Ethical Capital, Coordination, and the Correction of Externalities

Harshini Shanker[*]

London Business School

November 14, 2025

Abstract

This paper builds a general equilibrium model of ethical investing and coordinated externality mitigation. Ethical behavior often departs from Nash rationality by relying on normative principles rather than belief-based best responses. I introduce the *Kant equilibrium* as a complement to Nash, and show that reasoning heterogeneity and coordination technology, rather than preferences alone, select unique coordination equilibria, explain minority catalysis in activism and predict when one-share-one-vote regimes aid or impede externality correction. I test model predictions on a global panel of mutual funds and ETFs, using textual analysis of filings to identify ethical mandates. Using the staggered rollout of the *Climate Action 100+* initiative as a positive coordination shock, I find that engagement-oriented funds increase ownership in targeted high-externality firms relative to other firms and non-engagement funds. Dynamic estimates from a stacked event study confirm that reallocations are persistent. Using 2SLS, I show that ethical capital causally improves firm-level externality outcomes through coordinated engagement channels.

Keywords: ethical investing, shareholder activism, coordination, externalities, Kant equilibrium

JEL Classification: G11, G12, G23, G34, D62

[*] This paper has benefitted greatly from discussions with Vikas Agarwal, Brad Barber, Suleyman Basak, Jean-Pierre Benoît, Rebecca De Simone, Alex Edmans, Julian Franks, Roberto Gomez Cram, Christopher Hennessy, Zoran Ivkovich, Emre Ozdenoren, Anna Pavlova, Uday Rajan, Henri Servaes, Hanbin Yang, and Ayako Yasuda.

1 Introduction

Institutional investors often target externalities, using exit strategies to raise the cost of capital for harmful firms (Heinkel et al., 2001; Hong & Kacperczyk, 2009) and voice and engagement to push for change from within (Broccardo et al., 2022). Success through either channel turns on collective action, yet the coordination mechanisms that help or hinder such action remain underexplored. This paper asks how ethical commitments and coordination mechanisms shape real outcomes—in particular, under what circumstances does collective investor action succeed in correcting externalities, and what governance structures best facilitate it?

I build a general equilibrium model that introduces heterogeneity in not only investor preferences but also reasoning logic, bringing insights from the philosophy of ethics into a strategic portfolio-choice setting. The model shows that investors' reasoning logic—rather than preferences alone—determines when collective action succeeds. It explains persistence in activism despite coordination failures, predicts minority catalysis effects, and delineates when one-share-one-vote regimes aid or impede externality correction.

I test the model's predictions on global mutual fund data from 2015–2025, using textual analysis of regulatory filings to classify funds by ethical mandate. I use a triple-difference design to show that when coordination costs fall, as during the *Climate Action 100+* initiative, engagement-oriented funds increase ownership in targeted firms relative to both non-targets and funds without engagement mandates, validating the model's coordination channel. Dynamic estimates from a stacked event study confirm that these reallocations persist over time. I instrument for engagement-oriented firm ownership using the coordination shock as well as flow-driven shifts in ethical capital supply, and estimate the causal effect of ethical capital on firm-level externalities. A one-percentage-point rise in engagement-oriented ownership in a firm reduces emissions intensity by roughly 20 percent.

These results position ethical orientation as a missing state variable in asset pricing and corporate governance, linking preference heterogeneity, reasoning modes, and coordination technology with real economic impact. To formalize these mechanisms, the theoretical analysis begins with a model in which investors allocate wealth toward an externality-generating firm, trading off pecuniary returns against ethical disutility from social harm.

How externalities enter an investor's objective is not obvious *ex ante*, and I draw on the philosophy of ethics to provide microfoundations for investor behavior in this setting. Investors differ in their *moral imperatives*—some engage to minimize harm (utilitarians), while others exclude or divest to avoid personal complicity in harm (deontologists). Investors also differ in their *strategic imperatives*—Nash agents take choices of others as given and respond optimally, while Kant agents act on normative principles that do not depend on beliefs about others. This two-by-two taxonomy yields four investor archetypes and a coordination game that endogenously determines equilibrium allocations, prices, and externality levels.

Kant agent behavior is central to the model. The Kant agent evaluates the moral permissibility of an action by imagining a world where everyone acts similarly. Termed the *universalizability principle* by Kant (1797), this mode of reasoning asks 'would I endorse this action if everyone chose it?', rather than 'is this action optimal given what others actually do?' By definition, Kant reasoning departs from best-response logic. Further, counterfactual reasoning violates the Nash requirement that beliefs be correct in equilibrium, as the Kant agent forms 'what-if' beliefs that differ from actual play. Existing models cannot account for these behaviors without invoking ad hoc preferences or belief distortions.

This paper views Kant-type reasoning as a distinct mode of strategic reasoning rooted in normative principles and defines the Kant equilibrium solution concept to formalize it. That said, Kant reasoning is not incompatible with Nash; and the foremost contribution of this paper is to formalize a model where Kant and Nash reasoning coexist in one unified setting. Three core insights follow.

First, Kant investors, who act on principled commitments rather than best response, can catalyze coordination even when in the minority, as they provide credible commitments that shift others' incentives to act. Using evolutionary game theory concepts, I demonstrate conditions in which Kant types can invade and persist, even when initially rare. A small fraction of such principled investors can therefore catalyze large-scale abatement even without wealth or voting power—a minority-catalysis result impossible in standard Nash frameworks.

To be sure, Stackelberg models predict that large investors can lead collective action by setting precedents. Fahlenbrach et al. (2023) find supporting evidence in Norges Bank Investment Management (NBIM) pre-declaring its voting intentions. However, the circumstances under which principled *minorities* catalyze *majority* action, is both theoretically and empirically unexplored, despite being common in practice. For example, Candriam, a US-based asset manager, pre-declares its votes on its website despite an AUM of only USD150B[1]. Standard models cannot rationalize such behavior. The Kant equilibrium framework provides a natural explanation.

Second, democratizing shareholder influence has non-monotonic welfare effects: one-share-one-vote regimes outperform one-investor-one-vote regimes when wealth is concentrated among exclusion-oriented investors, but neither dominates when wealth lies with engagement-oriented investors. Efforts to democratize shareholder influence must consider general equilibrium feedback between wealth, motives, and coordination institutions.

Third, heterogeneity in preferences and reasoning modes together break the multiplicity of equilibria endemic to coordination games. Nash utilitarians exhibit strategic substitutability, Nash deontologists complementarity, and Kant agents act according to fixed-point logic, independent of others' observed play. Even without introducing payoff uncertainty or other perturbations, uniqueness emerges from the fact that (i) simultaneous presence of complementarity and substitutability dampens rather than amplifies feedback loops, and (ii) principled commitment anchors expectations and limits self-fulfilling belief cascades. The model thus highlights an alternative mechanism for equilibrium selection, complementing global-games approaches.

The model is robust to bounded rationality and local distortions of universalizability, showing that the core insights hold even when Kant agents make systematic errors in their counterfactual beliefs or when all investors exhibit mild levels of behavioral bias. The framework thus provides a general-purpose toolkit for analyzing norm-based behavior in strategic environments.

I test the model's predictions using the universe of global open-ended mutual funds and ETFs from 2015–2025. I classify funds as engagement-oriented (utilitarian), exclusion-oriented (deontological),

[1] NBIM manages over USD1.8T for comparison. Big Three manage USD5-10T each. Data retrieved June 2025.

or neutral. The classification challenge is to isolate engagers that target social and environmental externalities from those that engage on, say, only corporate governance concerns. Similarly, ethically motivated exclusion screens must be distinguished from financially motivated ones. I use textual analysis of prospectuses and stewardship statements to identify not just broad engagement or exclusion mandates, but also the specific issues targeted. The analysis reveals that exclusion-oriented funds systematically avoid environmentally or socially controversial holdings, while engagement-oriented funds retain exposure to them, consistent with an intent to improve practices from within.

To identify coordination effects, I exploit the staggered rollout of the *Climate Action 100+* initiative as a quasi-natural experiment that created exogenous focal points for collective engagement. Assignment of firms to CA100+ targeting depended on predetermined and publicly known firm characteristics such as emissions scale and industry relevance, not on contemporaneous performance or investor composition, providing plausibly exogenous treatment variation. Using a triple-difference framework, I find that engagement-mandate funds increase ownership following the shock in targeted high-externality firms relative to both non-targets and non-engagement funds, consistent with the model's predictions about coordination channels.

To trace dynamics while avoiding bias from staggered treatment timing, I estimate a stacked event study that re-indexes each firm's treatment date and pools post-announcement windows. The stacked estimator (Callaway & Sant'Anna, 2021) prevents contamination from already-treated units, allowing identification of the average dynamic response to coordination shocks. The results rule out pre-trends and show a sharp and sustained increase in engagement-oriented funds' ownership of targeted firms, consistent with the model's predictions.

These shocks provide the first-stage variation for a two-stage least squares design estimating the causal effect of engagement-oriented ownership on firm-level externalities. I use two instruments to generate plausibly exogenous shifts in engagement-oriented ownership: (i) firm-level CA100+ targeting, capturing the coordination channel, and (ii) flow-driven shocks to investable capital, capturing the ethical capital supply channel. Together they isolate ownership changes unrelated to firm fundamentals, mitigating concerns of reverse causality or omitted firm-level trends.

The identification challenge is to control for confounding factors that could affect both ownership and externalities, such as media attention or pressure from other stakeholder groups on targeted firms. I address the exclusion restriction by exploiting the staggered rollout of the initiative, controlling for cohort-time and industry-time fixed effects in both stages to absorb alternative channels of influence.

The evidence shows that engagement-oriented ownership causally improves firm-level externality outcomes. A one-percentage-point increase in engager ownership reduces emissions intensity by roughly 20 percent over a two-year horizon, an economically meaningful effect comparable to those achieved through major regulatory interventions or carbon pricing adjustments of similar scale. The analysis also speaks to the relative importance of two complementary channels through which the Climate Action 100+ initiative operates: one, directly lowering coordination costs and incentivizing increased ownership among engagement-oriented investors; and two, enhancing the salience of climate risks in targeted firms to all investors. Taken together, the theory and empirics demonstrate that reasoning logic and coordination technology—rather than moral preference alone—govern when ethical capital is able to effect real change.

Literature

A growing literature examines how heterogeneity in moral and social preferences shapes portfolio choice and asset pricing. Investors exhibit an intrinsic willingness to sacrifice returns for social impact (Bénabou et al., 2020; Hartzmark & Sussman, 2019), modulated by motives such as warm-glow utility (Carpenter, 2021), reputation concerns in activist behavior (T. L. Johnson & Swem, 2021), and social signaling (Riedl & Smeets, 2017).

Theoretical models embed such preferences into equilibrium asset-pricing frameworks, showing that socially responsible assets can command valuation premia and that firms excluded by certain investors face higher costs of capital (Edmans et al., 2023; Heinkel et al., 2001; Pástor et al., 2021; Pedersen et al., 2020). Empirical work confirms that asset prices and capital flows respond to shifts in these preferences—green assets enjoy valuation premia while carbon-intensive assets face higher expected returns (Bolton & Kacperczyk, 2021; Hong & Kacperczyk, 2009; Pástor et al., 2022). Pedersen (2025) offers a review of this literature. However, cost of capital effects alone are limited in their ability to induce real corporate change (Broccardo et al., 2022). This limitation motivates the governance channel—where preferences operate through voice rather than prices.

A parallel literature explores *voice*, in contrast to *exit*, as mechanisms for investor influence, showing that voice tends to dominate when ownership is concentrated or when information asymmetries allow informed intervention (Becht et al., 2019; Berk & van Binsbergen, 2025; Edmans & Manso, 2011); see (Denes et al., 2017) for a survey. Theoretical models further demonstrate that shareholder engagement can enhance both firm value and social welfare (Albuquerque et al., 2022).

One branch of this literature studies *private* engagements, negotiated directly between investors and firms. Passive investors can exert influence through voting and monitoring (Appel et al., 2016), especially when engagement incentives are explicitly linked to performance fees (Becht et al., 2023). Hedge-fund activism shows that assertive but often non-public interventions generate real operational and performance gains (Brav et al., 2015). Clinical evidence from the Hermes Focus Fund documents that confidential dialogues improved governance outcomes without public confrontation (Becht et al., 2009). Subsequent empirical studies generalize this evidence, showing that active ownership through private engagements raises environmental, social, and governance standards and can increase firm value (Barko et al., 2017; Dimson et al., 2015; Hoepner et al., 2024; Yang & Yasuda, 2023), while broader cross-country evidence links institutional ownership to stronger CSR outcomes (Dyck et al., 2019).

A smaller but growing strand highlights *coordinated* engagements—collective action by multiple investors pursuing common goals. Collaborative investor coalitions can achieve environmental improvements beyond what unilateral activists accomplish (Becht et al., 2017; Dimson et al., 2025). These studies point to coordination technology—the institutional and informational structures enabling dispersed ethical capital to act collectively—as a central but understudied determinant of investor impact. This paper contributes to this emerging literature by formalizing how coordination technology interacts with investor preferences and reasoning modes to shape collective action outcomes.

This paper unifies these literatures—moral-preference theory and shareholder engagement—by embedding both reasoning-mode heterogeneity and coordination technology in a single equilibrium framework. While prior work emphasizes taste heterogeneity, I show that differences in reasoning logic generate distinct equilibrium outcomes that cannot be captured by preferences alone, including minority catalysis effects. Modeling coordination technology explicitly links shareholder-governance structures with collective-action outcomes.

This paper also contributes to the corporate governance literature on shareholder power and voting regimes. In traditional one-share-one-vote (OSOV) systems, influence scales with wealth. Emerging pass-through voting and investor assembly reforms aim to realign corporate control with the preferences of ultimate beneficiaries (Bainbridge, 2005; Fisch & Schwartz, 2023; Gramitto Ricci et al., 2025; Zingales et al., 2025). While these reforms aim to purify OSOV—ensuring that each share's vote reflects the will of its true owner—this paper questions the normative premise that a fully realized OSOV system is necessarily welfare-enhancing. By modeling how coordination and wealth distribution jointly determine influence, I show that more "democratic" representation of shareholder preferences can, in some equilibria, suppress rather than amplify socially beneficial engagement.

Finally, this paper contributes to theories of equilibrium selection in environments with strategic complementarities. Global-games models show how incomplete information and higher-order beliefs govern equilibrium multiplicity (Angeletos et al., 2007; Carlsson & van Damme, 1993; Goldstein & Huang, 2016; Morris & Shin, 1998, 2002). Introducing heterogeneity in reasoning modes offers a new selection mechanism that unites moral extensions of utility with general-equilibrium coordination, explaining how minority investors can catalyze collective action even when purely strategic models predict inertia.

The paper proceeds as follows. Section 2 develops the portfolio-choice environment and formalizes the coordination mechanism. Section 3 establishes equilibrium existence, uniqueness, and characterizes the key comparative statics. Section 4 presents empirical evidence on a subset of the model's testable predictions, including causal evidence of the effects of ethical orientation on real outcomes. Section 5 examines theoretical extensions and robustness to behavioral distortions, bolstering the model as a general-purpose toolkit for analyzing norm-based behavior in strategic environments. Section 6 concludes.

2 Model

A firm raises capital for an investment that yields a stochastic cash flow θ and generates a non-stochastic negative externality λ. An ethical investor allocates wealth between a risk-free asset and the firm's stock. The externality imposes no direct cost on the firm or investor, but causes moral disutility to investors. Shareholders can pressure the firm to offset part of the externality, with the resulting reduction modeled as the outcome of a coordination game.

Externalities matter differently depending on an investor's ethical outlook. The philosophy of ethics distinguishes two central imperatives that map to economic preferences: a duty to avoid personal complicity in harm (deontology) and a duty to minimize aggregate harm (utilitarianism).

Definition 1 (Moral Imperative)**.** *A* moral imperative *is an obligation to act (or refrain from acting) according to a conception of right and wrong.*

1. Utilitarianism*: maximize aggregate well-being or minimize aggregate harm.*

2. Deontology*: avoid personal complicity in harm.*

Deontological investors exclude or reduce holdings in harmful firms to avoid moral contamination. Utilitarian investors hold such assets only if their participation can induce abatement of the harm. Loosely speaking, deontologists may be viewed as the exit (or exclude) camp and utilitarians as the voice (or engage) camp in the ethical investing literature.

Ethical philosophy also guides investor response to the anticipated actions of others, distinguishing between Nash best-response and Kant universalizability.

