, since a firm in the primary sample may have follow-through events linked to commitments that predate the event-study window. The interaction remains positive and significant at the 1% level under this restriction (1.42*** and 1.37***). ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 17: Follow-On Disclosure Response After Climate Actions (H3: actions trigger updates in the credibility proxy)

Dependent variable: $\hat{\pi}_{i,t+1}$	(1) Firm + year FE
Climate action $A_{i,t}$ (any climate action, binary)	0.064***
	(0.015)
Log assets	0.011
	(0.008)
Leverage	-0.019
	(0.014)
Book-to-market	-0.007
	(0.009)
Firm FE	Yes
Year FE	Yes
Observations	17,940
R^2	0.26

This table tests whether climate actions are associated with higher subsequent credibility scores, conditional on firm fixed effects, consistent with within-firm belief updating. The dependent variable is $\hat{\pi}_{i,t+1}$ constructed from ESG disclosures anchored to December 31 of year $t+1$ (Equation (29)). The sample of 17,940 firm-years is smaller than the full panel of 19,804 because constructing the lead credibility measure $\hat{\pi}_{i,t+1}$ requires dropping the final observation for each firm, as the last year has no subsequent period for the lead variable. The specification regresses the level of $\hat{\pi}_{i,t+1}$ on the action indicator with firm fixed effects, which absorb time-invariant firm differences. Because the calendar-year anchor creates variation in post-action measurement horizons (early-year events have up to 24 months of post-action disclosures in the dependent variable, while late-year events have fewer), we re-estimate using a fixed 12-month post-event window anchored to the event date. This fixed-window robustness (Table OA-13, Column (2)) yields a coefficient of 0.058*** (SE 0.014), confirming that the result is not driven by calendar-year measurement. Results are also robust to a first-differenced specification ($\Delta\hat{\pi}_{i,t+1}$ as dependent variable). The key explanatory variable is an indicator for a climate action at time t. Standard errors are clustered at the firm level (rather than double-clustered; see inference policy in Section 6). ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Note: standard errors clustered at the firm level; firm-year panel structure precludes event-level double clustering.

Table 18: Credibility-Pricing Channel Attenuates After Disclosure Standardization (EU CSRD, California SB 253/261)

Dependent variable: CAR [-1,+1]	(1) Main	(2) 2022 Placebo
ActionStrength	0.56***	0.57***
	(0.17)	(0.17)
Credibility prior $\hat{\pi}_{i,e-1}$	0.31*	0.32*
	(0.18)	(0.18)
ActionStrength $\times\hat{\pi}_{i,e-1}$	1.19***	1.22***
	(0.31)	(0.32)
$Post_t \times Treated_i \times ActionStrength \times \hat{\pi}_{i,e-1}$	−0.74**	0.04
	(0.33)	(0.35)
Controls	Yes	Yes
Firm FE	Yes	Yes
Year FE	Yes	Yes
Observations	4,812	4,812
R^2	0.22	0.22

This table presents the quasi-experimental identification test based on disclosure-regulation shocks. *Post* equals one after the relevant disclosure standardization regime takes effect. Column (1) uses $Post_t = 1$ for observations after the 2023 CSRD/California implementation dates. Column (2) implements a placebo test using the November 2022 European Parliament vote on CSRD, defining $Post_t = 1$ for observations after November 2022. The corresponding interaction is economically small and statistically insignificant (0.04, SE 0.35), indicating that the attenuation in the credibility-pricing interaction does not precede the 2023 implementation of standardized disclosure requirements. *Treated* identifies firms with above-median EU revenue exposure (from Compustat geographic segment data, which reports revenues by broad geographic regions including Europe; post-Brexit UK, Norway, and Switzerland are included in this measure, introducing some attenuation error in treatment assignment relative to a precise EU-member-country definition), making them exposed to CSRD supply-chain requirements; results are robust to an alternative treatment indicator based on U.S. state-level climate-disclosure mandates. Note that the CSRD entered into force in January 2023 but first reporting deadlines for large EU companies apply to FY2024 (reports due 2025); the supply-chain requirements on non-EU suppliers take effect even later. The treatment therefore captures *anticipatory* exposure to future compliance requirements, not operative legal obligations in 2023H2. The "early-implementation" framing in the text reflects this: the documented attenuation reflects market anticipation of disclosure standardization rather than firms already subject to binding CSRD requirements. Proposition 4's mechanism (text variance compression reducing $\sigma^2_{\eta,G}$ toward $\sigma^2_{\eta,S}$) characterizes the long-run steady-state effect of operative standardization. The documented early-implementation attenuation captures the *anticipatory* version of the same channel: investors who anticipate that future disclosures will be standardized begin discounting the narrative credibility signal in current disclosures before compression has mechanically occurred. Both the anticipatory and mechanical channels predict attenuation of $\hat{\beta}_3$; the distinction affects the speed of adjustment but not the direction of the effect or its policy interpretation. Both *Post* and *Treated* are predetermined as of the shock year. The coefficient on the four-way interaction captures whether disclosure standardization reduces the pricing role of inferred credibility. This specification uses firm and year fixed effects; the baseline interaction of 1.19 reported here therefore differs from the preferred estimate of 1.17 in Table 15 Column (3), which uses industry-by-year fixed effects. The four-way interaction implies approximately 62% attenuation relative to the full-sample base interaction of 1.19 (firm and year fixed effects specification). The preferred event-time estimate of 38–52% uses the pre-period treated-firm baseline: 38% compares the $k = -3$ level (1.27) to the $k = +1$ level (0.79); 52% compares the $k = -3$ level (1.27) to the $k = +3$ level (0.61). See Table 19 for the full dynamic treated–control differentials. For completeness, the regression includes the full set of lower-order constituent terms implied by the saturated interaction specification: $ActionStrength_{i,e}$, $\hat{\pi}_{i,e-1}$, and all two-way and three-way interactions among the four variables ($ActionStrength_{i,e}$, $\hat{\pi}_{i,e-1}$, $Post_t$, $Treated_i$). The main effect of $Treated_i$ is absorbed by firm fixed effects. The main effect of $Post_t$ is partially (but not fully) absorbed by year fixed effects, since $Post_t$ switches within calendar 2023; the unabsorbed component is included as an explicit regressor in the lower-order term set. All remaining lower-order interaction terms are included in the estimation but omitted from the displayed table for space reasons. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 19: Credibility-Pricing Interaction Across Event Time: Treated–Control Differential and Control-Group Levels