Definition 2 (Strategic Imperative)**.** *A* strategic imperative *prescribes how an agent forms expectations about others' behavior and chooses their own.*

1. Nash best response*: act optimally given others' actions.*

2. Kant universalizability*: act only on maxims that one could will as universal laws.*

Termed the universalizability principle, Kant's doctrine requires the agent to first evaluate the moral permissibility of a proposed action by imagining a world where everyone adopts that action. The question is ”would I endorse this action if everyone chose it?”, rather than ”is this action optimal given what others actually do?” This thought experiment tests whether the action can be consistently willed as a universal law without contradiction or undesirable outcomes. Only if the action passes this test does the Kant agent admit it into their choice set.[2] Kant reasoning is thus an alternative to Nash best-response reasoning, as it emphasizes principled commitment to norms over contingent adaptation to beliefs about others' actions. This structure mirrors rule-based institutional mandates, where action is selected based on a prescribed logic, not observed behavior.

However, Kantian reasoning is not incompatible with instrumental rationality; the foremost contribution of this paper is to formalize a model where Kant and Nash reasoning coexist as distinct modes of strategic thought in one unified setting.

Strategic imperatives are modeled as two equilibrium concepts—Nash and Kant equilibria. Moral imperatives are captured by moral disutility functions: $v(\cdot)$ for utilitarians (ends-focused) and $\chi(\cdot)$ for deontologists (means-focused). Together, they define four investor archetypes:

Table 1: Investor Types Defined by Moral and Strategic Imperatives

	Moral Imperative	
	Outcome Maximization	Complicity Avoidance
Strategic Imperative		
Best-response	Nash Utilitarian	Nash Deontologist
Universalizability	Kant Utilitarian	Kant Deontologist

[2] This form of reasoning is rooted in Kant's universalizability principle:

> Act only according to that maxim whereby you can, at the same time, will that it should become a universal law. – Kant (1797)

For brevity, formal properties of Kant equilibria are detailed in the Appendix. Section Section B.1 motivates theoretical foundations, Section B.4 compares Nash and Kant reasoning in the two-player prisoner's dilemma.

2.1 Utilitarian Moral Imperative

The utilitarian investor experiences a moral cost from exposure to activities that degrade collective well-being. For illustration, assume the externality in question is carbon emissions. Let λ denote the firm's baseline externality level and $\Delta\lambda_i$ denote the externality reduction that investor i perceives would result from collective activism. $\Delta\lambda_i$ is formally defined in the next section, but may be viewed for now as an investor-type-dependent function capturing the counterfactual or factual externality reduction from investor i's point of view.

The utilitarian investor's moral disutility is modelled as a convex function of residual levels of the externality: $v(\lambda - \Delta\lambda)$ with $v(0) = 0$, $v' > 0$, and $v'' > 0$. This modeling choice reflects empirical work in moral psychology (Kubany & Watson, 2003) showing that agents often experience disproportionately greater guilt or discomfort from marginal increases in perceived harm. Convex moral cost captures this intensifying psychological burden. Further, integrated assessment models such as DICE (Barrage & Nordhaus, 2023) and recent empirical work (Jarvis & Forster, 2024) suggest that economic damages from emissions are convex in total output—often quadratic. Since temperature increases are approximately linear in cumulative emissions, residual emissions generate convex welfare loss.

The model's predictions do not hinge on a specific functional form as long as it is continuously differentiable and convex. However, for illustrative purposes, a quadratic moral cost function is a natural way to model a utilitarian's disutility that is rooted in the social harm caused by emissions, and is used in this paper's numerical illustrations:

$$v(\lambda - \Delta\lambda_i) = \frac{1}{2}a(\lambda - \Delta\lambda_i)^2 \tag{1}$$

with $a > 0$ governing steepness of moral disutility.

2.2 Deontological Moral Imperative

In its purest form, deontology holds that certain actions are intrinsically wrong, regardless of consequences. While deontology is often associated with Kant (Kant, 1788, 1797), whose categorical imperative forbids complicity in wrongdoing without exception, this paper relies on the William David Ross school (Ross, 2002), which allows for competing duties and moral trade-offs. As mutual funds often have fiduciary obligations to deliver returns, diversification imperatives, or other practical considerations, a Rossian framework better captures real-world investor behavior.

Empirical evidence shows deontological investors are less responsive to financial incentives than utilitarian ones, supporting a more sharply convex moral cost profile (Niszczota et al., 2024). Related findings on taboo trade-offs (Tetlock et al., 2000) and protected values (Baron & Spranca, 1997) support the notion that moral preferences often exhibit non-compensatory and sharply nonlinear structure.

As with utilitarians, the model's predictions do not depend on a specific functional form, but the following quadratic specification is used to motivate the required properties of the deontological moral cost function:

$$\chi(\phi_i, \Delta\lambda_i) = \frac{1}{2} b_1 \lambda \phi_i^2 - I_i^N b_2 \frac{\Delta\lambda_i}{\lambda} \phi_i \tag{2}$$

with $b_1, b_2 > 0, b_1 > b_2$, and $I_i^N = 1$ if i is a Nash agent, 0 otherwise.

The first term reflects intensifying blame with greater complicity (ϕ_i) and harm (λ). Deontological ethics prohibit investors from justifying wrongful acts by appealing to good consequences. Therefore, deontologists do not participate in activism, and their disutility scales with *ex-ante* externality levels that would prevail absent any activism.

The second term captures moral relief from reduced externality due to activism by others, and is included for Nash but not Kant deontologists. This is consistent with the logic of each strategic imperative; Kant agents cannot claim moral relief from outcomes they did not will universally, but Nash agents who take circumstances as given can do so without contradiction.

The expression $\frac{\Delta\lambda_i}{\lambda}\phi_i$ in the second term represents the fraction of one's original complicity (measured by ϕ_i) that is mitigated by the share of externality successfully offset. Note that the moral relief experienced by Nash deontologists is strictly limited: the weight on this term, b_2, is always less than the complicity penalty weight b_1, ensuring that moral relief can never eliminate the primary cost of complicity, which is the primary marker of a deontological agent. This structure reflects the Rossian logic in deontology: it permits limited moral tradeoffs without abandoning the core deontological imperative. Even in the best-case scenario—where all generated externalities are offset—complicity is not erased, only reduced.

The model is agnostic to the exact functional form of $\chi(\cdot)$ as long as it adheres to the qualitative properties outlined above.

2.3 Coordination Regimes and Offset Rule

The economy consists of a representative firm and four representative investors, one of each type: Nash Utilitarian (NU), Nash deontologist (ND), Kant Utilitarian (KU), and Kant deontologist (KD). Let I_i^τ be an indicator function that takes the value 1 if i is of type τ and 0 otherwise. Let w_i represent investor i's initial wealth endowment.

Investors allocate a fraction ϕ_i of their endowed wealth to the firm's stock and the remainder to a risk-free asset yielding gross return R_f. Utilitarian investors who hold the firm's equity exert pressure on the firm to reduce its externality generated, in this example carbon emissions. In response, the firm purchases carbon offsets at a fixed price q per ton to reduce emissions by $\Delta\lambda \in [0, \lambda]$. The firm is a price-taker in the offsets market. The purchase of these offsets subtracts from the cash flow available to all investors. There is no direct cost to investors who influence the firm except the reduced return on invested capital, which is borne by activist and non-activist shareholders alike.

The offset quantity is assumed to increase linearly with the level of utilitarian support. Two canonical coordination regimes translate individual holdings into collective abatement: share-weighted coordination, in which one share, one vote applies and offsets are proportional to utilitarian capital as a fraction of firm value; and headcount-weighted coordination, in which one agent, one vote applies and offsets are proportional to utilitarian headcount relative to total investors in the firm. This paper adopts a convex combination of the two:

$$\Delta\lambda_i = \left(\mu \frac{\sum_j I_j^U \tilde{\phi}_j^i w_j}{P} + (1-\mu) \frac{\sum_j I_j^U I(\tilde{\phi}_j^i > 0)}{n} \right) \lambda. \tag{3}$$

In this expression, $\mu \in [0,1]$ allows for tractable comparative statics on the relative influence of capital versus participation; I_j^U is an indicator function that equals 1 if j is a utilitarian investor and 0 otherwise; $\tilde{\phi}_j^i$ is the portfolio weight of investor j as perceived by investor i; and $I(\tilde{\phi}_j^i > 0)$ is an indicator function that equals 1 if investor j holds a positive portfolio weight in the risky asset as perceived by investor i, and 0 otherwise.

While corporate voting is almost universally one-share-one-vote (OSOV), there is an active debate about broadening the voice in stewardship. Large asset managers (BlackRock, Vanguard, State Street) experiment with client-directed voting and pass-through programs. Coalition platforms such as Principles of Responsible Investment (PRI) compete on participation counts and promote disclosure norms that emphasize how many investors supported a resolution rather than only what percentage of shares did. This paper views the debate as one about raising the salience of breadth. Accordingly, μ may be read as a reduced-form interpretation of that salience, used by the model to anticipate how democratizing influence would reshape activism outcomes even in systems that remain formally one-share-one-vote.

Given these ingredients—preferences, beliefs, coordination technology, and offset rules—the investor's problem can now be formally defined.

Definition 3 (Investor Problem)**.** *Given market price P of the firm's stock, distribution of gross cash flow θ, pre-activism externality level λ, and the portfolio weights of all other investors $\boldsymbol{\phi}_{-i}$, investor i's problem is to choose a portfolio weight ϕ_i that solves*

$$\max_{\phi_i \in [0,1]} \mathbb{E}\left[u(w_i')\right] - I_i^U v(\lambda - \Delta\lambda_i) - (1 - I_i^U)\chi\left(\phi_i, \Delta\lambda_i\right), \tag{4}$$

subject to

$$w_i' = w_i\left((1-\phi_i)R_f + \phi_i R\right), \tag{5}$$

$$R = \frac{\theta - q\Delta\lambda}{P}, \tag{6}$$

$$\Delta\lambda_i = \left(\mu \frac{\sum_j I_j^U \tilde{\phi}_j^i w_j}{P} + (1-\mu) \frac{\sum_j I_j^U I(\tilde{\phi}_j^i > 0)}{n} \right) \lambda, \tag{7}$$

$$\tilde{\phi}_j^i = \begin{cases} \phi_j, & \text{if } I_i^N = 1, \\ \phi_i, & \text{otherwise}, \end{cases} \tag{8}$$

where all variables and functions are as defined previously.[3]

2.4 Optimal Portfolios

The first-order conditions (FOC) for the investor problem reflect the trade-off between pecuniary and moral considerations:

$$\mathbb{E}\left[\frac{\partial u(w_i')}{\partial w_i'}\frac{\partial w_i'}{\partial \phi_i}\right] = \begin{cases} \frac{\partial v(\lambda - \Delta\lambda_i)}{\partial \Delta\lambda_i}\left(\frac{\partial \Delta\lambda_i}{\partial \phi_i}\right)_{NU} & \text{if } i \in \text{Nash-Util}, \\ \frac{\partial \chi(\phi_i, \lambda - \Delta\lambda_i)}{\partial \phi_i} & \text{if } i \in \text{Nash-Deon}, \\ \frac{\partial v(\lambda - \Delta\lambda_i)}{\partial \Delta\lambda_i}\left(\frac{\partial \Delta\lambda_i}{\partial \phi_i}\right)_{KU} & \text{if } i \in \text{Kant-Util}, \\ \frac{\partial \chi(\phi_i, \lambda)}{\partial \phi_i} & \text{if } i \in \text{Kant-Deon}. \end{cases} \tag{9}$$

where $\frac{\partial \Delta\lambda_i}{\partial \phi_i}$ is the marginal effect of increasing portfolio weight on the investor's perceived reduction in externality levels.

Each investor computes the marginal pecuniary benefit of exposure to the risky asset (increased expected return adjusted for risk, LHS above) against the marginal moral cost of that exposure (either from residual externality levels or complicity, RHS above). Note that the marginal pecuniary benefit of risky asset exposure is structurally identical across investor types, and varies only in magnitude based on the investor's risk aversion and wealth. The marginal moral cost, however, differs qualitatively by both moral and strategic imperatives.

Proposition 1. *Let C_τ be the marginal moral cost of increasing portfolio weight for an agent of type $\tau \in \{NU, ND, KU, KD\}$. The ordering of C_τ by type for any given level of ϕ_i is*

$$C_{KU}(\phi_i) < C_{NU}(\phi_i) < 0 < C_{ND}(\phi_i) < C_{KD}(\phi_i). \tag{10}$$

Proof. See Appendix Section A. □

Corollary 1 (Participation Incentives by Investor Type)**.** *Let ϕ_τ^* be the optimal portfolio weight of an agent of type $\tau \in \{NU, ND, KU, KD\}$, w_τ the agent's wealth, γ the agent's risk aversion, and μ the weight on the share-based coordination regime. Under the assumption that w_τ and γ are identical across types, the ordering of ϕ_τ^* by type for any given level of γ and μ is*

$$\phi_{KU}^* \geq \phi_{NU}^* \geq \phi_{ND}^* \geq \phi_{KD}^*. \tag{11}$$

Proof. See Appendix Section A. □

Absent any complicity penalty, utilitarians increase their holding of the firm's stock until pecuniary risk outweighs moral benefit from externality reductions. Deontologists, saddled with a complicity penalty, only participate if pecuniary benefits outweigh moral complicity costs. The complicity cost,

[3] $\Delta\lambda_i$ reflects the investor's perceived externality reduction under their epistemic frame, while $\Delta\lambda$ (without the subscript) is the true externality reduction resulting from the collective actions of all utilitarian investors. For Nash investors, these are identical.

Table 2: Parameterization for Simulation Results

Parameter	Value	Rationale
Primitives		
n	4	Number of agents (one rep for each type)
θ	1	Aggregate output of firm (normalized)
w_i	1	Initial wealth per investor (normalized)
Empirically Calibrated		
λ	0.0045	Based on global avg. 1.5 tons per \$1000 output
$\mathbb{E}(R)$	1.1	Long-run mean of gross return
$\mathbb{V}(R)$	0.15	Long-run variance of return
R_f	1.02	Long-run mean of gross risk-free rate
Contextually Calibrated		
a	$\sim 10^6$	Moral disutility 10–25% of pecuniary
b_1	$\sim 10^3$	Moral disutility 10–25% of pecuniary
b_2	$\sim 10^{-2}$	Complicity alleviation less than b_1 by construction
Baseline Benchmarks		
μ	0.5	Share-based coordination weight
γ	4	Risk aversion parameter

Note: While broadly consistent with empirical observation, parameterization choices are meant for illustrating the model's analytical predictions, and are not meant to reflect precise real-world estimates.

being invariant to the actions of others for Kant investors but not Nash investors, causes Kant deontologists to be the least incentivized to participate.

Based on this hierarchy of participation incentives, participation from each investor type can be mapped out in the model's parameter space. Simulating the model using CRRA utility from wealth and the parameterization summarized in Table 2, Fig. 1 maps each investor's optimal portfolio weight in the $\gamma \times \mu$ space. The figure illustrates that Kant utilitarians hold the highest portfolio weight for any given level of γ and μ, while Kant deontologists are the key bottleneck for broad-based participation.

Note, further, the difference in investor responses to the coordination regime. Utilitarians exhibit an increase in holdings with μ, the weight on share-based coordination, albeit with diminishing sensitivity and eventual reversal. On the other hand, deontological investors do not partake in activism and are thus insensitive to the coordination regime in the direct sense. However, strategic interdependence and price feedback effects (discussed later) can induce indirect sensitivity to μ. Deontological investor participation appears to first decrease and then increase with μ. The relationship between the coordination regime, type-specific participation and equilibrium abatement levels is complex and is explored further in Section 3.2.

2.5 Strategic Interdependence

The universalizability principle precludes Kant investors from experiencing any strategic interdependence in their activism. In contrast, Nash investors can experience strategic complementarity or substitutability, that is, their marginal incentive to participate changes with the participation of others.