Event time k	-3	-2	0	$+1$	$+2$	$+3$
Treated–control differential						
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Treated_i$	1.27***	1.24***	1.22***	0.79**	0.68**	0.61**
	(0.45)	(0.44)	(0.40)	(0.31)	(0.29)	(0.28)
Control-group level (for reference)						
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$	−0.02	0.04	0.01	−0.03	0.02	−0.01
	(0.11)	(0.10)	(0.09)	(0.08)	(0.09)	(0.08)
Controls and fixed effects			Yes			
Pre-period equality test (H_0: $\delta_{-3} = \delta_{-2}$)			$p = 0.74$			
Pre-period levels test (H_0: $\delta_{-3} = \delta_{-2} = 0$)			$p < 0.001$			
Post-period equality test (H_0: $\delta_{+1} = \delta_{+2} = \delta_{+3}$)			$p = 0.58$			

Notes. This table reports event-time coefficients from a dynamic specification that interacts $ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Treated_i$ with event-time indicators $k \in \{-3, -2, 0, +1, +2, +3\}$, where $k = 0$ captures events in 2023H1 through the regulatory announcement dates (a transition bin); attenuation is first detected at $k = +1$. The coefficients δ_k capture the differential pricing of climate actions for treated firms relative to control firms at each event time. The pre-period coefficients (approximately 1.25) are significantly positive, reflecting higher baseline credibility pricing among treated firms prior to the disclosure shock; accordingly, the joint test of equality to zero is rejected ($p < 0.001$). The relevant identifying assumption is stability of the treated–control differential across pre-period bins rather than equality to zero; consistent with this, we fail to reject equality of the pre-period coefficients ($p = 0.74$), supporting parallel trends. The post-period equality test ($p = 0.58$) fails to reject equality across the three post-treatment coefficients ($k = +1, +2, +3$), indicating that the attenuation is stable across event time.

Pre-versus-post-shock attenuation. The reported tests bracket the attenuation question (within-period stability before and after the shock) but do not constitute a formal pre-versus-post Wald test on the dynamic event-time coefficients themselves. The pre-versus-post test is implemented separately via two pooled DiD specifications: Table 18 reports the full-sample four-way interaction (−0.74, SE 0.33, $p < 0.05$, attenuation ≈62% of the 1.19 baseline), and Table 21 reports the donut version that explicitly collapses the pre-shock bins $\{-3, -2\}$ against the post-shock bins $\{+1, +2, +3\}$ while excluding the $k = 0$ transition window (−0.56, SE 0.24, $p < 0.05$). The donut pooled DiD is the most direct test of pre-versus-post equality for the dynamic-specification coefficients reported in this table; the dynamic event-time figure should be read as a parallel-trends and stable-attenuation diagnostic rather than as the primary attenuation test.

All specifications include the full set of controls (log assets, book-to-market, leverage, prior returns, emissions intensity, analyst coverage), firm fixed effects, and year fixed effects. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Inference and sparse event-time bins. The coefficients δ_k are obtained from a single pooled regression estimated on the full sample ($N = 4{,}812$), rather than from separate regressions within each event-time bin. Each δ_k identifies the absolute treated-minus-control differential within bin k, not a contrast relative to an omitted reference period; the $k = -1$ observations are dropped from the sample (rather than serving as a reference category) because their three-day CAR windows risk contamination from early market anticipation of the regulatory shock. As a result, the variance of $\hat{\delta}_k$ depends on the full design matrix, fixed effects, and cluster structure, and does not scale mechanically with the inverse square root of the number of treated observations in bin k. Although each post-shock bin contains approximately 40–45 events, the reported standard errors (0.28–0.31) reflect full-sample residual variation and clustering rather than bin-specific estimation. Pre-shock bins ($k = -3, -2$) each span approximately 12 months (2021–2022); post-shock bins ($k = +1, +2, +3$) each span approximately 2 months (partitioning July–December 2023). In the pre-period ($k = -3, -2$), the control-group interaction is economically small and statistically insignificant (coefficient range [−0.04, 0.06]), implying that the treated-group levels are well approximated by the reported differentials. The elevated pre-period differential (≈1.25) reflects persistent cross-sectional differences in baseline reliance on narrative credibility signals between treated and control firms; the absence of pre-trends supports the validity of the research design.