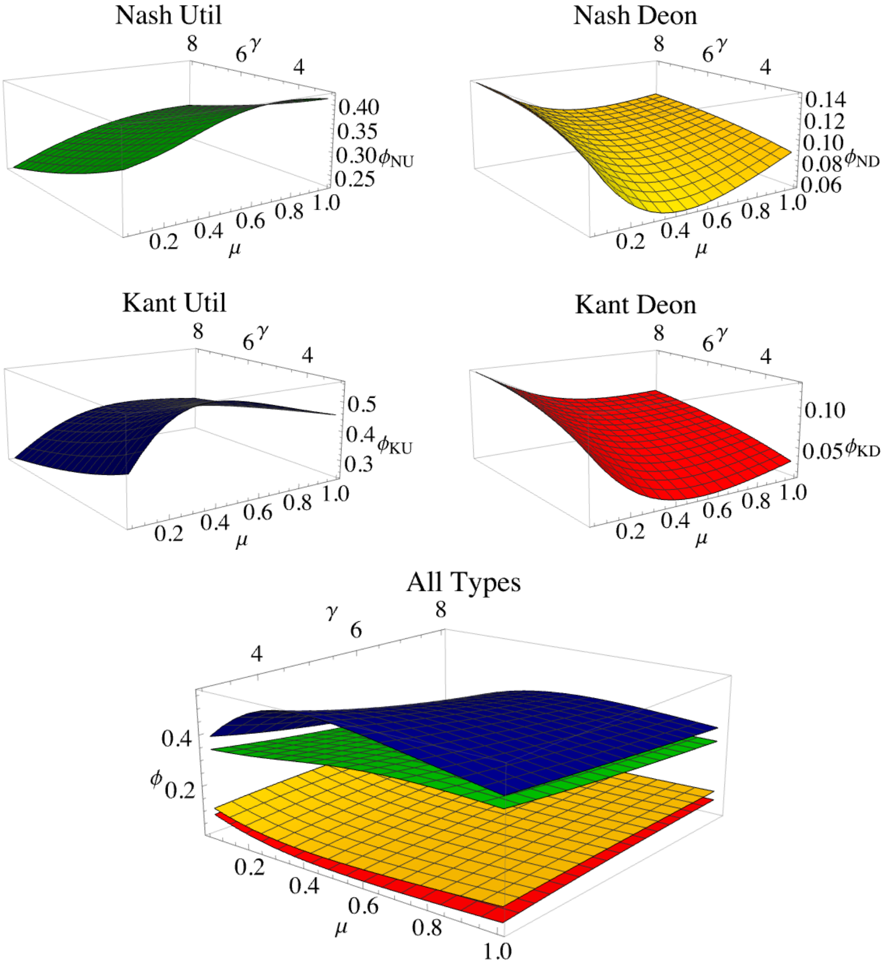


Figure 1: Optimal Portfolio Weight by Investor Type as a Function of Risk Aversion and Coordination Regime

Note: Vertical axis represents portfolio weight ϕ_i^* for each investor type. Horizontal axes: share-based coordination regime weight μ, and risk aversion parameter γ. Source: Simulation using parameterization from Table 2; view as directionally indicative rather than quantitatively precise.

Proposition 2 (Strategic Interdependence by Investor Type)**.** *An increase in the portfolio weight of investor j increases (decreases) investor i's marginal moral cost of raising their own portfolio weight if investor i is a Nash utilitarian (Nash deontologist). Kant investors experience no strategic interdependence. Table 3 summarizes.*

Proof. See Appendix Section A. □

Table 3: First Derivative and Cross Partial of Moral Cost by Investor Type

Investor Type	Marginal Moral Cost	Cross Partials
Nash Utilitarian (NU)	$\frac{\partial v}{\partial \Delta\lambda_i}\frac{\partial \Delta\lambda_i}{\partial \phi_i} < 0$	$\frac{\partial^2 v}{\partial \Delta\lambda_i^2}\frac{\partial \Delta\lambda_i}{\partial \phi_i}\frac{\partial \Delta\lambda_i}{\partial \phi_j} > 0$
Nash Deontologist (ND)	$\frac{\partial \chi}{\partial \phi_i} + \frac{\partial \chi}{\partial \Delta\lambda_i}\frac{\partial \Delta\lambda_i}{\partial \phi_i} > 0$	$\frac{\partial^2 \chi}{\partial \phi_i \partial \Delta\lambda_i}\frac{\partial \Delta\lambda_i}{\partial \phi_j} < 0$
Kant Utilitarian (KU)	$\frac{\partial v}{\partial \Delta\lambda_i}\frac{\partial \Delta\lambda_i}{\partial \phi_i} < 0$	$\frac{\partial^2 v}{\partial \Delta\lambda_i^2}\frac{\partial \Delta\lambda_i}{\partial \phi_i}\frac{\partial \Delta\lambda_i}{\partial \phi_j} = 0$
Kant Deontologist (KD)	$\frac{\partial \chi}{\partial \phi_i} > 0$	$\frac{\partial^2 \chi}{\partial \phi_i \partial \Delta\lambda_i}\frac{\partial \Delta\lambda_i}{\partial \phi_j} = 0$

Note: Cross partials reflect the effect of an increase in investor j's portfolio weight on investor i's marginal moral cost of increasing their own portfolio weight. Derivations in Appendix Section A.

For Nash deontologists, reduced externalities from others' activism reduces their marginal complicity cost. This increases their marginal incentive to hold the asset, all else equal, and reflects strategic complementarity in holdings.[4]

Nash utilitarians, on the other hand, face a lower marginal gain from increasing holdings when others increase theirs, reflecting strategic substitutability. Holding price fixed, the convexity of the Nash utilitarian's disutility function ($v'' > 0$) means that as aggregate participation rises, the marginal benefit of the investor's own holding on externality levels (weakly) decreases.

The simultaneous presence of complementarity and substitutability in the model ensures that feedback loops between Nash utilitarians and Nash deontologists are mutually cancelling rather than reinforcing. This is a key feature of the model, as it enables uniqueness without the need for additional perturbations, and is exploited in the proof of uniqueness presented later.

2.6 Market Clearing

Let $\phi_\tau^*(P)$ be the solution for optimal portfolio weight for investor type τ at price P. Closed-form demand curves are possible under specific functional forms for utility and moral cost. For example, under CRRA utility from wealth and quadratic moral costs, closed-form solutions for $\phi_\tau^*(P)$ are provided in Appendix Section A.

Equilibrium is characterized by the portfolio weights for each type and the market clearing price P^*. The quantity of offsets purchased by the firm in equilibrium is given by substituting the optimal portfolio weights chosen by each investor into the offset rule. The market price of the firm's equity

[4] Conventionally, strategic complementarity corresponds to an increase in marginal benefit with others' increased participation. Here, as moral considerations are modeled as costs rather than benefits, complementarity corresponds to a decrease in marginal cost with others' increased participation, and vice versa for substitutability.

is determined by the market-clearing condition below, which sets demand to the unit supply of the risky asset:

$$P^* = \sum_{\tau} \phi_{\tau}^*(P^*) w_{\tau}. \tag{12}$$

The model's equilibrium is pinned down by a fixed point in demand:

$$P^* = \Phi(P^*), \tag{13}$$

where $\Phi(P^*)$ is defined as

$$\Phi(P^*) = \sum_{\tau} \phi_{\tau}^*(P^*) w_{\tau}. \tag{14}$$

3 Equilibrium

Definition 4 (Static Equilibrium with Heterogeneous Reasoning)**.** *A static equilibrium consists of a portfolio profile $\boldsymbol{\phi}^* = (\phi_1^*, \ldots, \phi_n^*) \in [0,1]^n$ and a market-clearing price $P^* \in \mathbb{R}_+$ such that:*

1. Type Consistent Optimization*: For each agent i, the portfolio weight ϕ_i^* solves the agent's maximization problem given their moral and strategic imperative: Nash agents maximize utility taking ϕ_{-i}^* as given; Kant agents optimize as if all agents were to adopt ϕ_i^*.*

2. Perception Consistency*: Each agent's perceived externality reduction $\Delta\lambda_i\left(\boldsymbol{\phi}^*, P^*\right)$ is consistent with their type-specific logic: Nash agents evaluate the true offset based on observed weights; Kant agents assess the externality offset under the universal adoption of their own action.*

3. No Regret*: Given their respective reasoning frameworks, no agent has a profitable deviation from ϕ_i^*, that is:*

$$\phi_i^* \in \arg\max_{\phi_i \in [0,1]} \; \mathbb{E}[u(w_i')] \; - \; I_i^U \cdot v(\lambda - \Delta\lambda_i) \; - \; \left(1 - I_i^U\right) \cdot \chi(\phi_i, \Delta\lambda_i)\,. \tag{15}$$

4. Market Clearing: *The market-clearing condition holds:*

$$P^* \; = \; \sum_{i=1}^{n} \phi_i^* \, w_i. \tag{16}$$

In equilibrium, no investor—Kant or Nash—can profitably change her portfolio weight given the decision rule she actually uses. We therefore treat equilibrium as a fixed point in deviations, not necessarily in beliefs. Kant investors act on their internalized counterfactual offset function and, by construction, will not change their action even if the true aggregate offset differs. Nash investors optimize against the objective offset and will not deviate so long as their perceived offset equals reality. Equilibrium does not require belief–truth alignment, only internal consistency and no profitable deviation.

Proposition 3 (Existence and uniqueness of equilibrium)**.** *Under the following assumptions:*

1. There exists a representative agent for each type with a strictly concave, continuously differentiable utility function combining pecuniary and moral considerations.

Table 4: Comparative Statics: Direction of Change for Each Type

Parameter	ϕ_{NU}	ϕ_{ND}	ϕ_{KU}	ϕ_{KD}	P	$\Delta\lambda$
Risk aversion γ	$\downarrow$	$\downarrow^{*}$	$\downarrow$	$\downarrow^{*}$	$\downarrow$	$\downarrow$
Util sensitivity a	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\uparrow$
Deon complicity b_1	–	$\downarrow$	–	$\downarrow$	$\downarrow$	$\uparrow$
Deon relief b_2	–	$\uparrow$	–	–	$\uparrow$	–
Offset price q	$\downarrow$	$\downarrow^{*}$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Project risk $\mathbb{V}(\theta)$	$\downarrow$	$\downarrow^{*}$	$\downarrow$	$\downarrow^{*}$	$\downarrow$	$\downarrow$
Share-Coordination Weight μ	$\downarrow\uparrow^{\dagger}$	$\downarrow\uparrow^{\dagger}$	$\downarrow\uparrow^{\dagger}$	$\downarrow\uparrow^{\dagger}$	$\downarrow\uparrow^{\dagger}$	$\uparrow$

*May locally reverse via price feedback. †Non-monotonic, effect depends on population composition and distribution of wealth.

2. The moral cost functions $v(\cdot)$ and $\chi(\cdot)$ are convex and differentiable.

3. The externality offset rule $\Delta\lambda(\cdot)$ is continuous and bounded.

4. Each agent reasons according to either Nash or Kant logic.

a unique static equilibrium $(\boldsymbol{\phi}^*, P^*)$ exists.

Proof. See Appendix Section A.4. □

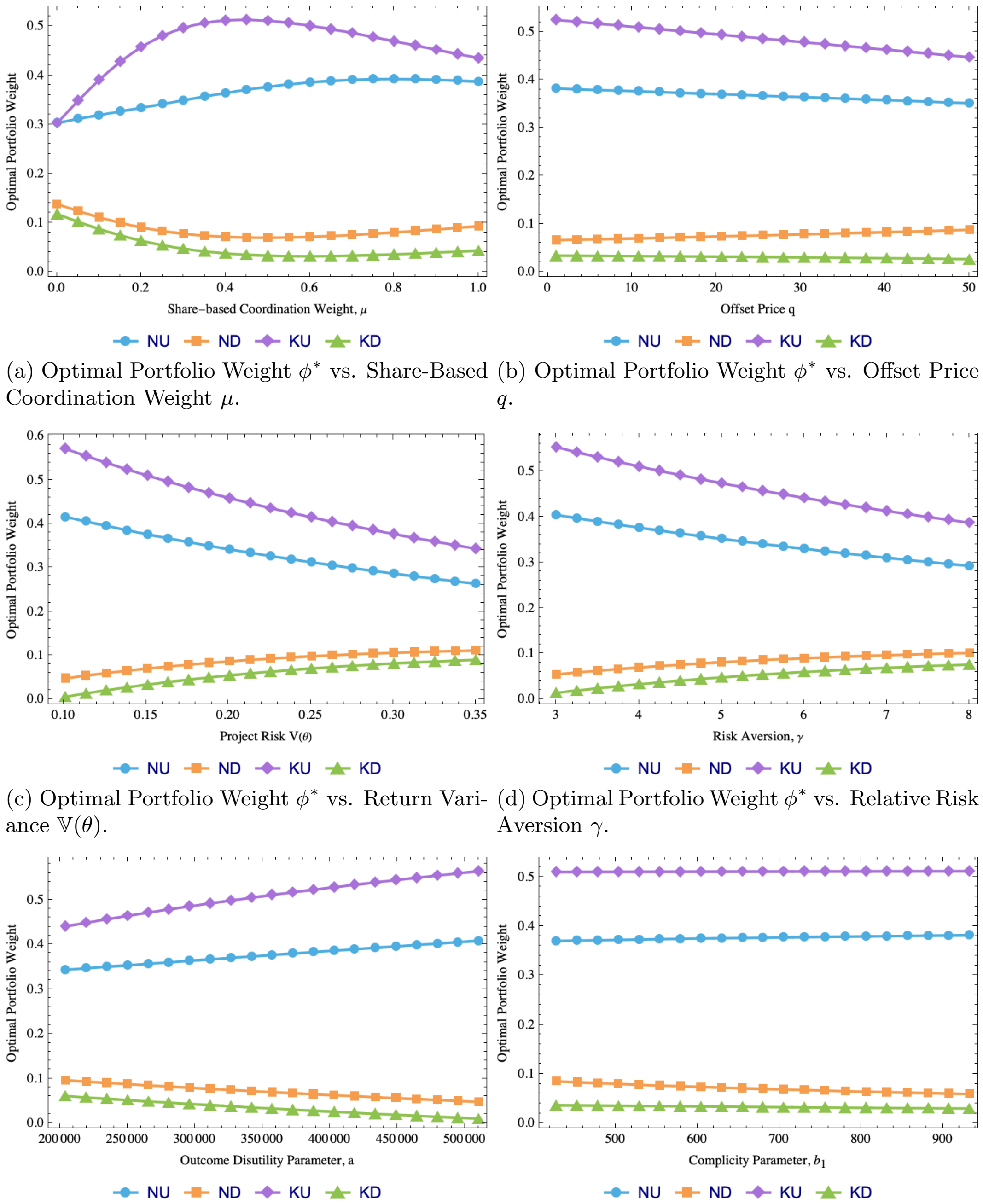
Uniqueness of equilibrium emerges from the simultaneous presence of strategic complementarity (Nash deontologists) and substitutability (Nash utilitarians), which ensures feedback loops dampen rather than amplify, while Kant agents' strategic invariance anchors expectations and stabilizes outcomes. This allows uniqueness to be achieved under broad conditions, without need for payoff uncertainty, noise, or other equilibrium selection devices.

3.1 Comparative Statics

To illustrate the model's implications concretely, I simulate equilibrium numerically with four representative investors, CRRA pecuniary utility, quadratic moral costs and baseline parameter values summarized in Table 2. Table 4 provides a summary of analytically derived comparative statics results, and Figs. 2 to 4 plot numerical illustrations for each type's optimal portfolio weight, equilibrium price and expected return, and externality reductions as each parameter is varied around the baseline. For this illustration, I assume the externality in question is carbon emissions, so that $\Delta\lambda/\lambda$ represents fraction of externality abated.

Rising risk aversion γ, offset price q, and project risk $\mathbb{V}(\theta)$ have the standard negative impact on risky asset allocations, price, and abatement: as cost and/or uncertainty rises, all types reduce their holdings, lowering equilibrium demand and Externality Reduction correspondingly. While the charts and table summarize all comparative statics, I highlight two results of particular interest below.

First, in contrast to models in which moral disutility arises from complicity (and is thus escapable via exclusion), utilitarian investors here suffer from a public bad—aggregate externality—regardless of participation. As the moral cost parameter a rises, their only means of reducing disutility is to



(a) Optimal Portfolio Weight ϕ^* vs. Share-Based Coordination Weight μ.

(b) Optimal Portfolio Weight ϕ^* vs. Offset Price q.