Table 20: Credibility-Pricing Interaction Across Event Time: Donut DiD (Excluding $k = 0$ Bin)

Event time k	-3	-2	$+1$	$+2$	$+3$
Treated–control differential					
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Treated_i$	1.27***	1.24***	0.79**	0.68**	0.61**
	(0.45)	(0.44)	(0.31)	(0.29)	(0.28)
Control-group level (for reference)					
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$	-0.02	0.04	-0.03	0.02	-0.01
	(0.11)	(0.10)	(0.08)	(0.09)	(0.08)
Controls and fixed effects			Yes		
Pre-trends test (H_0: $\delta_{-3} = \delta_{-2}$)			$p = 0.74$		
Observations			4,412		

Notes. The dependent variable is CAR $[-1, +1]$. This table re-estimates the event-time specification from Table 19 after excluding all observations in the transition bin ($k = 0$), corresponding to 2023H1 events around the regulatory announcement dates. The resulting donut sample contains 4,412 observations (excluding 400 events from the full sample of 4,812). The 400 events in the $k = 0$ transition bin reflect a 2023H1 event-sample rate of approximately 67 climate-action events per month (400 events over 6 months), substantially above the full-sample average of approximately 21 events per month over 2005–2022. This 2023H1 spike is consistent with event clustering around the announcement and adoption of major disclosure-standardization regimes (EU CSRD entered into force in January 2023; California SB-253 was advancing through the legislature). Firms anticipating standardized future disclosure rules frequently announce climate commitments before the cross-firm narrative variation that supports those commitments is compressed by mandatory templates. The post-shock bins ($k = +1, +2, +3$, 2023H2) revert to roughly 20–22 events per month, comparable to the historical average; we treat the $k = 0$ rate as elevated by anticipatory announcement clustering rather than by sample-restriction differences across calendar 2023, and exclude this bin entirely from the donut specification. The specification compares pre-shock bins $k \in \{-3, -2\}$ to post-shock bins $k \in \{+1, +2, +3\}$ without using the transition window. Pre-trends remain statistically indistinguishable, and the attenuation in the $ActionStrength \times \hat{\pi}$ interaction for treated firms persists in the post-shock period. All specifications include the same controls and fixed effects as Table 19, with standard errors double-clustered by firm and year. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

Table 21: Donut DiD: Pooled Post-Shock Effect (Excluding $k = 0$ Bin)

	(1)
Treated × Post × Credibility interaction	
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Treated_i \times Post$	-0.56^{**}
	(0.24)
Lower-order terms	
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Treated_i$	1.25^{***}
	(0.44)
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1} \times Post$	-0.03
	(0.09)
$ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$	0.78^{***}
	(0.11)
Controls and fixed effects	Yes
Observations	4,412
Within R^2	0.19

Notes. The dependent variable is CAR $[-1, +1]$. This table reports a pooled difference-in-differences specification that collapses post-shock bins $k \in \{+1, +2, +3\}$ into a single indicator ($Post$) and compares them to pre-shock bins $k \in \{-3, -2\}$, excluding the transition bin ($k = 0$). The coefficient of interest captures the change in the pricing of the $ActionStrength \times \hat{\pi}$ interaction for treated firms after disclosure standardization (-0.56^{**}, SE 0.24). The coefficient on $ActionStrength \times \hat{\pi}_{i,e-1}$ implies an untreated pre-period baseline of 0.78^{***}. The treated pre-period interaction is therefore 2.03 ($= 0.78 + 1.25$), indicating that the credibility-pricing channel is present in both groups but stronger among treated (EU-exposed) firms prior to the shock. The negative DiD coefficient (-0.56^{**}) reflects a relative attenuation of this interaction for treated firms in the post period. All lower-order interaction terms are included in estimation. The sample excludes 2023H1 events (400 observations) to eliminate potential anticipation or announcement-window contamination. Specifications include the same controls and fixed effects as Table 19, with standard errors double-clustered by firm and year. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

Table 22: Heterogeneity in the Credibility-Pricing Channel

	Full	Carbon intensity		Analyst coverage		Earnings uncertainty	
	sample	High (hard)	Low (easy)	Low (hard)	High (easy)	High (hard)	Low (easy)
ActionStrength $\times\hat{\pi}_{i,e-1}$	1.17***	1.68***	0.92***	1.54***	1.03***	1.72***	0.89***
	(0.30)	(0.41)	(0.28)	(0.39)	(0.27)	(0.44)	(0.26)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,812	1,920	2,892	1,606	3,206	1,488	3,324
Within R^2	0.19	0.24	0.18	0.23	0.19	0.25	0.18

Notes. The dependent variable is CAR $[-1,+1]$. This table reports the coefficient on $ActionStrength_{i,e} \times \hat{\pi}_{i,e-1}$ from the preferred specification (industry×year fixed effects with controls). The full-sample coefficient is 1.17***. Each pair of columns splits the full sample along the relevant dimension: high vs. low carbon intensity (industry-based classification), low vs. high analyst coverage, and high vs. low earnings uncertainty. Within each pair, the subsamples are mutually exclusive and collectively exhaustive (each pair sums to 4,812 events). The interaction is larger in settings where intent inference is more difficult (carbon-intensive: 1.68***, low analyst coverage: 1.54***, high uncertainty: 1.72***); in the complementary subsamples the coefficients remain positive and significant but smaller in magnitude (0.92***, 1.03***, 0.89***, respectively), indicating the channel operates across all settings but is amplified where credibility assessment is hardest.
Carbon-intensive industries are defined using SIC codes 10–14 (mining), 26 (paper), 28 (chemicals), 29 (petroleum), 33 (primary metals), and 49 (utilities). Analyst coverage and earnings uncertainty cutoffs are determined at the firm-year median and then applied to events; because more climate-announcement events come from firms with higher analyst coverage and lower earnings uncertainty, event-level subsample sizes are unbalanced (low-coverage: 1,606; high-coverage: 3,206; high-uncertainty: 1,488; low-uncertainty: 3,324). Low analyst coverage is below the firm-year median I/B/E/S analyst count in the prior year; high uncertainty is above the firm-year median dispersion in analyst earnings forecasts (standard deviation scaled by absolute mean forecast). All specifications include the baseline controls (log assets, book-to-market, leverage, prior returns, emissions intensity, analyst coverage). Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 23: Credibility Measures Predict Future Emissions Intensity and Green Innovation

Dependent variable	(1) Emissions intensity$_{t+1}$	(2) Green patents$_{t+1}$
Credibility prior $\hat{\pi}_{i,t}$	-0.094***	0.072***
	(0.021)	(0.018)
Log assets	-0.012	0.033*
	(0.010)	(0.017)
Leverage	0.026	-0.014
	(0.016)	(0.013)
Firm FE	Yes	Yes
Year FE	Yes	Yes
Observations	18,200	18,200
R^2	0.31	0.27

Dependent-variable units: Column (1) emissions intensity is scope-1 CO_2e per unit sales, log-transformed. Column (2) green patents is the natural log of 1+ (count of green-classified patents applied for in year $t+1$), using the Haščič–Migotto green-patent classification applied to USPTO patent data.
Notes. This table examines whether inferred credibility predicts subsequent real environmental outcomes. The independent variable is $\hat{\pi}_{i,t}$ and all dependent variables are measured in year t+1. Column (1) uses emissions intensity as the dependent variable; Column (2) uses green patenting activity. Higher credibility priors are associated with lower future emissions intensity and greater green innovation. Standard errors are clustered at the firm level (rather than double-clustered; see inference policy in Section 6). ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 24: Salience-Based Amplification of the Credibility-Pricing Channel

Dependent variable: CAR [-1,+1]	(1) No controls	(2) With controls
ActionStrength	0.51***	0.47***
	(0.17)	(0.16)
$\hat{\pi}_{i,e-1}$	0.33*	0.29*
	(0.19)	(0.18)
ActionStrength $\times$ $\hat{\pi}_{i,e-1}$	1.08***	1.02***
	(0.29)	(0.27)
$Attention_t$	0.21	0.17
	(0.15)	(0.14)
ActionStrength $\times$ $Attention_t$	0.26	0.22
	(0.20)	(0.19)
$\hat{\pi}_{i,e-1} \times Attention_t$	0.17	0.14
	(0.18)	(0.17)
ActionStrength $\times$ $\hat{\pi}_{i,e-1} \times Attention_t$	0.59**	0.54**
	(0.29)	(0.27)
Controls	No	Yes
Firm FE	Yes	Yes
Year FE	Yes	Yes
Observations	4,812	4,812
Within R^2	0.21	0.24

The dependent variable is the cumulative abnormal return over the $[-1, +1]$ event window. $Attention_t$ is a monthly index of climate-related keyword frequency in major financial news outlets and earnings call transcripts, using the climate-keyword dictionary of Sautner et al. (2023) aggregated to monthly market-wide frequency (rather than the firm-quarter frequency at which their own measure is constructed). Column (1) includes no additional controls beyond fixed effects. Column (2) adds firm-level controls (log assets, book-to-market, leverage, prior returns, emissions intensity, and analyst coverage). All specifications use firm and year fixed effects rather than the industry-by-year fixed effects of the preferred baseline (Table 15 Column (3)). This departure is necessary because $Attention_t$ is measured at the monthly level: including industry-by-year fixed effects would create severe event sparsity within industry-year cells, preventing clean identification of the triple interaction with the monthly $Attention_t$ index. Results are robust to including industry fixed effects without year-FE absorption. The baseline interaction ($ActionStrength \times \hat{\pi}_{i,e-1}$) in this table is 1.02–1.08, modestly lower than the 1.19 reported in Table 18 (which also uses firm and year fixed effects). The difference reflects the inclusion of $Attention_t$ and its lower-order interactions ($ActionStrength \times Attention_t$ and $\hat{\pi}_{i,e-1} \times Attention_t$), which together span variation that is correlated with the baseline interaction. Mechanically, $\hat{\pi}_{i,e-1} \times Attention_t$ captures cross-sectional-times-time variation: firms with higher pre-event credibility beliefs tend to announce in higher-salience months (plausibly because more credible firms time actions to salient periods), so this lower-order term partially absorbs variation that otherwise enters β_3. The appropriate comparator for the amplification coefficient (0.54–0.59) is the table's own β_3 baseline of 1.02–1.08, not 1.17 or 1.19; comparing the amplification against a different specification's baseline would conflate specification effects with the salience result. This means the regulation and salience designs do not share an identical baseline. The appropriate comparator is each specification's own baseline: on that within-specification basis, standardization attenuation (≈62%: 0.74/1.19 ≈ 62%, pooled estimate) and salience amplification (≈53%: 0.54/1.02 in the preferred with-controls specification) are of comparable magnitude and together constitute an on/off test of the credibility channel, as described in Section OA-2.2. Standard errors are double-clustered at the firm and year levels, consistent with the paper's default convention. Year-level clustering is particularly important here because $Attention_t$ is a time-series index common to all firms in a given month; without year clustering, standard errors would not account for the Moulton-type inflation arising from this aggregate regressor (firm-level clustering alone does not correct for cross-sectional correlation induced by a common time-series shock). We note that $Attention_t$ is constructed partly from earnings-call text produced by the firms in the sample, so it may partly reflect supply-side climate disclosure intensity rather than purely exogenous investor attention; as a robustness check, the index is also reconstructed from financial news only (excluding earnings calls), with qualitatively similar results (see Section OA-2.2 for details). Within R^2 is reported. The lower-order interactions with $Attention_t$ are insignificant, isolating the amplification effect to the triple interaction. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 25: Robustness Tests