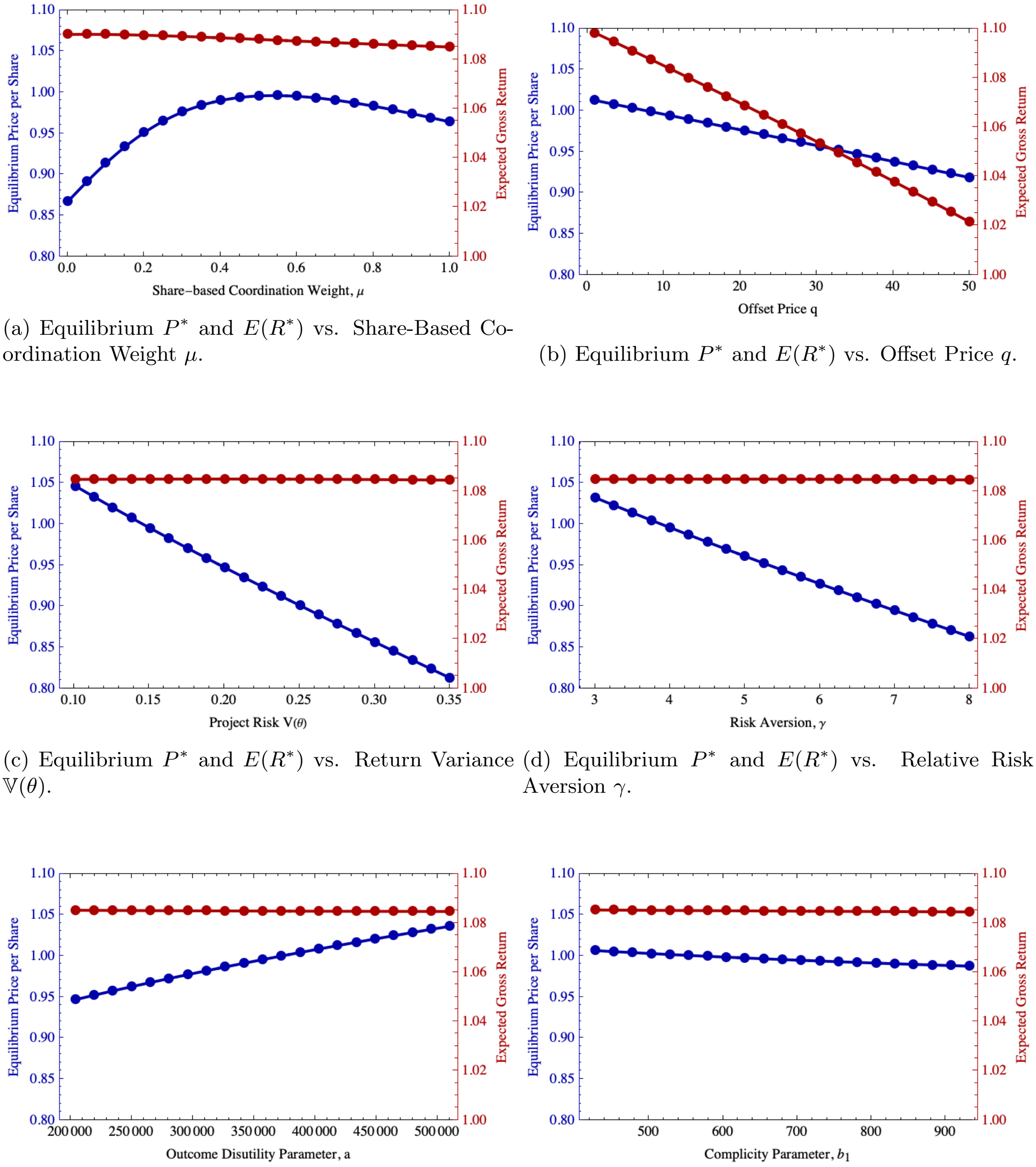
(c) Optimal Portfolio Weight ϕ^* vs. Return Variance $\mathbb{V}(\theta)$.

(d) Optimal Portfolio Weight ϕ^* vs. Relative Risk Aversion γ.

(e) Optimal Portfolio Weight ϕ^* vs. Outcome Sensitivity a.

(f) Optimal Portfolio Weight ϕ^* vs. Complicity Sensitivity b_1.

Figure 2: Optimal Portfolio Weight Comparative Statics.



(a) Equilibrium P^* and $E(R^*)$ vs. Share-Based Coordination Weight μ.

(b) Equilibrium P^* and $E(R^*)$ vs. Offset Price q.

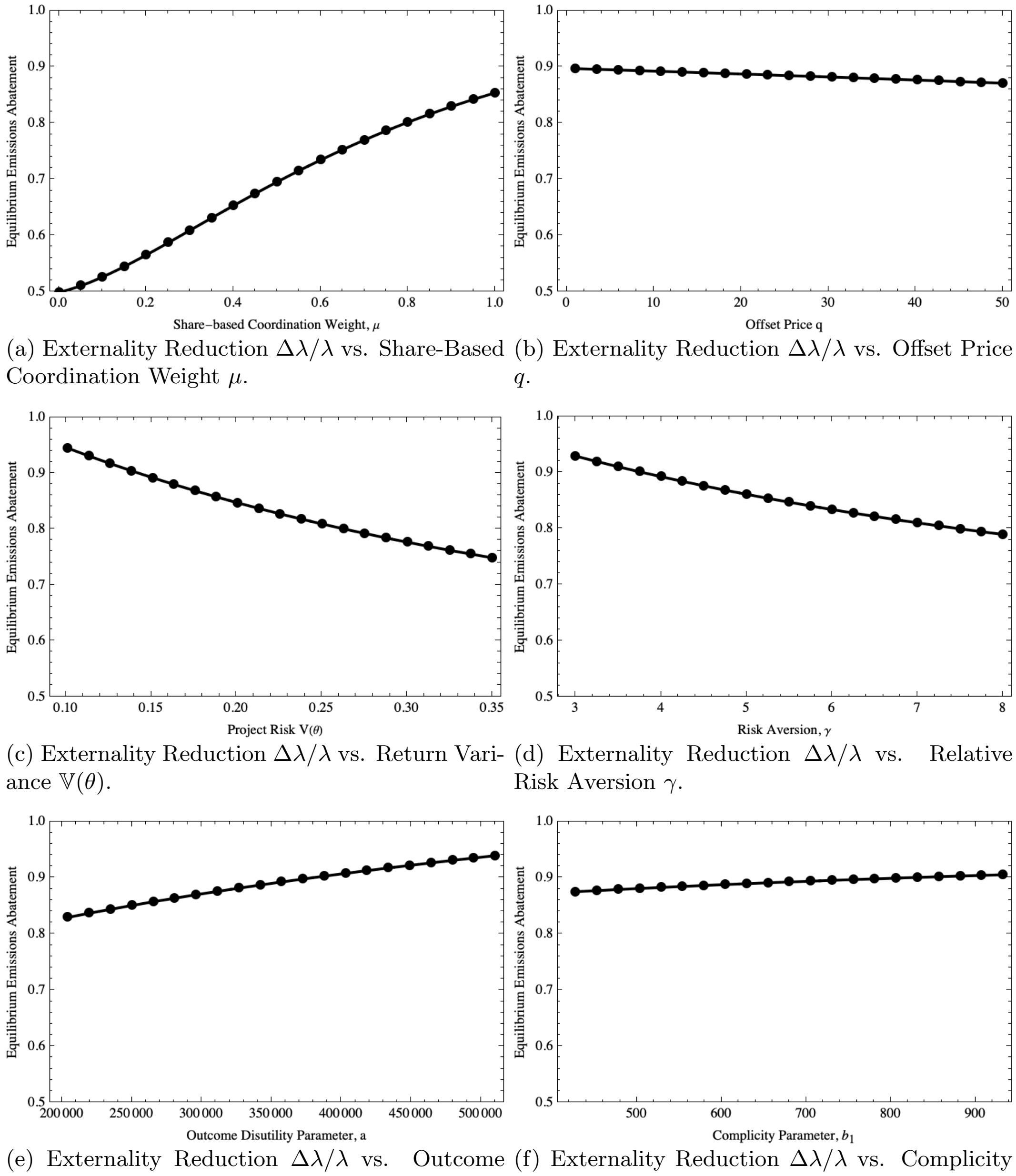
(c) Equilibrium P^* and $E(R^*)$ vs. Return Variance $\mathbb{V}(\theta)$.

(d) Equilibrium P^* and $E(R^*)$ vs. Relative Risk Aversion γ.

(e) Equilibrium P^* and $E(R^*)$ vs. Outcome Sensitivity a.

(f) Equilibrium P^* and $E(R^*)$ vs. Complicity Sensitivity b_1.

Figure 3: Equilibrium Price and Gross Expected Return Comparative Statics.



(a) Externality Reduction $\Delta\lambda/\lambda$ vs. Share-Based Coordination Weight μ.

(b) Externality Reduction $\Delta\lambda/\lambda$ vs. Offset Price q.

(c) Externality Reduction $\Delta\lambda/\lambda$ vs. Return Variance $\mathbb{V}(\theta)$.

(d) Externality Reduction $\Delta\lambda/\lambda$ vs. Relative Risk Aversion γ.

(e) Externality Reduction $\Delta\lambda/\lambda$ vs. Outcome Sensitivity a.

(f) Externality Reduction $\Delta\lambda/\lambda$ vs. Complicity Sensitivity b_1.

Figure 4: Externality Reduction Comparative Statics.

increase activism (i.e., risky asset demand) to induce externality abatement. In equilibrium, this raises the risky asset's price, partially offsetting the increased demand, but does not reverse the direction of the effect: risky asset demand, price, and abatement all rise with utilitarian moral sensitivity.

Second, raising the deontological complicity penalty b_1 induces both Nash and Kant deontologists to reduce their risky asset holdings, as expected. More importantly, total market-clearing requires that the shares vacated by deontologists are absorbed by utilitarians, causing the utilitarian ownership fraction to rise, and with it, equilibrium externality abatement. While the positive effect of increased deontological complicity on abatement in the model is a mechanical consequence of the share-weighted abatement rule and fixed supply, this logic reflects a plausible real-world dynamic. If divestment by less-committed (deontological) investors is not immediately offset by entry of neutral or amoral buyers, activist investors (utilitarians) can end up controlling a larger fraction of shares, increasing their capacity to influence firm behavior. Thus, the model highlights a channel by which selective exit can—counterintuitively—strengthen activist impact via concentration of voting power among activist types.

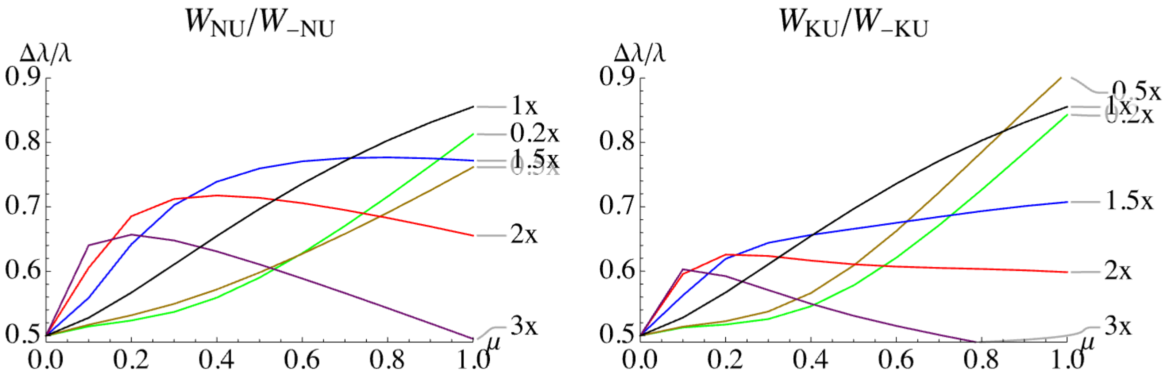
3.2 Wealth Distribution and Coordination Dynamics

The coordination mechanism and offset rule are crucial determinants of equilibrium outcomes. To explore their interaction with wealth distribution, I simulate equilibrium abatement across a continuum of coordination regimes, from pure headcount ($\mu = 0$) to pure share-based ($\mu = 1$), under varying wealth concentrations among investor types. The analysis generates novel insights, summarized in the proposition below and expanded upon thereafter.

Proposition 4. *Wealth distribution across types modulates the relationship between coordination regime (share-based vs. headcount-based influence) and equilibrium externality reduction. Specifically:*

1. *When wealth is concentrated among deontologists, share-weighted influence maximizes abatement.*
2. *When wealth is concentrated among utilitarians (Nash or Kant), neither headcount- nor share-weighted influence necessarily maximizes abatement; a coordination mechanism that allows for both individual- and wealth-based influence may be optimal.*
3. *Wealth concentration within utilitarian groups bends the relationship between coordination regime and abatement itself, not just abatement levels.*

The analysis of the previous section assumes that wealth is homogeneous across investor types. In such an economy, as the share-based coordination weight μ increases, the influence of each utilitarian-held share on abatement rises, making activism more potent. Yet the market price of the risky asset also increases with demand, as does the total cost incurred to offset the externality, generating diminishing returns to participation (Fig. 2a). Both utilitarian types—Nash and Kant—face this price feedback, but differ in how quickly their moral cost scales with μ. The Kant investor's universalizability test amplifies the perceived gains from abatement, but also accelerates saturation as prices rise and the marginal moral return flattens. Nash utilitarians experience a steadier increase that, depending on wealth distribution, may or may not saturate within the feasible range of μ.



(a) Nash-utilitarian wealth varied from 0.2x–3x baseline

(b) Kant-utilitarian wealth varied from 0.2x–3x baseline

Figure 5: Equilibrium abatement across coordination regimes and wealth concentrations.

With homogeneous endowments, greater share-weighting ($\mu \uparrow$) unambiguously increases aggregate abatement (Fig. 4a). Share-based coordination allows committed investors to exert influence in proportion to their willingness to bear cost, without redistributing power across types. When every agent has equal economic power, share-based mechanisms neither generate inequity nor undermine collective action—they simply harness the full firepower of moral investors. Coordination failures arise only when wealth concentration is misaligned with moral motive, a theme explored below.

Fig. 5 illustrates the impact of wealth heterogeneity on the relationship between equilibrium Externality Reduction and coordination regimes. Each investor type is, in turn, endowed with a multiple of the baseline wealth, while the wealth of other types is held constant at baseline. Wealth multiples range from 0.2x to 3x of baseline wealth. The resulting equilibrium abatement is computed for various values of the share-based coordination weight μ, ranging from 0 (pure headcount) to 1 (pure share-based). Note that varying deontological wealth does not shift the relationship between μ and abatement, and the relationship in this case is identical to the homogeneous wealth distribution case labelled 1x in both panels of Fig. 5.

It is important to separate how wealth distribution affects abatement *levels* from how it affects the *shape* of the abatement–coordination regime relationship. I address the former first.

Level effects. When utilitarians—Nash or Kant—are wealthy (curves 2x and 3x in Fig. 5), a purely headcount regime fails to leverage their full potential. With $\mu = 0$, maximum abatement is limited by utilitarians' population share (50% in this illustration). As μ rises, abatement improves but faces diminishing returns as price feedback dampens participation. The abatement-maximizing μ balances leveraging utilitarian wealth against overpaying for abatement; formally, it is the μ that maximizes equilibrium abatement given endogenous prices, and lies between 0 and 1.

When deontologists are wealthy (corresponding to utilitarian curves 0.2x and 0.5x), the pattern reverses. Because deontologists abstain from activism, headcount weighting dilutes utilitarian influence. Increasing μ then reallocates decision weight toward moral investors and generally raises abatement. Only when utilitarian wealth is extremely low does headcount weighting outperform share weighting. In short, the effectiveness of a plutocratic coordination regime hinges on whether moral conviction and financial power are aligned.

Shape effects. The *shape* of the abatement–μ relationship depends on which type holds the wealth. If deontological wealth were varied in the simulation instead of utilitarian wealth, the abatement–μ relationship remains unchanged from the homogeneous wealth distribution case. Doubling utilitarian wealth relative to deontological wealth, for example, is not equivalent to halving deontological wealth relative to utilitarian wealth, in terms of its effect on the *curvature* of the abatement–μ relationship. Therefore, relative wealth concentration among types is not a sufficient statistic for predicting how abatement may respond to changes in the coordination regime, even if it is sufficient for predicting abatement *levels*.

Minority catalysis. Kant utilitarians can catalyze abatement even when financially small. In Fig. 5b, curves 0.2x and 0.5x show that as $\mu \to 1$, equilibrium abatement exceeds not only what would be possible in the baseline case, but also what would be possible if Kant utilitarians themselves were wealthy.

Although counterintuitive at first glance, this finding is a direct consequence of the strategic invariance of Kant utilitarian reasoning. Kant utilitarian participation influences equilibria in two ways. First, their strategic invariance stabilizes expectations and actions among Nash utilitarians. By committing to a fixed level of participation that is invariant to the choices of other agents, Kant utilitarians provide a reliable anchor for Nash utilitarians' beliefs about aggregate activism. This anchoring effect reduces uncertainty and strategic volatility, encouraging Nash utilitarians to increase their own participation in response to the predictable activism of Kant utilitarians.

Second, Kant utilitarians, on account of their universalization logic, tend to participate more aggressively than Nash utilitarians for a given level of wealth and price. Their universalization logic causes them to internalize the positive externality of their activism more fully than Nash utilitarians, who may view their impact as diluted by non-participation. However, their aggressive participation also carries the negative side-effect of driving up prices, which can dampen Nash utilitarian participation.

The net effect on total abatement depends on the relative strength of these two effects. When Kant utilitarians are relatively poor, the first effect—stabilizing and encouraging Nash utilitarian participation—tends to dominate as Kant utilitarians lack the financial clout to significantly raise prices. This leads to higher overall abatement as Nash utilitarians respond positively to the predictable activism of Kant utilitarians. On the other hand, when Kant utilitarians are wealthy, their aggressive participation can substantially increase prices, which may discourage Nash utilitarians from participating as much. It is in this scenario that aggregate abatement falls short of the level that would be achieved under a more balanced wealth distribution.

While wealthy Nash utilitarians can also influence expectations and actions of other Nash utilitarians, their impact is less pronounced than that of their Kant counterparts. As a result, Externality Reduction levels are generally higher when Nash utilitarians are wealthy compared to when Kant utilitarians are wealthy, all else equal. However, without Kant utilitarians in the mix, the system lacks the stabilizing influence that leads to a unique equilibrium.