	(1) Alt. events	(2) Alt. intent measure	(3) Placebo	(4) Firm FE	(5) IPW
Dependent variable: CAR [-1,+1]					
ActionStrength	0.57***	0.54***	0.49***	0.51***	0.54***
	(0.17)	(0.16)	(0.18)	(0.19)	(0.17)
ActionStrength $\times\hat{\pi}_{i,e-1}$	1.19***	1.14***	0.12	1.08***	1.19***
	(0.33)	(0.31)	(0.28)	(0.36)	(0.34)
Controls	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	Yes	Yes	Yes	No	Yes
Year FE	No	No	No	No	No
Firm FE	No	No	No	Yes	No
IPW	No	No	No	No	Yes
Observations	4,812	4,812	4,812	4,812	4,812
Within R^2	0.20	0.19	0.11	0.23	0.20

Notes. This table reports robustness tests for the baseline interaction result from Table 15. Column (1) uses an alternative classification of climate-related corporate actions (sample size held constant by reclassification). Column (2) replaces the FinBERT-based credibility measure with a dictionary-based alternative. Column (3) is a *non-climate-disclosure placebo* using non-ESG corporate disclosures; the interaction is both statistically and economically insignificant (point estimate an order of magnitude smaller than baseline), while the standalone *ActionStrength* coefficient captures general announcement effects unrelated to credibility beliefs formed from climate-specific communication. Column (4) adds firm fixed effects (replacing industry-by-year fixed effects) and restricts identification to within-firm variation over time. The interaction coefficient in Column (4) (1.08) is lower than the within-firm estimate reported in Table OA-2 Column (1) (1.28) because the two specifications absorb time variation differently. Table OA-2 Column (1) includes both firm fixed effects *and* year fixed effects alongside the full control set. Column (4) here, by contrast, replaces the preferred industry-by-year fixed effects with firm-only fixed effects and does *not* include separate year fixed effects, leaving aggregate time-series variation unabsorbed. This unabsorbed variation enters the residual, increasing residual variance and reducing the precision of the interaction estimate. The direction of any resulting bias depends on the correlation between aggregate year-level shocks and the interaction regressor; year fixed effects in Table OA-2 Column (1) remove this common year-level variation entirely, yielding a cleaner estimate. The gap between 1.28 and 1.08 is therefore most naturally interpreted as the additional noise and potential omitted-variable contamination introduced by excluding year fixed effects, rather than a systematic bias in a known direction. Both specifications use the identical control vector. The three specifications (1.28 with firm+year FE, 1.08 with firm FE only, 1.17 preferred with industry×year FE) are internally consistent: all deliver a large, statistically significant positive interaction. The ordering (1.28 > 1.17 > 1.08) is not monotone in the number of fixed effects, which might seem surprising. The key is that firm+year FE (Table OA-2 Column (1), yielding 1.28) and industry×year FE (Column (3), yielding 1.17 in Table 15) absorb *different* sources of variation: firm+year FE removes persistent cross-sectional variation in $\hat{\pi}$ and shifts identification to within-firm fluctuations. Under classical measurement error, both specifications are biased toward zero; the difference between 1.28 and 1.17 reflects the composition of identifying variation rather than the removal of upward bias from time-invariant confounders. If positive bias from such confounders dominated, the within-firm coefficient would be expected to *decline* relative to the cross-sectional benchmark, which is not what we observe. This pattern is consistent with both estimates understating the underlying credibility-pricing effect if bias runs toward zero, though time-varying confounding cannot be fully ruled out; accordingly, we interpret the stability of the within-firm coefficient as evidence against persistent firm-quality confounding rather than as a universal lower-bound argument against all forms of bias. Separately, industry×year FE absorbs industry-level time shocks but does not remove within-firm time trends. Firm-FE-only (Column (1) here, yielding 1.08) strips out firm-level fixed differences but not year-level common shocks, which introduces omitted-variable contamination that attenuates the interaction estimate. The preferred industry×year FE specification (1.17) sits between the two because it addresses industry-time confounds without the additional within-firm restrictions of the firm+year FE design. Column (5) re-estimates the baseline using inverse probability weights (IPW) constructed from a probit model of event-sample inclusion on log total assets (firm size), book-to-market, leverage, emissions intensity, analyst coverage, and industry fixed effects. All specifications include the full set of controls (log assets, book-to-market, leverage, prior returns, emissions intensity, and analyst coverage). The main effect of $\hat{\pi}_{i,e-1}$ is included in all specifications but suppressed from the displayed rows for brevity; its coefficient is positive and consistent in sign with Table 15 across all columns. Standard errors are double-clustered at the firm and year levels. Within R^2 is reported throughout. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 26: Within-Firm Updates in Investor Beliefs and the Pricing of Climate Actions

	(1) Baseline	(2) + Trend	(3) + Level	(4) + Lagged	(5) + Future level	(6) Residualized	(7) **Lead placebo**
ActionStrength	0.58***	0.55***	0.53***	0.51***	0.50***	0.48***	0.51***
	(0.19)	(0.18)	(0.18)	(0.17)	(0.17)	(0.17)	(0.17)
$\Delta\hat{\pi}_{i,e-1}$	0.39**	0.34**	0.31*	0.28*	0.26*		
	(0.19)	(0.18)	(0.17)	(0.17)	(0.16)		
ActionStrength $\times\Delta\hat{\pi}_{i,e-1}$	1.45***	1.39***	1.33***	1.27***	1.24***		
	(0.41)	(0.39)	(0.38)	(0.37)	(0.36)		
$\hat{\pi}_{i,e-1}$ (level)			0.25	0.23	0.21	0.19	0.20
			(0.20)	(0.19)	(0.19)	(0.18)	(0.19)
ActionStrength $\times\hat{\pi}_{i,e-1}$ (level)			0.56*	0.53	0.51	0.48	0.49
			(0.33)	(0.32)	(0.31)	(0.31)	(0.32)
$\Delta\hat{\pi}_{i,e-2}$ (lagged)				0.06	0.05	0.04	0.05
				(0.17)	(0.16)	(0.16)	(0.17)
$\Delta\hat{\pi}_{i,e+1}$ (future)					0.09	0.08	0.07
					(0.18)	(0.18)	(0.18)
$\Delta\hat{\pi}^{res}_{i,e-1}$						0.22	
						(0.18)	
ActionStrength $\times\Delta\hat{\pi}^{res}_{i,e-1}$						1.16***	
						(0.34)	
ActionStrength $\times\Delta\hat{\pi}_{i,e+1}$							**0.06**
							(0.33)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	No	No	No	No
Firm-specific linear trend	No	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	No	No	No	Yes	Yes	Yes	Yes
Observations	4,812	4,812	4,812	4,812	4,812	4,812	4,812
Within R^2	0.22	0.25	0.26	0.27	0.27	0.28	0.28

Notes. This table examines within-firm variation in inferred credibility and its relation to announcement returns. In this table, e indexes event-quarters (as distinct from its use as an event index in the primary event-study tables); $e-1$ denotes the quarter immediately preceding the event-quarter. Columns (1)–(6) follow the same progressive specification structure as the prior version. Column (7) adds the lead placebo interaction $ActionStrength_{i,e} \times \Delta\hat{\pi}_{i,e+1}$, where $\Delta\hat{\pi}_{i,e+1}$ is the change in the credibility measure over the 12 months *following* the event. The estimated coefficient is small and statistically insignificant (0.06, SE 0.33), in contrast to the pre-event interaction in Columns (1)–(5) (1.24–1.45, all significant at the 1% level). This placebo rules out the interpretation that the within-firm interaction effect is driven by anticipatory or mechanically related changes in post-event disclosure. Because $\hat{\pi}_{i,e-1}$ is constructed at the firm-quarter level, multiple events from the same firm within the same quarter share identical prior values of $\hat{\pi}_{i,e-1}$; results are robust to restricting to one event per firm-quarter. All explanatory variables are constructed using information available strictly prior to the announcement window. Columns including firm-specific linear trends (Columns (2)–(7)) estimate these trends within-firm using the available event observations; identification in those columns comes from firms with at least three events. Standard errors are double-clustered at the firm and year levels. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Lead placebo: future changes in inferred credibility. As a complementary placebo within the within-firm specification, Column (7) introduces the interaction between *ActionStrength* and the future change in inferred credibility, $ActionStrength_{i,e} \times \Delta\hat{\pi}_{i,e+1}$. If firms planning stronger climate actions simultaneously adjusted their communication dynamics in anticipation, this future interaction would load positively on announcement returns. Instead, the estimated coefficient is small and statistically indistinguishable from zero (0.06, SE 0.33), in contrast to the pre-event interaction in Columns (1)–(5) (1.24–1.45, all significant at the 1% level). Together with the level placebo in Column (5) of Table 26 ($\Delta\hat{\pi}_{i,e+1}$ main effect: 0.09, SE 0.18), this result indicates that neither levels nor changes in post-event credibility measures predict contemporaneous returns, consistent with the interpretation that markets price ex-ante beliefs formed from prior communication.