In sum, Kant utilitarians can serve as catalysts for change even when they lack majority wealth or voting power, provided the coordination regime allows their influence to be felt. Ironically, while the Kant agent's strategic invariance provides stability and predictability that can enhance collective action, their rigidity also prevents them from adapting to changing circumstances. When they are wealthy, for example, they participate more aggressively than would be optimal given the responses of others, leading to higher prices, reduced participation from others, and ultimately lower abatement levels than would be possible with a more flexible strategy.

Taken together, these results underscore that equilibrium abatement depends not just on investor preferences, reasoning styles or wealth levels, but also on coordination mechanisms and price feedback, each of which in turn varies in efficacy depending on how wealth is distributed. The welfare and financial-market implications of impact-oriented investing cannot be inferred from investor-level reasoning or partial equilibrium logic, or indeed, even from coordination technology alone. General equilibrium effects—driven by utility curvature, price feedback and strategic dependencies—fundamentally shape the curvature of otherwise intuitive relationships, and can even reverse their direction.

3.3 Testable Predictions

The model developed thus far generates several empirically falsifiable predictions on themes of central interest in finance and political economy. It highlights the conditions under which a small minority of principled investors can catalyze large-scale change, the mechanisms through which shareholder stewardship and coordinated engagement translate into real outcomes, and the implications of wealth concentration for collective action and welfare. The following testable hypotheses link individual moral orientation to market-level outcomes, and provide a roadmap for empirical validation.

Hypothesis 1 (Composition/Imprint). *Across otherwise comparable funds, portfolio composition differs systematically by moral type: engagement-oriented (utilitarian) funds allocate more to firms with higher ex-ante externality levels than exclusion-oriented (deontological) funds, consistent with an intent to improve rather than avoid.*

Hypothesis 2 (Coordination Mechanism). *Following firm-specific coordination shocks that increase the salience or feasibility of collective engagement (e.g., credible public commitments, coalition focal points), engagement-oriented funds increase holdings of target firms relative to comparable non-target firms and relative to non-engagement-oriented funds.*

Hypothesis 3 (Firm Outcomes). *Conditional on pre-trends and firm fundamentals, greater ownership by engagement-oriented funds is associated with subsequent improvements in firm-level externality outcomes relative to otherwise similar firms.*

It is worth emphasizing that the above hypotheses are joint predictions of the model under the assumption that wealth is homogeneously distributed across types. However, wealth concentration among different types can materially alter the efficacy of coordination regimes, as explored in the following hypotheses.

Hypothesis 4 (Coordination Plutocracy vs Democracy). *Plutocratic coordination regimes amplify moral impact when wealth is homogeneously distributed or when wealth is concentrated among exclusion-*

oriented investors. If wealth is concentrated among engagement-oriented investors, the optimal coordination is a combination of headcount and share-weighting.

Hypothesis 5 (Strategic Voting Behavior)**.** *Holding beliefs and incentives constant, Kant and Nash investors take measurably different actions (e.g., pre-declared, principle-consistent voting versus contingent, best-response voting).*

Hypothesis 6 (Minority Catalysis)**.** *A minority of universalizing (Kant) investors can catalyze coordinated real-economy change, even when they lack majority wealth or voting power.*

Hypothesis 7 (Transparency, Commitment, and Coordination)**.** *Greater transparency about investor philosophy and credible commitment mechanisms (e.g., pre-declarations, binding stewardship pledges) shift equilibria toward higher abatement via improved coordination.*

Hypothesis 8 (Dynamics and Stability)**.** *When reputation and transparency are high, Kantian activism is dynamically stable (e.g., gains AUM share over time), whereas in low-transparency environments Nash logic dominates.*

The next section tests the first of the aforesaid hypotheses, with the remaining ones deferred to future work.

4 Evidence

This section tests the empirical predictions of the model using global mutual fund data. The analysis proceeds in three steps that parallel the theoretical argument.

First, I verify that funds' stated moral orientations manifest in observable portfolio characteristics. This step confirms that deontological funds avoid exposure to controversial firms, while utilitarian funds, who claim an engagement motive, do not differ materially from neutral peers, consistent with their willingness to hold problematic firms in order to influence them.

Second, I examine whether engagement-oriented funds respond to exogenous coordination opportunities, using the staggered rollout of Climate Action 100+ (CA100+) as a natural experiment. These tests validate that funds classified as utilitarian behave in a manner consistent with the model's coordination mechanism, increasing holdings of targeted firms when collective engagement becomes more feasible.

Finally, I test whether these induced ownership changes translate into measurable improvements in firm-level externality outcomes, using a two-stage least squares (2SLS) design that instruments utilitarian ownership with both coordination shocks and flow-driven shifts in capital. Together, these steps build from intent to action to outcome, tracing the causal chain predicted by the theory.

4.1 Data and Fund Classification

The empirical analysis draws on the universe of open-ended mutual funds and exchange-traded funds (ETFs) covered by Morningstar between 2015–2025. To classify funds as deontological, utilitarian, or

Table 5: Summary statistics by fund moral type, 2015–2025 quarterly panel

	Utilitarian		Deontological		Neutral	
	Mean	SD	Mean	SD	Mean	SD
Funds	2148		602		7459	
Fund Families	278		155		729	
Funds with Matched Holdings	957		276		2151	
Portfolio Characteristics						
Actively Managed	0.48	0.50	0.51	0.50	0.58	0.49
Log AUM	19.49	2.20	18.90	2.14	19.20	2.57
Quarterly Return (Percent)	1.88	7.27	1.83	7.42	1.62	7.67
Quarterly Flow (Percent)	5.16	21.76	6.68	22.87	3.35	20.43
Growth Tilt (Percent)	28.09	18.86	28.84	20.50	25.37	21.92
Small Cap Tilt (Percent)	11.25	20.15	11.14	20.43	10.43	19.98
Log Avg Market Cap	10.15	1.49	10.34	1.53	10.02	1.44
ESG and Externality Profile						
Wtd. Avg. Carbon Intensity Scope 1, 2	219.63	596.59	176.83	385.82	223.16	457.34
Carbon Risk Score (Lower=Greener)	9.75	5.53	8.61	4.79	10.24	5.36
Env. Risk Score (Lower=Greener)	10.17	15.67	9.42	15.24	11.34	16.64
Social Risk Score (Lower=Better)	13.48	14.23	13.04	13.74	14.52	15.08
Morningstar ESG Risk Rating (Higher=Greener)	3.09	1.05	3.36	1.14	3.01	1.05
Percent AUM in Oil/Gas	1.66	4.53	1.45	3.83	2.05	5.01
Percent AUM with High/Severe Controversies	6.12	6.61	5.59	7.28	6.63	7.24
Observations	55566		13312		252259	

neutral, I collect all U.S. Securities and Exchange Commission (SEC) filings made by mutual funds and ETFs in the period 2015–2025, including prospectuses (Forms 497 and 485), annual reports (Forms N–CSR and N–CEN), proxy voting records and amendments (Form N–PX), and Statements of Additional Information (SAI). While the SEC covers primarily U.S.-domiciled funds, many international funds that market to U.S. investors also file with the SEC, allowing for a broader sample. For funds not covered by the SEC, I collect documentation from Bloomberg, Morningstar, London Stock Exchange Group (LSEG)/Refinitiv, and other public sources, albeit with less historical depth. For example, Morningstar and LSEG/Refinitiv report fund objective summaries that often include whether a fund employs exclusions or states an engagement objective. Classification relies only on fund documentation, taken directly from filings where possible and from data provider summaries otherwise, and does not use holdings or portfolio characteristics to avoid mechanical biases.

For each fund–quarter, fund-level characteristics and performance data are taken from Morningstar Direct at a monthly frequency. Fund holdings are obtained from the Center for Research in Security Prices (CRSP) for U.S. funds and FactSet for non-U.S. funds. Holdings are further characterized by metrics on firm-level environmental, social, and governance (ESG) characteristics from Sustainalytics (quarterly) and emissions data from Trucost (yearly).

The final sample covers approximately 3,500 funds and approximately 30,000 firms held by these funds over 2015–2025, with quarterly observations that amount to over 30 million fund–firm–quarter triplets. While fund-level data are available for over 11,000 funds, not every fund could be matched to holdings. However, analyses that only use fund-level data leverage the full sample. Summary statistics are presented in Table 5.

Funds are classified along the moral dimension into deontological, utilitarian, and neutral types based on exclusionary criteria and engagement objectives disclosed in fund documentation. Classification relies on algorithmic keyword searches followed by manual validation. Appendix Section C.1 describes the classification procedure in detail, with examples of fund language that characterize each type as well as examples of keywords used in the initial screening. Neutral funds are the fallback category for funds that do not express either an exclusion or an engagement-oriented mandate. Among funds that disclose an engagement mandate, only funds that explicitly mention social or environmental concerns, broadly defined, are classified as utilitarian. This restriction excludes funds that engage with management on purely financial or corporate governance issues, as such topics do not constitute an externality that causes ethical concern.

4.2 Portfolio Composition by Moral Type

Figures 6 and 7 provide descriptive evidence on Hypothesis 1, which predicts that deontological and utilitarian funds should differ in portfolio characteristics. Across all outcome measures, portfolio-level ESG characteristics differ by fund type. Deontological funds appear consistently "greenest," with higher Morningstar ESG ratings, lower carbon intensity, and less exposure to oil and gas or severe controversies. Utilitarian funds occupy an intermediate position: along select metrics—notably ESG scores—they are somewhat greener than neutral funds but not as green as deontological funds. On carbon risk, intensity, and controversy exposure, utilitarian funds are statistically indistinguishable from neutral funds, consistent with their strategy of retaining exposure to controversial firms in order to engage.

To formalize these descriptive differences, I estimate regressions of the form

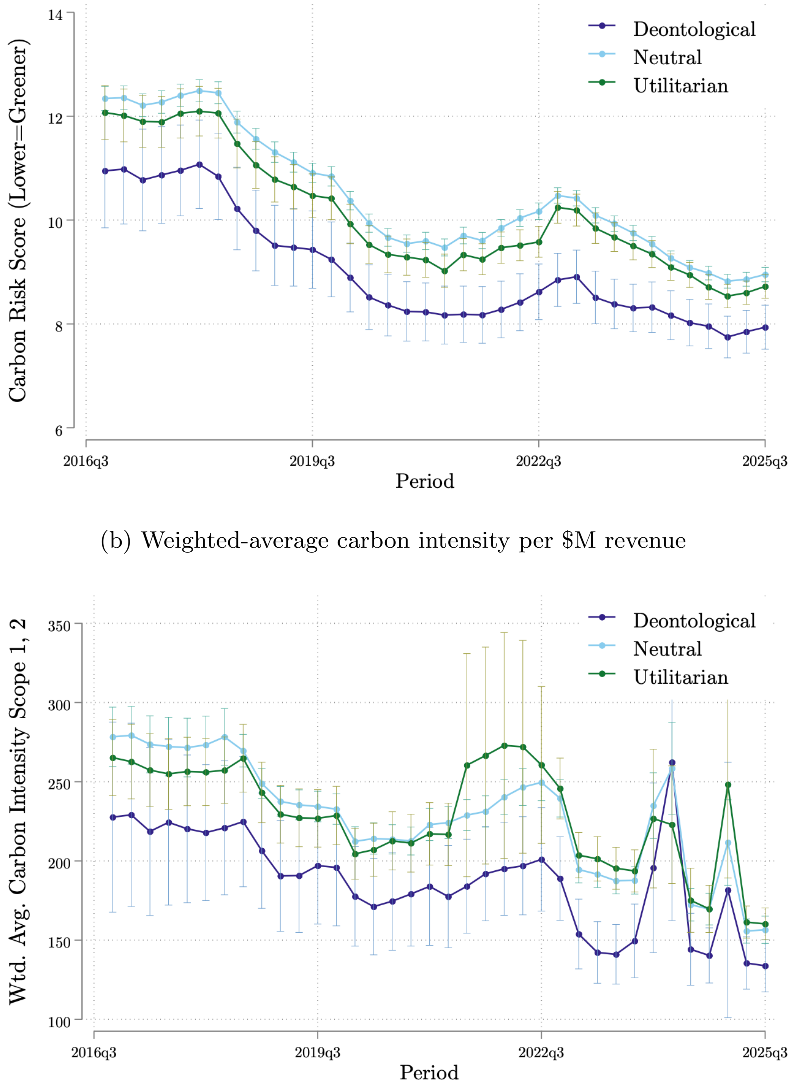
$$Y_{it} = \alpha + \beta \, \text{Type}_i + \gamma' X_{it} + \phi_{(f)t} + \psi_{(s)t} + \varepsilon_{it}. \tag{17}$$

Here, Y_{it} is a portfolio-level characteristic of fund i in quarter t. Dependent variables include the percentage of assets under management (AUM) exposed to high or severe controversies, the weighted-average carbon intensity of the portfolio, the share of AUM in oil and gas, and Morningstar's carbon risk score. The key explanatory variable, Type_i, captures pairwise differences between two fund types (for example, deontological versus neutral), with one of the types omitted as the baseline. X_{it} is a vector of time-varying controls capturing fund size (log AUM), portfolio composition (log average market capitalization of holdings, growth and small-cap tilts), and flows and returns in the current quarter.

All regressions absorb fund family-by-quarter and style-by-quarter fixed effects. A fund family refers to the parent asset management firm that sponsors the fund (for example, BlackRock or Vanguard). Including family-by-quarter fixed effects ensures that any time-varying policies, mandates, or ESG initiatives at the parent firm are accounted for. The style variable combines the fund's primary asset class (for example, equity or fixed income), its geographical and sectoral focus, its style tilt (for example, growth, value, small-cap), domicile, and region of sale. Including style-by-quarter fixed effects therefore controls for highly granular shocks to investor demand or market conditions affecting specific asset-class–region–style combinations. Identification thus relies on differences between funds of different moral types that share the same parent family, investment style, and time period. Standard errors are clustered at the fund level to allow for arbitrary serial correlation within funds.

Tables 6 and 7 present pooled regression results for pairwise comparisons of deontological versus neutral funds and utilitarian versus neutral funds, respectively. While deontological funds hold significantly greener portfolios than neutral funds across all four metrics, utilitarian funds do not differ sig-

Figure 6: Portfolio-level differences in carbon-related metrics by fund type, 2015–2025. Bars indicate 95% confidence intervals.



(a) Carbon risk score

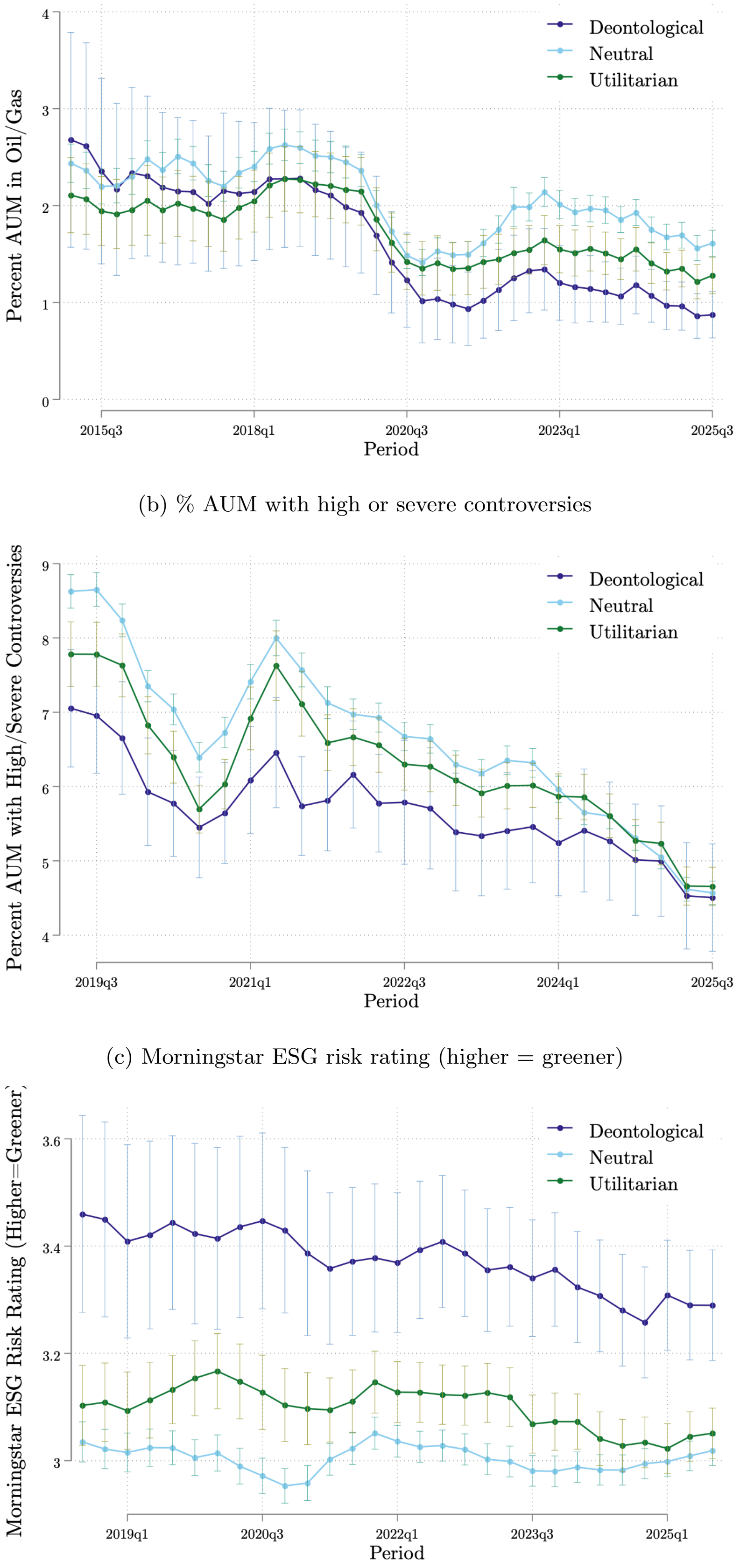
(b) Weighted-average carbon intensity per $M revenue

nificantly from neutral funds once fund-level characteristics, family–quarter shocks, and style–quarter shocks are accounted for.

4.3 Coordination Shocks and Type-Specific Reallocation

Hypothesis 2 concerns whether coordination shocks alter the behavior of different fund types. The CA100+ initiative provides a natural setting: in three successive waves, firms were publicly designated as targets for collective engagement. These announcements act as coordination shocks by reducing uncertainty about focal points for engagement and lowering the cost of joint action. The model predicts that utilitarian funds should increase their holdings of targeted firms relative to otherwise similar firms and relative to non-utilitarian funds, whereas deontological funds, who rely on exclusion

Figure 7: Portfolio-level differences in ESG-related metrics by fund type, 2015–2025. Bars indicate 95% confidence intervals.



(a) % AUM in oil and gas (Integrated)

(b) % AUM with high or severe controversies

(c) Morningstar ESG risk rating (higher = greener)

Table 6: Portfolio characteristics by fund moral type: Deontological vs. neutral

	Controversies	Carbon Intensity	AUM in Oil/Gas	Carbon Risk Score
Deon	-1.366***	-35.79***	-0.902***	-1.213***
	(-3.55)	(-3.83)	(-3.00)	(-7.15)
Log AUM	0.0661*	-5.015***	-0.0181	-0.0358
	(1.74)	(-2.96)	(-0.72)	(-1.34)
Log Avg Market Cap	2.704***	-52.85***	0.648***	-0.964***
	(17.74)	(-9.18)	(8.77)	(-10.47)
Quarterly Flow (Percent)	0.00188	-0.0775	-0.00143	0.000841
	(1.08)	(-1.16)	(-1.59)	(0.96)
Quarterly Return (Percent)	0.0141	0.850	0.00552	0.0151*
	(1.47)	(1.49)	(1.06)	(1.78)
Growth Tilt (Percent)	-0.0657***	-2.917***	-0.0186*	-0.0977***
	(-9.50)	(-9.55)	(-1.65)	(-17.75)
Small Cap Tilt (Percent)	0.0210***	-2.103***	0.00499	-0.00982
	(2.79)	(-4.14)	(1.17)	(-1.56)
Constant	-20.93***	948.3***	-3.714***	23.37***
	(-11.96)	(12.16)	(-4.06)	(19.91)
Observations	90,798	102,411	139,764	101,447
Funds	4776	4694	4959	4689
Adjusted R^2	0.551	0.362	0.549	0.692

Portfolio characteristics by fund type: pairwise comparison of Deontological (included) and Neutral (omitted category) funds. Dependent variables: (1) % AUM in high or severe controversies, (2) carbon intensity per $M revenue, (3) % AUM in oil and gas, and (4) Morningstar carbon risk score. Fixed effects: family×quarter and style×quarter. Standard errors clustered by fund. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7: Portfolio characteristics by fund moral type: Utilitarian vs. neutral

	Controversies	Carbon Intensity	AUM in Oil/Gas	Carbon Risk Score
Util	-0.162	6.493	-0.715	-0.227
	(-0.67)	(0.46)	(-1.39)	(-1.33)
Log AUM	0.0953***	-1.552	-0.00231	0.0118
	(2.89)	(-0.96)	(-0.07)	(0.51)
Log Avg Market Cap	2.438***	-56.92***	0.645***	-0.854***
	(19.41)	(-7.55)	(9.19)	(-10.87)
Quarterly Flow (Percent)	0.000246	-0.127**	-0.00223***	0.000770
	(0.19)	(-2.03)	(-2.70)	(1.04)
Quarterly Return (Percent)	0.0113	0.407	0.00997*	0.0189***
	(1.53)	(0.81)	(1.94)	(2.71)
Growth Tilt (Percent)	-0.0574***	-2.869***	-0.0208	-0.0955***
	(-9.38)	(-10.36)	(-1.62)	(-19.82)
Small Cap Tilt (Percent)	0.0109*	-2.231***	-0.00420	-0.00311
	(1.69)	(-4.61)	(-0.73)	(-0.55)
Constant	-19.06***	926.1***	-3.759***	21.18***
	(-13.09)	(11.06)	(-4.02)	(21.65)
Observations	110,037	124,756	168,048	123,612
Funds	5832	5727	6013	5721
Adjusted R^2	0.556	0.333	0.490	0.705

Portfolio characteristics by fund type: pairwise comparison of Utilitarian (included) and Neutral (omitted category) funds. Dependent variables: (1) % AUM in high or severe controversies, (2) carbon intensity per $M revenue, (3) % AUM in oil and gas, and (4) Morningstar carbon risk score. Fixed effects: family×quarter and style×quarter. Standard errors clustered by fund. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

rather than engagement, should not display the same reallocation.

4.3.1 Institutional background

CA100+ was launched in December 2017 at the One Planet Summit in Paris as a collaborative engagement platform for institutional investors. At inception, the coalition comprised roughly 225 signatories and targeted around 100 of the world's largest corporate greenhouse gas emitters. The list of target firms was subsequently expanded in additional waves in 2018 and 2023, in total covering over 160 firms. As of 2025, CA100+ counts over 600 investor signatories representing tens of trillions of dollars in AUM, coordinated across five investor networks (AIGCC, Ceres, IGCC, IIGCC, and PRI). The initiative does not mandate voting or portfolio reallocations, but rather provides focal points for engagement by identifying target companies, publishing benchmark metrics of climate transition performance, and providing a platform for sharing information and best practices. In doing so, CA100+ lowers strategic uncertainty about which firms investors should prioritize for climate engagement and provides reputational salience that amplifies the effect of investor pressure.

CA100+ offers a quasi-natural experiment that is particularly well suited to the present analysis. First, it provides sharp, dated shocks at the firm level: firms were designated as targets in discrete waves, creating staggered adoption events across the sample. This feature enables an event-study design that exploits variation in the timing of treatment. Further, because it targets firms rather than funds, the shock is exogenous to any single fund's revealed portfolio.

Second, the treatment is firm specific, in contrast to initiatives such as the Principles for Responsible Investment (PRI), which represent broad commitments at the investor level without discrete shocks at the firm level. PRI sign-ups are diffuse and self-selected and less plausibly exogenous, whereas CA100+ announcements are public, salient, and widely covered in financial media.

Third, CA100+ is explicitly about coordination: its purpose is to reduce strategic uncertainty and facilitate collective engagement. It facilitates information exchange and best practices among signatories but does not mandate coordination. This aligns directly with the theoretical notion of a coordination shock. Other initiatives, while important, lack the same event-like structure or explicit coordination mandate.

The empirical analysis does not condition on whether a fund or its parent firm is a signatory to CA100+. This choice is deliberate. First, the benefits of a coordination shock extend beyond formal signatories: by lowering uncertainty and creating focal points, CA100+ can influence the behavior of non-signatory investors as well.

Second, signatory status is often defined at the parent-firm level rather than the fund level, and the extent of active participation by any individual fund within a family is uncertain. Including a signatory dummy would therefore introduce measurement error and risk misclassification.

Finally, by showing that utilitarian funds respond to coordination shocks even without conditioning on membership, the analysis demonstrates that the coordination mechanism operates broadly rather than being confined to formal insiders. This is consistent with the model's prediction that coordination shocks lower strategic uncertainty for all agents who share the relevant preferences and reasoning modes.

4.3.2 Identification strategy

A natural concern is whether the designation of target firms is truly exogenous. It is plausible that investors, particularly large asset managers, had some influence in determining which firms were included on the CA100+ list, especially in the later waves. To the extent that inclusion reflects firms' prior salience or controversy, selection may correlate with unobserved determinants of fund reallocation. However, the shock I exploit is one of timing, not firm selection. The firms targeted by CA100+ were already well known to be large emitters prior to designation, and many had already been the subject of investor engagement efforts. The announcement did not introduce new information about these firms so much as it provided a focal point for collective action.

Nevertheless, designation may increase public salience or media coverage of targeted firms, which threatens identification. Two features of the empirical design mitigate these concerns. First, the empirical design incorporates firm-by-quarter fixed effects in the cross-type comparisons, ensuring that identification comes from differential reallocation across fund types holding the same firm in the same quarter. This ensures that any firm-level time-varying shocks, such as increased media coverage or regulatory scrutiny, are differenced out. Second, the within-type event study exploits staggered adoption across firms, absorbing fund-by-quarter and fund–firm fixed effects so that identification derives from changes within a fund's own portfolio around the event.

I first estimate the effect of coordination shocks separately within each fund type. For fund i of type τ, firm j, and quarter t, the specification is

$$w_{ijt} = \beta\left(\text{Target}_j \times \text{Post}_{jt}\right) + \phi_{it} + \mu_{ij} + \varepsilon_{ijt}. \tag{18}$$

Here, w_{ijt} is fund i's ownership of firm j at time t, measured as the fraction of the firm's total outstanding shares held by the fund; as both numerator and denominator are in shares, this measure is not affected by stock price movements. The indicator Target_j equals one if firm j is ever a CA100+ target. The variable Post_{jt} is one for all quarters at or after the announcement date for the targeted firm's wave, and for non-targets Post_{jt} assigns placebo announcement dates drawn from the empirical distribution of actual announcement dates. The coefficient β captures the change in fund i's ownership of firm j following the coordination shock, relative to other firms held by the same fund in the same quarter. All regressions absorb fund-by-quarter fixed effects ϕ_{it} to capture time-varying shocks to each fund's overall portfolio and fund–firm fixed effects μ_{ij} to absorb persistent match quality between a fund and a firm. Standard errors are two-way clustered by fund and firm.

Next, I estimate cross-type differences by pooling pairs of fund types. For a given pairwise comparison (for example, Deontological versus Neutral), the specification is

$$w_{ijt} = \beta\big(\text{Target}_j \times \text{Post}_{jt} \times \mathbb{1}\{i \in g\}\big) + \phi_{it} + \mu_{ij} + \kappa_{jt} + \varepsilon_{ijt}. \tag{19}$$

Here, w_{ijt}, Target_j, and Post_{jt} are defined as before. The indices g and h denote, respectively, the focal and baseline fund types in the given pairwise comparison (for example, g = Deontological and h = Neutral in the Deontological versus Neutral contrast). The indicator $\mathbb{1}\{i \in g\}$ equals one if fund i is of type g (the focal type) and zero if it is of type h (the baseline type). All regressions absorb fund-by-quarter fixed effects ϕ_{it}, fund–firm fixed effects μ_{ij}, and firm-by-quarter fixed effects κ_{jt}. Lower-order interaction terms are absorbed by the fixed effects; for example, κ_{jt} absorbs $\text{Target}_j \times \text{Post}_{jt}$. Note that the staggered timing of treatment breaks collinearity between $\text{Post}_{jt} \times \mathbb{1}\{i \in g\}$ and the fixed effects, so the triple interaction survives alongside the main regressor of interest. Standard errors are two-way clustered by fund and firm.

The coefficient of interest, β, captures a three-way difference: the change in ownership of the focal fund type in the CA100+ target firm post-announcement (i) relative to the same fund's pre-announcement ownership in the same firm, (ii) relative to non-target and not-yet-targeted firms held over the same period, and (iii) relative to the baseline fund type. This isolates whether the focal-type fund reallocates differently from the baseline-type fund in response to the coordination shock.

The two designs are complementary. The within-type specification isolates the overall response of each moral type to coordination shocks using staggered adoption for identification. The cross-type specification sharpens the contrast by directly estimating the incremental reallocation of the focal-type fund relative to the baseline-type fund for the same firm in the same quarter, net of fund-wide shocks and fund–firm match quality. Together, these tests evaluate whether utilitarian funds exhibit a response to coordination shocks that is distinct from other fund types and is thus attributable to their ethical orientation.

The within-type regressions in Table 8 indicate that utilitarian funds increase their holdings of CA100+ target firms following the announcement by approximately 0.016% of the firm's outstanding equity. Aggregated across the set of utilitarian investors, these basis-point reallocations cumulate into a substantial reshaping of the shareholder base of targeted firms, strengthening the voice of engagement-oriented investors in governance and voting outcomes. The cumulative effect is explored in the 2SLS analysis presented later. By contrast, deontological funds exhibit no systematic change in holdings. Neutral funds also tilt modestly into targets, suggesting that the coordination shock generates spillovers to the broader market.

The cross-type regressions in Table 9 sharpen these contrasts. The most robust finding is the utilitarian–deontological comparison: utilitarians increase their ownership of CA100+ targets by roughly 0.01% of outstanding equity relative to deontological funds, which corresponds to an incremental increase of about 20% in the number of shares held. Although other pairwise contrasts are less precisely estimated, the utilitarian–deontological gap provides direct evidence that the composition of the shareholder base shifts in favor of utilitarian funds when coordination opportunities arise, consistent with the model's predictions.

4.3.3 Robustness: stacked event-study design

Traditional two-way fixed-effect models in staggered-adoption settings pose well-documented challenges for causal inference, including the possibility of negative weighting and contamination from heterogeneous treatment effects (Goodman-Bacon, 2021). However, it is possible to recover valid average causal effects by replacing the standard two-way fixed-effect estimator with an average of cohort-specific difference-in-differences estimates (Borusyak et al., 2024; Sun & Shapiro, 2022).

Table 8: Fund response by type to positive coordination shocks

	Util	Deon	Neut
Post=1	-0.000804	0.0125	-0.000361
	(-0.82)	(1.03)	(-0.45)
Target=1 × Post=1	0.0119***	-0.00121	0.00827***
	(3.34)	(-0.31)	(3.24)
Constant	0.0401***	0.0317***	0.0475***
	(55.57)	(3.27)	(85.95)
Observations	6,383,268	1,151,144	19,226,524
Funds	897	252	2007
Firms	23956	11096	23703
Adjusted R^2	0.720	0.514	0.704

Effect of a positive coordination shock on holdings of each fund type. The coefficient on the interaction term measures ownership of the fund in CA100+ target firms post-announcement in excess of pre-announcement levels, relative to non-target and not-yet-targeted firms held by the same fund. Placebo treatment dates assigned to non-target firms are drawn from the empirical distribution of true treatment dates. Dependent variable: percent of total shares outstanding of firm j held by fund i in quarter t. Fixed effects: fund×quarter and fund×firm. Standard errors clustered by fund and firm. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