Table 27: Credibility-Pricing Heterogeneity and Benchmark-Rejection Tests

	Panel A: Industry × Year FE		Panel B: Firm + Year FE	
	(1) HighCost	(2) Joint	(3) HighCost	(4) Joint
Baseline credibility interaction				
ActionStrength $\times\hat{\pi}_{i,e-1}$	1.09***	0.91***	1.25***	1.08***
	(0.30)	(0.32)	(0.36)	(0.34)
Cost-asymmetry benchmark test				
ActionStrength $\times\hat{\pi}_{i,e-1} \times HighCost_{i,e}$	0.82***	0.71**	0.95***	0.84***
	(0.31)	(0.32)	(0.35)	(0.33)
ActionStrength $\times HighCost_{i,e}$	0.45**	0.38*	0.52**	0.44**
	(0.20)	(0.21)	(0.23)	(0.22)
$\hat{\pi}_{i,e-1} \times HighCost_{i,e}$	0.19	0.15	0.22	0.17
	(0.14)	(0.13)	(0.16)	(0.15)
Consistency amplification				
ActionStrength $\times\hat{\pi}_{i,e-1} \times Consistency_{i,e-1}$		0.62**		0.71**
		(0.29)		(0.31)
Salience-by-diagnosticity interaction				
ActionStrength $\times\hat{\pi}_{i,e-1} \times Attention_t \times HighCost_{i,e}$		0.48**		0.55**
		(0.23)		(0.25)
Controls	Yes	Yes	Yes	Yes
Fixed effects	Ind×Yr	Ind×Yr	Firm + Yr	Firm + Yr
Observations	4,812	4,812	4,812	4,812
Within R^2	0.23	0.25	0.28	0.29

Notes. Saturated specification (Equation (OA-7)). The coefficient on $ActionStrength \times \hat{\pi}_{i,e-1} \times HighCost_{i,e}$ tests the fixed-precision Bayesian benchmark (Section OA-1.1). Under the benchmark this triple interaction should be zero or negative conditional on the two-way $ActionStrength \times HighCost$ term; positive coefficients indicate amplification consistent with inverse planning. Panel B replaces industry-by-year fixed effects with firm and year fixed effects, identifying variation from within-firm changes in action costliness and credibility. $HighCost_{i,e} = 1$ for actions with large verifiable capital commitments or science-based targets; $Consistency_{i,e-1}$ is within-firm cross-venue cosine similarity in the prior year; $Attention_t$ is the monthly climate-salience index. All required lower-order interactions are included in estimation. Controls include log assets, book-to-market, leverage, prior returns, emissions intensity, and analyst coverage. Standard errors are double-clustered at the firm and year levels and reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 28: Continuous Capex-Intensity Gradient

	(1) Industry × Year FE	(2) Firm + Year FE
ActionStrength $\times\hat{\pi}_{i,e-1} \times \log_cost_{i,e}$	0.67***	0.79***
	(0.24)	(0.28)
Controls	Yes	Yes
Fixed effects	Ind×Yr	Firm + Yr
Observations	4,812	4,812
Within R^2	0.24	0.29

Notes. Continuous specification using standardized $\log(capex/assets)$ in place of the binary *HighCost* indicator from Table 27. A positive coefficient on the triple interaction indicates monotone amplification of the credibility-pricing slope with cost intensity, against the fixed-precision benchmark which predicts zero or negative. All required lower-order interactions are included in estimation. Standard errors are double-clustered at the firm and year levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 29: Alternative Cost Measure: Third-Party Verifiability

	(1) Industry × Year FE	(2) Firm + Year FE
ActionStrength $\times\hat{\pi}_{i,e-1} \times$ *Verifiable*$_{i,e}$	0.66***	0.74***
	(0.25)	(0.27)
Controls	Yes	Yes
Fixed effects	Ind×Yr	Firm + Yr
Observations	4,812	4,812
Within R^2	0.22	0.27

Notes. *Verifiable*$_{i,e} = 1$ for actions with mandatory third-party assurance (SBTi-validated targets, externally audited emissions reductions, certified offset programs). Replaces *HighCost* in the saturated specification of Table 27. Confirms robustness of the benchmark rejection across cost operationalizations. All required lower-order interactions are included in estimation. Standard errors are double-clustered at the firm and year levels. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

B Proofs

Scope of all propositions. All propositions are stated over the empirically relevant support $\mathcal{D}$ under Assumption 1. Dominance claims refer to the quality-spread component outweighing bounded negative contributions from the belief-revision channel on this support. These results are therefore local rather than global and do not rely on monotonicity of $\pi_1(D)$ outside the observed range of the data.

Let $Q \equiv p\ell + (1-p)$ throughout, so $h(p,\ell) = p(1-p)(\ell-1)/Q$. Note the useful identity

$$Q - p(\ell-1) = \big[p\ell + (1-p)\big] - \big[p\ell - p\big] = p\ell + 1 - p - p\ell + p = 1, \tag{39}$$

which is used repeatedly in the proofs below to simplify expressions involving $\partial h/\partial \ell$ and $\partial \pi_2/\partial \ell$.

B.1 Proof of the Pricing Cross-Partial

Dominance of the quality-spread term. The dominance condition in Proposition 1 is not a global analytical result but is imposed over the empirically relevant support $\mathcal{D}$ via Assumption 1. Empirically, the distribution of $\hat{\pi}$ is concentrated below the threshold at which $\partial h/\partial p$ becomes negative (median $\hat{\pi} \approx 0.46$), and $\Delta(D)$ is increasing in D, so the direct quality-spread channel is quantitatively dominant over the observed support. This restriction is sufficient for the directional comparative static and aligns the theoretical prediction with the empirical specification.