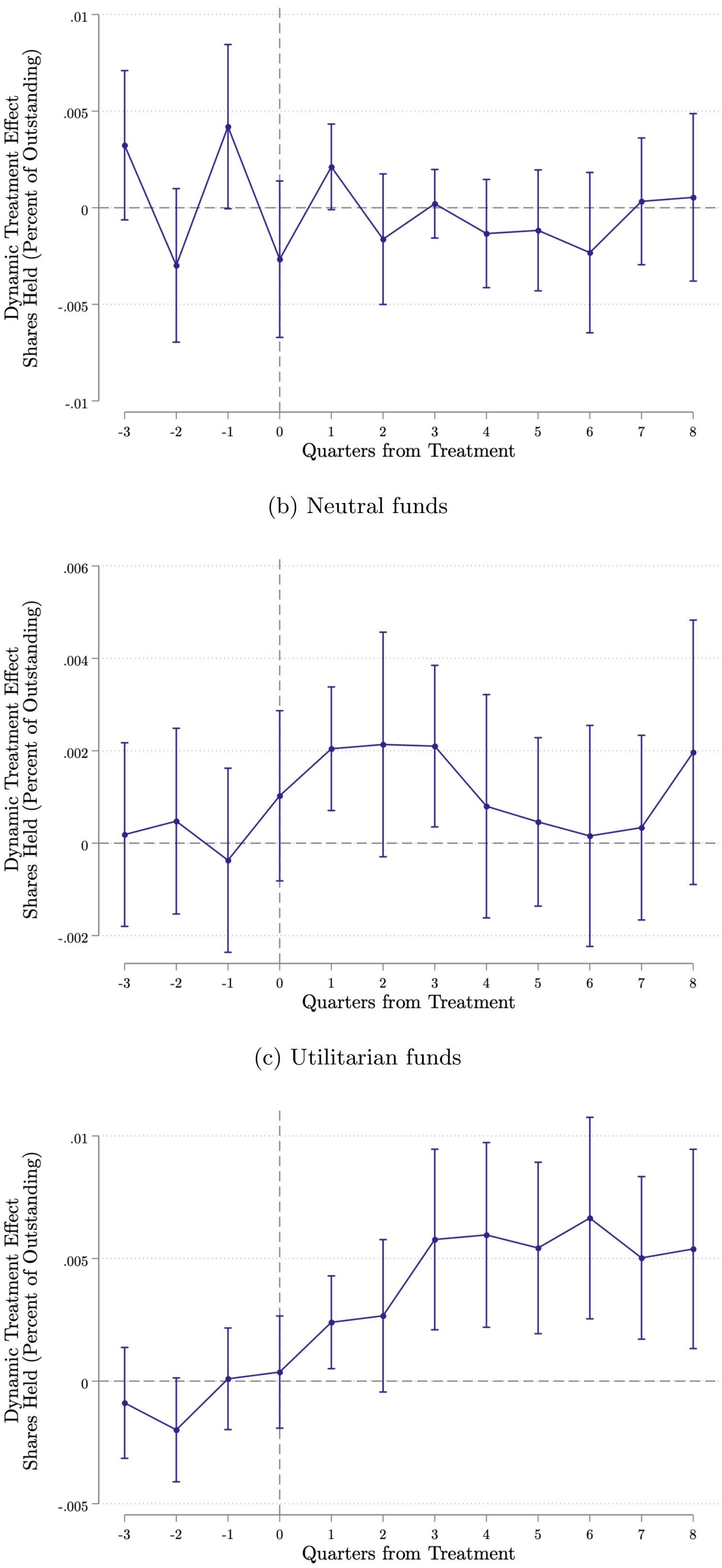
I implement a stacked event-study design using the cohort-specific difference-in-differences estimator of Callaway and Sant'Anna (2021). Appendix Section C.3 provides full details on the implementation. This approach constructs event-time cohorts based on the timing of treatment, that is, the CA100+ announcement date for each firm. The event study provides dynamic, cohort-clean estimates of how ownership by different moral types adjusts around coordination shocks, complementing the contemporaneous triple-interaction regressions above. In particular, it shows when utilitarian ownership begins to diverge from others, whether pre-trends are flat by cohort, and how effects evolve over several quarters.

Fig. 8 plots dynamic treatment effects for each fund type. The pattern is consistent with the theoretical predictions. For utilitarian funds, the coefficients display a clear upward trajectory in the quarters following the coordination shock, with post-event estimates significantly above zero and confidence intervals excluding the null across multiple horizons. This indicates that utilitarians systematically expand their ownership in targeted firms once coordination opportunities arise. By contrast, deontological funds show no clear trend; post-event coefficients are imprecise and confidence intervals uniformly cover zero, suggesting no systematic adjustment. Neutral funds exhibit a slight upward tendency, though weaker than for utilitarians. Overall, these results provide dynamic evidence of the predicted coordination mechanism.

4.4 From Ownership to Outcomes: 2SLS Evidence on Coordinated Abatement

The model implies that engagement-oriented (utilitarian) ownership influences firm outcomes through a coordination-based abatement channel. When collective engagement becomes more feasible, utilitarian investors are predicted to expand ownership in targeted firms, lowering coordination costs and ultimately increasing real abatement effort. Testing this mechanism empirically requires distin-

Figure 8: Event study of changes in holdings around positive shocks to coordinated engagement opportunities



(a) Deontological funds

(b) Neutral funds

(c) Utilitarian funds

Note: Confidence intervals at 95% level.

Table 9: Pairwise comparison of fund-type responses to a positive coordination shock

	Deon v Neut	Util v Neut	Util v Deon
Post=1 × Deon=1	0.00315^{**} (2.29)		
Target=1 × Post=1 × Deon=1	-0.00574 (-1.54)		
Post=1 × Util=1		-0.000119 (-0.11)	
Target=1 × Post=1 × Util=1		0.00452 (1.24)	
Post=1 × Util=1			-0.00161 (-1.02)
Target=1 × Post=1 × Util=1			0.00852^{**} (2.04)
Constant	0.0437^{***} (704.43)	0.0436^{***} (229.49)	0.0348^{***} (35.59)
Observations	20,327,035	25,549,869	7,440,449
Funds	2248	2886	1135
Firms	19395	21900	16706
Adjusted R^2	0.734	0.733	0.681

Pairwise comparison of fund-type responses to a positive coordination shock: Deontological versus Neutral, Utilitarian versus Neutral, and Utilitarian versus Deontological. The coefficient on the triple interaction term measures the three-way difference described above. Placebo treatment dates assigned to untreated firms are drawn from the empirical distribution of true treatment dates. Dependent variable: percent of total shares outstanding of firm j held by fund i in quarter t. Fixed effects: fund×quarter, firm×quarter, and fund×firm. Standard errors clustered by fund and firm. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

guishing changes in utilitarian ownership that are driven by coordination opportunities or fund-level inflows from those that merely reflect endogenous firm fundamentals. This section isolates plausibly exogenous variation in utilitarian ownership and estimates its causal effect on firms' emissions outcomes using a 2SLS design.

Two complementary sources of variation in a firm's shareholder composition provide the basis for identification.

1. *Coordination shocks:* The staggered rollout of CA100+ created sudden opportunities for collective engagement with a subset of large emitters. Treated firms (targets) experienced an exogenous reduction in coordination costs, leading engagement-oriented funds to increase ownership. This variation captures the narrow channel of coordinated abatement envisioned in the model.
2. *Fund flows:* Exogenous inflows into utilitarian funds mechanically raise the overall supply of utilitarian capital available for allocation across portfolio firms, shifting ownership composition independently of firm-level fundamentals. Flow shocks thus capture broader shifts in engagement-oriented ownership that are unrelated to any specific coordination event.

The two instruments therefore capture distinct but related mechanisms. The coordination-shock instrument isolates the causal impact of ownership changes specifically induced by coordination opportunities, while the Bartik-style flow instrument identifies the broader effect of engagement-oriented ownership, regardless of whether coordination is explicit or implicit. Together, they yield both a direct test of the model's coordination mechanism and a comprehensive assessment of utilitarian ownership's causal influence on firm outcomes.

The structural equation of interest is

$$Y_{jt} = \beta \, \mathrm{UtilOwn}_{jt} + \gamma \big(\mathrm{Target}_j \times \mathrm{Post}_{jt} \big) + \phi_j + \psi_{(c)t} + \zeta_{(s)t} + \varepsilon_{jt}. \tag{20}$$

Here, Y_{jt} is the outcome for firm j in year t, either (i) carbon emissions intensity (tons of greenhouse gases per million U.S. dollars of revenue) or (ii) the monetized cost of emissions per million U.S. dollars of revenue. The variable $\mathrm{UtilOwn}_{jt}$ denotes the percentage of shares held by all utilitarian funds in year t. The term $\mathrm{Target}_j \times \mathrm{Post}_{jt}$ captures the direct effect of CA100+ treatment. The vectors ϕ_j are firm fixed effects, $\psi_{(c)t}$ are cohort-by-year fixed effects controlling for differential time trends across target cohorts, and $\zeta_{(s)t}$ are industry-by-year fixed effects capturing global shocks to specific sectors. Standard errors are clustered by firm.

Because ownership may be endogenous to firm performance or anticipated policy changes, I estimate the causal effect of utilitarian ownership using the 2SLS estimator. The first stage is

$$\mathrm{UtilOwn}_{jt} = \pi_1 \big(\mathrm{Target}_j \times \mathrm{Post}_{jt} \big) + \pi_2' Z_{jt} + \phi_j + \psi_{(c)t} + \zeta_{(s)t} + \nu_{jt}, \tag{21}$$

where Z_{jt} denotes the vector of excluded instruments:

$$Z_{jt} = \begin{bmatrix} \mathrm{BartikFlows}_{jt} \\ \mathrm{Target}_j \times \mathrm{Post}_{jt} \times \mathrm{PreExp}_j \end{bmatrix}. \tag{22}$$

Here, $\mathrm{BartikFlows}_{jt}$ is the Bartik-style flow instrument constructed by interacting aggregate annual net flows into utilitarian funds with each firm's pre-shock exposure to those funds. The term $\mathrm{Target}_j \times \mathrm{Post}_{jt} \times \mathrm{PreExp}_j$ is the triple interaction between CA100+ target status, post-announcement indicator, and pre-shock exposure of utilitarian funds to the firm's industry (leave-one-out average ownership by utilitarian funds). This instrument captures the intensity of the coordination effect: firms in industries already held by utilitarian investors experience stronger predicted increases in utilitarian ownership after the shock.

The Bartik-style flow instrument is constructed by interacting aggregate quarterly net flows into utilitarian funds with each firm's pre-shock (leave-one-out) exposure to those funds. Because fund flows largely reflect investor demand shifts unrelated to individual firm fundamentals, this instrument satisfies the exclusion restriction so long as flows do not directly alter firm behavior other than through ownership.

The coordination-shock instrument exploits exogenous variation from the CA100+ initiative. Target assignment was based on pre-determined size and emissions criteria, not contemporaneous performance, ensuring that treatment is orthogonal to firm-specific shocks. The interaction with pre-shock utilitarian exposure captures the intensity of the potential coordination effect: firms in industries already held by utilitarian investors experience stronger predicted increases in utilitarian ownership after the shock.

The instruments are highly relevant: the first-stage regressions (see Table 10) yield F-statistics comfortably above conventional thresholds for weak-instrument concerns. Sign directions match theoretical expectations; both coordination shocks and fund inflows significantly increase utilitarian ownership in targeted firms.

The second-stage results, reported in Tables 11 and 12, show that increases in utilitarian ownership lead to economically and statistically significant reductions in firms' emissions outcomes. The estimated coefficient indicates that a one–percentage-point increase in utilitarian ownership reduces emissions intensity by at least 70 tons of greenhouse gas emissions per million U.S. dollars of revenue. For comparison, the mean firm in the sample emits around 270 tons, while the mean firm targeted by CA100+ emits approximately 1600 tons per million U.S. dollars of revenue per year.

The results substantiate the model's key prediction that the effectiveness of ethical capital depends on coordination. Ownership shifts by utilitarian funds translate into measurable real outcomes only when coordination frictions are reduced, confirming that moral intent alone is insufficient: organization and timing matter. The estimated magnitudes suggest that modest reallocations of capital by engagement-oriented investors can generate abatement effects comparable to those achieved through major regulatory interventions or carbon-pricing adjustments of similar scale.

Taken together, the empirical evidence traces the full causal chain articulated by the model. The results establish that the capacity of ethical capital to correct externalities hinges on the collective dimension of ownership, validating the theoretical mechanism of coordination-based abatement.

Table 10: First stage of 2SLS: instrumenting utilitarian ownership with coordination shock and fund flows (shift–share Bartik)

Panel A: Includes all firm–years with non-missing ownership data.

	Util	Deon	Neut
Bartik (Percent of Firm)	1.560***	0.638	3.156***
	(4.97)	(1.15)	(8.92)
Target × Post × Pre-Shock Ind. Exp.	0.652***	1.043	0.0762
	(3.94)	(1.54)	(0.52)
Target × Post	-0.283***	-0.0204	-0.0222
	(-3.55)	(-0.64)	(-0.16)
Constant	0.413***	0.0692***	1.416***
	(950.65)	(141.69)	(1722.34)
Observations	89555	49199	80482
Firm Clusters	8883	4470	8006
F-statistic	14.849	1.430	26.599
Adjusted R^2	0.666	0.486	0.863

Panel B: Sample restricted to firm–years with matched emissions data for outcome analysis.

	Firm ownership (%) by fund type		
	Engage	Exclude	Neutral
Bartik Flows (% of Firm)	1.725***	0.871	1.255***
	(3.76)	(1.07)	(2.73)
Target × Post × Pre-Shock Ind. Exp.	0.518**	1.022	-0.0148
	(2.29)	(1.47)	(-0.13)
Target × Post	-0.206**	-0.0151	0.0391
	(-2.13)	(-0.42)	(0.37)
Constant	0.461***	0.0722***	1.464***
	(483.74)	(78.92)	(372.07)
Observations	60525	34797	54681
Firm Clusters	8408	4397	7539
F-statistic	6.815	1.648	2.490
Adjusted R^2	0.748	0.531	0.912

Dependent variable: percent of total shares outstanding of firm j held by all funds of a given type in year t. Fixed effects: firm, cohort×year, and industry×year. Standard errors clustered by firm. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 11: 2SLS estimates: effect of utilitarian ownership on emissions intensity

	GHG emissions		
	+1Y	+2Y	+3Y
Engager Ownership (% of Firm)	-70.19	-82.64*	-72.40
	(-1.34)	(-1.91)	(-1.08)
Target × Post	-234.9**	-274.6**	-217.2*
	(-2.15)	(-2.24)	(-1.75)
Observations	60525	54775	47055
Firm Clusters	8408	8173	7859
K-P rk Wald F statistic	10.196	15.716	13.474
Hansen J statistic	1.180	1.259	1.530
p-value	0.277	0.262	0.216

Dependent variable: emissions intensity in tons of GHG per $M in revenue, led by 1 year, 2 years, and 3 years. Excluded instruments: (i) shift–share Bartik instrument based on fund flows into utilitarian funds, expressed in percentage of firm-level ownership and (ii) triple interaction between CA100+ target status, post-announcement indicator, and pre-shock exposure to the firm's industry, computed as a leave-one-firm-out average ownership across all utilitarian funds. Included instrument: $\text{Target}_j \times \text{Post}_{jt}$. Fixed effects: firm, cohort×year, and industry×year. Standard errors clustered by firm. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 12: Effect of utilitarian ownership on emissions costs

	+1Y	+2Y	+3Y
Util. Ownership (Percent of Firm)	-0.316	-0.369**	-0.327
	(-1.53)	(-2.03)	(-1.13)
Target × Post	-0.763*	-0.962**	-0.754
	(-1.88)	(-2.00)	(-1.48)
Observations	60525	54775	47055
Firm Clusters	8408	8173	7859
K-P rk Wald F statistic	10.196	15.716	13.474
Hansen J statistic	1.125	1.226	1.532
p-value	0.289	0.268	0.216

Dependent variable: emissions cost per $M in revenue, led by 1 year, 2 years, and 3 years. Excluded instruments: (i) shift–share Bartik instrument based on fund flows into utilitarian funds, expressed in percentage of firm-level ownership and (ii) triple interaction between CA100+ target status, post-announcement indicator, and pre-shock exposure to the firm's industry, computed as a leave-one-firm-out average ownership across all utilitarian funds. Included instrument: $\text{Target}_j \times \text{Post}_{jt}$. Fixed effects: firm, cohort×year, and industry×year. Standard errors clustered by firm. t-statistics in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5 Theoretical Extensions

The theoretical analysis thus far reveals a core lesson: when moral reasoning is heterogeneous and strategic interdependence is endogenous, both the level and pattern of activism—and the resulting market and social outcomes—are dictated not by averages, but by the participation thresholds and interactions among types. Abatement, price, and risky asset allocations are not mere functions of preferences, but of the marginal agent's reasoning logic and the structure of coordination.

In what follows, I discuss the extent to which Kant equilibrium can be recovered from modified Nash games, examine model robustness to behavioral distortions such as imperfect applications of the universalizability principle, and finally examine conditions under which Kant reasoning can emerge endogeneously and persist via evolutionary replicator dynamics.

Additionally, Appendix Section B.1 compares Kant equilibrium to models of fairness (Rabin, 1993), team reasoning (Sugden, 2003) and psychological Nash (Geanakoplos et al., 1989). Appendix Section B.6 addresses the unsuitability of self-confirmation arguments for Kant equilibria and proposes 'self-enforcement' as an alternative property to characterize stability. Appendix Section Section B.7 discusses off-equilibrium plausibility and refinements.

Together with the aforesaid appendices, this section formalizes the Nash-Kant hybrid equilibrium as a necessary extension to the existing toolkit of economics and finance, arguing that without it, norm-based behaviors observed in ethical investing and real-world coordination games remain irreducible anomalies.

5.1 Recovering Kant Equilibrium from a Modified Nash Game

One might ask whether Kant equilibrium could be recovered from a transformed Nash game. For instance, consider a two-player game with players i and j, and utility function $U(\phi_i, \phi_j)$. Define a transformed game by modifying the utility function as follows:

$$\tilde{U}(\phi_i, \phi_j) \;=\; U(\phi_i, \phi_j) \;-\; \omega \mathcal{I}(\phi_i \neq \phi_j) \tag{23}$$

where $\mathcal{I}(\cdot)$ is an indicator function that equals 1 if the argument is true and 0 otherwise, and $\omega > 0$ imposes a penalty for non-conformity. As $\omega \to \infty$, symmetric strategies are enforced by fiat, and the Nash equilibrium of the transformed game coincides with the Kant fixed point.