The announcement return is given by

$$R(D, A) = \Delta(D) \cdot h(\pi_1(D), \ell(A)),$$

where $\Delta(D)$ is the quality spread and $h(p,\ell) \equiv p(1-p)(\ell-1)/[p\ell + (1-p)]$ is the belief-revision function. The empirical interaction coefficient β_3 identifies the total derivative

$$\frac{d}{dD}\left(\frac{\partial R}{\partial A}\right),$$

which combines both the belief-revision channel and the quality-spread channel — a richer object than the structural cross-partial alone. This total derivative, with $\hat{\pi}$ serving as a sufficient statistic for disclosure information, is the object most directly mapped into the empirical specification and is the natural target of

estimation. It differs from the structural partial derivative

$$\frac{\partial^2 R}{\partial A \, \partial \pi_1},$$

which isolates only the belief-revision channel holding D fixed. The sign of the total derivative need not coincide with the sign of the belief-revision cross-partial alone. While $\partial^2 h / \partial \ell \, \partial p$ is positive only over a subset of the parameter space ($p < 1/(1+\ell)$), the empirical interaction coefficient captures the combined effect of both the quality-spread channel $\Delta'(D)$ and the belief-revision channel. A positive $\hat{\beta}_3$ therefore does not require the cross-partial of h to be globally positive, but rather reflects that the total derivative is expected to be positive over the empirically relevant region of the parameter space. To isolate the contribution coming purely from the belief-revision channel, we examine the cross-partial

$$\frac{\partial^2 h}{\partial \ell \, \partial p}.$$

Differentiating h with respect to ℓ and then with respect to p yields

$$\frac{\partial^2 h}{\partial \ell \, \partial p} = \frac{(1-p) - p\ell}{Q^3}, \quad \text{where } Q = p\ell + (1-p).$$

This cross-partial is positive whenever $p < 1/(1+\ell)$ (equivalently $\pi_2 < 1/2$). Multiplying by $\partial \ell / \partial A > 0$ (from the monotone likelihood ratio property) gives the contribution of the belief-revision channel to the total derivative.

Derivation of $\partial^2 h / \partial \ell \, \partial p$. With $Q \equiv p\ell + (1-p)$, using the quotient rule on $h(p, \ell) = p(1-p)(\ell - 1)/Q$:

$$\frac{\partial h}{\partial \ell} = \frac{p(1-p)\, Q - p(1-p)(\ell-1) \cdot p}{Q^2} = \frac{p(1-p)\big[Q - p(\ell-1)\big]}{Q^2} = \frac{p(1-p)}{Q^2} > 0$$

since $Q - p(\ell - 1) = 1$. Differentiating $p(1-p)/Q^2$ with respect to p (using $\partial Q / \partial p = \ell - 1$):

$$\frac{\partial^2 h}{\partial \ell \, \partial p} = \frac{(1-2p)Q - 2p(1-p)(\ell-1)}{Q^3},$$

which simplifies (substituting $Q = 1 + p(\ell-1)$, so $(1-2p)Q - 2p(1-p)(\ell-1) = (1-2p) + (\ell-1)[p(1-2p) - 2p(1-p)] = (1-2p) + (\ell-1)p(-1) = 1 - p - p\ell$) to

$$\frac{\partial^2 h}{\partial \ell \, \partial p} = \frac{(1-p) - p\ell}{Q^3},$$

which is positive if and only if $p < 1/(1+\ell)$, equivalently $\pi_2 < 1/2$. The proposition characterizes the cross-partial sign in the interior region where $\pi_2 < 1/2$. As discussed below, this condition holds for the majority of the empirical sample but not pointwise; we therefore state Proposition 2 as a local result and interpret $\hat{\beta}_3$ as a reduced-form average over the support of $\hat{\pi}_{i,e-1}$ rather than as a pointwise structural cross-partial.

Threshold comparison. Note that the threshold for $\partial h / \partial p > 0$ is $p < 1/(1+\sqrt{\ell})$ (Property (P2), Equation (A.4)), which is strictly larger than the cross-partial threshold $p < 1/(1+\ell)$ since $\sqrt{\ell} < \ell$ for $\ell > 1$. The empirical interaction β_3 captures the combined effect of belief updating and the disclosure-dependent quality spread $\Delta'(D) \cdot \partial h / \partial \ell > 0$; a positive $\hat{\beta}_3$ is therefore consistent with the total-derivative amplification

even when the cross-partial of h is negative for some observations.

A first-order-plus-cross-partial approximation of $R(D, A)$ around average values $(\bar{A}, \bar{\pi}_1)$ yields

$$R \;\approx\; \beta_0 + \beta_1(A - \bar{A}) + \beta_2(\pi_1 - \bar{\pi}_1) + \beta_3(A - \bar{A})(\pi_1 - \bar{\pi}_1) + \text{h.o.t.},$$

where this expansion is understood as a *reduced-form approximation* in the sufficient statistic $\pi_1(D)$, rather than a structural expansion treating π_1 as independent of D. Accordingly, β_3 should be interpreted as the local reduced-form interaction coefficient implied by the model. β_3 provides a reduced-form approximation to the total-derivative amplification

$$\beta_3 \;\approx\; \frac{d}{dD}\left(\frac{\partial R}{\partial A}\right),$$

in a local neighborhood of the empirical distribution, evaluated over the empirical distribution (with $\hat{\pi}$ serving as a reduced-form sufficient statistic for the disclosure information D). It therefore captures *both* channels—the variation in the quality spread $\Delta'(D)$ (the dominant component) and the dependence of the belief-revision function h on π_1—rather than the partial cross-derivative of h holding π_1 fixed. The pure quadratic terms are second-order and subsumed in the residual. The interaction term $(A - \bar{A})(\pi_1 - \bar{\pi}_1)$ therefore provides the key reduced-form restriction taken to the data, consistent with Proposition 2.