That said, the Kant solution concept, while recoverable as a limit case of this transformed game, is not equivalent to any finite-ω Nash equilibrium. Several important distinctions arise.

First, empirical observation finds that ethically-driven activism is persistent despite widespread coordination failures. The modified game Nash equilibrium may correctly predict activism in levels, but fails to capture its stability, as $\frac{\partial \tilde{U}_i}{\partial \phi_j} \neq 0$. The Nash agent's sensitivity to peer actions rises with ω, making equilibrium increasingly fragile to belief perturbations. In contrast, Kant equilibrium maintains stability by modeling principled commitment that does not waver with beliefs about others' actions. The Kant agent's action is stable as $\frac{\partial U_i}{\partial \phi_j} = 0$ by construction, thus capturing the observed persistence and resilience of norm-driven activism.

Second, in the modified Nash game, the choice of ω is ad hoc and its interpretation ambiguous absent a clear psychological or philosophical microfoundation. But this construction is behaviorally unmotivated—it does not explain *why* agents would reason universally, only that such reasoning could be forced through extreme penalties. In contrast, the Kant model embeds universalizability as a principle of ethical justification, not as a strategic constraint. Kant reasoning is grounded in a well-understood normative principle from a rich philosophical tradition, and further, it provides a precise agent-level rule for selecting actions.

Third, the Kant framework accommodates extensions to capture imperfect application of the normative principle and the dynamic evolution of norms based on perceived uptake. The modified Nash game does not lend itself easily to such extensions without losing interpretability.

Fourth, as ω varies, the equilibrium transitions discontinuously between Kant and Nash logic, and intermediate equilibria do not correspond to any meaningful hybrid of the two. The distinction between Kant and Nash reasoning is qualitative, not merely quantitative. The Kant agent's logic is structurally different from a Nash agent with a conformity preference.

In sum, while Kant reasoning can be seen as a limit case of team reasoning or psychological Nash, it cannot be recovered by any pure payoff modification in a standard Nash framework without also altering the belief formation. As such, Kant equilibrium is not merely a special case of Nash equilibrium but a distinct solution concept with its own behavioral and ethical implications.

5.2 Imperfect Universalizability and Behavioral Distortions

Canonical Kant strategies assume that each agent evaluates their action under the counterfactual that all other agents take the identical action. This "full universalizability" assumption is strong and potentially unrealistic. To relax it, we introduce a parameter $\epsilon \in [0, 1]$ that encodes the agent's subjective belief in the extent to which their action will be reciprocated by others: $\epsilon = 1$ recovers the canonical Kant logic; $\epsilon = 0$ recovers Nash behavior.

Define agent i's portfolio weight under partial application of the universalizability test as

$$\phi_i(\epsilon) \;=\; \epsilon\, \phi_i^K \;+\; (1-\epsilon)\, \phi_i^N \tag{24}$$

where ϕ_i^K is the optimal portfolio weight under full universalization, and ϕ_i^N is the Nash equilibrium allocation given prevailing market conditions. The agent's optimal action, therefore, is a solution to a best-response problem under this weighted belief. Here, ϵ is not merely a fuzzy universalizability parameter, but a moral epistemic threshold that reflects the agent's conviction in the generalizability of her action. Higher values reflect stronger normative commitments or institutional mandates (e.g., faith-based funds, ESG principles signatories); lower values reflect pragmatism, skepticism, or moral minimalism. This framework accommodates a wide range of real-world investor types: $\epsilon = 1$ represents unilateral moral leaders, $\epsilon \in (0.5, 1)$ rule-driven institutional investors with partial expectations of uptake, and $\epsilon \approx 0$ opportunistic virtue signalers.

We may allow ϵ to depend on environmental variables or institutional cues, such as the coordination technology available, salience of reputational incentives, signal strength, perceived norm uptake, or the agent's own past behavior. Such formulations enable the modeling of Kant agents who are strategic about their own normative thresholds, even if they are not belief-consistent in the Nash sense.

As ϵ decreases from 1 to 0, the allocation $\phi_i(\epsilon)$ interpolates smoothly from the Kant to the Nash equilibrium, and the aggregate abatement, market price, and the wedge between the Kant and Nash outcomes all decline monotonically with ϵ. Thus, imperfect universalizability systematically attenuates the intensity of Kant activism. As agents' beliefs about others' willingness to reciprocate recede toward selfishness, individual and aggregate investment in abatement diminishes, and the collective action mechanism weakens. This result is robust: any real-world deviation from perfect moral optimism—whether due to social context, institutional uncertainty, or psychological pessimism—erodes activism unless offset by external coordination mechanisms or policy interventions.

This result formally recovers Nash play as a limiting case and ensures that the model nests traditional equilibrium concepts. Next, to go beyond the simple reduction in universalizability, it is instructive to consider additional behavioral distortions, wherein a "Kant" agent misapplies or inconsistently implements the universalizability test. These distortions can be categorized as follows:

Local or Myopic Test of Universalizability. The agent universalizes only over a proper subset $A \subset N$ of agents (e.g., all funds within a particular market segment, coalition, or geographic region), Nash-best-responding to the remainder $A' = N \setminus A$. The perceived aggregate is thus $\phi_i(\epsilon) = \epsilon\phi_i^K(A) + (1-\epsilon)\phi_i^N(A')$, where the Kant component is restricted to the reference group. This represents bounded rationality, cognitive myopia, or a limited moral imagination. As the size of the reference group shrinks, the universalization effect attenuates proportionally—so long as the group is exogenously fixed and not behaviorally correlated with own action.

Reference-Dependent Test of Universalizability. The agent's universalization is anchored to a reference point, such as their own previous action or the status quo. That is, the counterfactual becomes "what if everyone acted as I did yesterday," or "what if all those already engaged intensified their action as I do?" This introduces temporal or cohort-based biases into moral reasoning and introduces inertia or path dependence in activism as a result of anchoring.

Biased Counterfactuals. The agent systematically overestimates or underestimates the probability that others will reciprocate their action. Such bias may arise from over-optimism (inflated perceived impact, resulting in over-investment in abatement) or pessimism (deflated impact, leading to under-investment). Here, the convex combination is itself miscalibrated, with ϵ not reflecting objective probabilities.

Most of these normative distortions can be recast as variants of imperfect universalizability, mathematically captured via the modified ϵ parameter. If an agent universalizes over a subgroup A and Nash-best-responds to the complement A', her best response is a convex combination of Kant logic (over A) and Nash logic (over A'). This framework also allows for agent-specific, probabilistic, or time-varying weights reflecting context-dependent beliefs or information.

Similarly, random or miscalibrated reference groups, status-quo anchoring, or dynamic learning all reduce to convex combinations of Nash and Kant logics, with contextually determined weights. When changes in social structure, external shocks, or learning dynamics alter the agent's perception of universalizability, ϵ can be treated as an endogenous variable.

Structurally, these forms of imperfect universalizability—local reference groups, optimism or pessimism, or subgroup moralization—all reduce to parameterizations of a unified "mixed motives" model. The result is a quantitative (rather than qualitative) attenuation of the Kant effect; that is, imperfect universalizability weakens, but does not fundamentally alter, the nature of moral activism in equilibrium.

Imperfect application of the universalizability test may also be interpreted as a source of intra-Kant heterogeneity, where agents differ in their beliefs about the scope and impact of their actions. This heterogeneity can lead to a distribution of ϵ values across the population, producing a mixed equilibrium where some agents act Kant, others Nash, and still others in between. This reintroduces strategic variation within the norm-driven class. The imperfect universalizability framework thus transforms the Kant from a fixed-type agent into a flexible, epistemically rich actor with bounded strategic interdependence. This allows for tractable comparative statics, endogenous norm dynamics, and realistic modeling of ESG-motivated investor heterogeneity.

Qualitative departures from this structure arise only if beliefs about others' behavior are nonlinearly correlated with own actions or with market features (e.g., via network externalities, dynamic feedback, or endogenous formation of reference groups), or if universalizability test is activated only above a critical threshold ("tipping point" universalizability). In such cases, the equilibrium may exhibit discontinuities, multiple steady states, or path dependence—avenues left for future research.

In summary, most empirically plausible distortions of Kant reasoning—whether through bounded rationality, reference dependence, or miscalibration—can be represented as a convex combination of Nash and Kant motives, parameterized by the agent's beliefs about the scope and impact of their actions. Thus, the principal effect of imperfect universalizability is to quantitatively diminish Kant activism, rather than to qualitatively reshape the collective action problem. True qualitative departures would require fundamentally nonlinear or dynamically evolving belief structures.

5.3 Evolutionary Dynamics

To understand the conditions under which Kantian universalizability can emerge and persist as a reasoning rule, I endogenize the distribution of strategic logics within the population by embedding them in an evolutionary framework. Rather than taking solution concepts (Nash vs. Kant) as fixed, I model them as heritable behavioral traits subject to evolutionary pressure. This allows for analysis of whether norm-based reasoning can persist, invade, or vanish depending on its relative payoff performance.

Let the population consist of a continuum of agents indexed by reasoning type $\tau \in \{\text{Nash}, \text{Kant}(\epsilon)\}$, where $\epsilon \in [0, 1]$ denotes the degree of universalizability in the Kantian heuristic. Each type τ is characterized by its distinct decision rule as defined previously. Let $x_\tau(t)$ denote the population share of reasoning type τ at time t, with $\sum_\tau x_\tau(t) = 1$.

Each type earns expected utility

$$U(\tau) = \mathbb{E}[u(w'_\tau)] - \text{MoralDisutility}_\tau(\phi_\tau), \tag{25}$$

where w'_τ denotes final wealth and ϕ_τ is the strategy adopted by type τ under its reasoning rule.

The evolution of types follows the standard replicator dynamic:

$$\frac{dx_\tau}{dt} = x_\tau(t)\left[U(\tau) - \bar{U}(t)\right], \quad \text{where } \bar{U}(t) = \sum_\tau x_\tau(t)U(\tau). \tag{26}$$

A population vector $\vec{x}^*$ is *evolutionarily stable* if no rare mutant type τ' can achieve higher expected utility when rare:

$$U(\tau') < \bar{U} \qquad \text{for all } \tau' \notin \text{supp}(\vec{x}^*). \tag{27}$$

In the portfolio-choice context, this implies that a minority of high-ϵ Kant types can persist if their norm-based strategies generate sufficient collective abatement benefits to raise overall welfare. As environmental externalities intensify or coordination technologies improve, Kant reasoning can *invade* a Nash-dominant population by producing larger marginal welfare gains. As $\epsilon \to 0$, Kant reasoning converges to Nash behavior, and the replicator dynamic collapses to a standard evolutionary process over preferences alone. These dynamics highlight a novel form of equilibrium selection: the relative fitness of reasoning rules depends jointly on their private payoffs and the externalities they jointly produce.

To avoid lock-in at boundary equilibria, introduce a mutation term $\nu > 0$ capturing experimentation or cultural diffusion. The augmented dynamics are:

$$\frac{dx_\tau}{dt} = x_\tau(t)\big(U(\tau) - \bar{U}(t)\big) + \nu\left[m(\tau) - x_\tau(t)\right], \tag{28}$$

where $m(\tau)$ is a baseline mutation distribution across reasoning types. This extension ensures persistent diversity and captures empirical observations that new institutional logics (e.g., ESG stewardship mandates) arise and diffuse even when initially suboptimal.

This evolutionary formulation transforms Kant reasoning from a static heuristic into a dynamic, selectable rule. It explains both the persistence and fragility of norm-driven investment behavior as emergent features of market evolution. Crucially, it endogenizes the moral–epistemic composition of the market: reasoning styles are not assumed—they evolve.

Alternative approach. Following Nowak (2006), cooperative (or conditionally cooperative) strategies become evolutionarily stable when the benefit-to-cost ratio of cooperation exceeds a threshold under five canonical mechanisms:

1. Kin selection: $r > c/b$ (genetic relatedness)
2. Direct reciprocity: $w > c/b$ (repeated interaction)
3. Indirect reciprocity: $q > c/b$ (reputation effects)

4. Network reciprocity: $b/c > k$ (local clustering)
5. Group selection: $b/c > 1 + n/m$ (group-level selection)

In this setting, indirect reciprocity—driven by transparency (q) and reputational rewards (b)—is the most relevant. When ESG disclosure is highly transparent (q large) and reputational benefits b exceed activism costs c, Kant reasoning becomes evolutionarily stable.

Formally, the payoff to a Kant agent is given by:

$$U_K = \pi + \psi\, M_K, \qquad M_K = q\, b\, x - c, \tag{29}$$

where π denotes the pecuniary return, $\psi > 0$ measures the weight placed on moral utility in portfolio choice, and x is the population share of Kant agents. The corresponding payoff to Nash agents is U_N.

Under a two-type reduction, the replicator dynamic simplifies to:

$$\dot{x} = x\left[U_K(x) - \bar{U}(x)\right], \tag{30}$$

$$\bar{U}(x) = x\, U_K(x) + (1 - x)\, U_N(x), \tag{31}$$

where x now denotes the share of Kant agents in the population.

A necessary and sufficient condition for Kant reasoning to invade when rare is:

$$U_K(x) > U_N(x)\big|_{x\to 0} \quad\Longrightarrow\quad q\, b > c. \tag{32}$$

If this inequality holds, Kant reasoning spreads from rarity, converging to a unique interior rest point:

$$x^* = \frac{c}{q\, b}. \tag{33}$$

Thus, the evolutionary stability of Kant reasoning depends on the balance between reputational rewards and activism costs, scaled by market transparency. In environments with strong disclosure (q large) and high reputational payoffs (b large), norm-based reasoning persists at equilibrium.

Solving the steady state of the replicator dynamic yields three stationary points: $x = 0$ (no Kant agents), $x = 1$ (all Kant agents), and the interior equilibrium x^*. For $q\, b > c$, the interior rest point is locally stable; for $q\, b < c$, Kant reasoning cannot invade and the Nash equilibrium is evolutionarily absorbing.

Finally, note that Kant equilibria do not, in general, Pareto-dominate Nash equilibria. While the presence of Kant types may increase aggregate welfare through abatement, Nash types can be worse off when norm-driven activism alters payoffs or prices against their interests. Pareto superiority is parameter-dependent and not guaranteed in the general case.

Table 13: Properties of Kant Equilibria

Property	Description	Remarks
Existence	Does equilibrium exist for all admissible environments?	Yes, by Brouwer Fixed Point Theorem. See Proposition 3
Uniqueness	Is the equilibrium unique?	Achievable under broad conditions. See Proposition 3
Self-Confirmation	Are beliefs correct on-path?	Inapplicable. Self-enforcement at the rule level is proposed as an alternative benchmark. See Section B.6
Off-Equilibrium Plausibility	Is equilibrium robust to off-path beliefs?	Standard definitions fail. Universalizability-consistent OEP is proposed; refinement remains open. See Section B.7
Structural Robustness	Are outcomes robust to parameter shifts or shocks?	Yes, provided smooth primitives. See Section 3.1
Robustness to Distortions	What if agents misapply norms or are boundedly rational?	Model accommodates imperfect universalizability, local moral reference groups, and biased reasoning—results in attenuation, not collapse. See Section 5.2
Evolutionary Stability	Can Kant reasoning persist under selection and mutation?	Yes, under high transparency and reputational incentives. See Section 5.3
Pareto Optimality	Is Kant allocation Pareto superior to Nash?	Not necessarily. Aggregate welfare rises do not guarantee individual improvements for all types.
Empirical Content	Does the model yield testable predictions?	Yes. Theoretical claims generate observable implications for activism patterns, coordination responses and impact. See Section 3.3

6 Conclusion

This paper proposes Kant equilibrium as a necessary extension to the economic and financial modeling toolkit. It captures a distinct class of ethical behavior that escapes standard Nash logic, offering a formal framework for how principled moral commitment can reshape market equilibria under strategic interdependence. Table 13 summarizes the main properties and open questions associated with Kant equilibria, distilling key insights and flagging directions for future research.

The model developed in this paper delivers a rich set of predictions about activism thresholds, belief-independent participation, and the role of coordination frictions in determining outcomes. It provides a richer conceptual toolkit for evaluating institutional design, investor typology, and the scope of collective action under ethical pluralism. The core insight is simple: strategic behavior does not collapse to belief-consistent optimization when agents act on principle. The result is a reconceptualization of equilibrium itself—its meaning, its structure, and its fragility.

References

Adomaitis, N. (2023). Norway's oil fund to vote against climate resolution at BP [newspaper]. *Reuters: Sustainable Business*.

Albuquerque, R., Fos, V., & Schroth, E. (2022). Value creation in shareholder activism. *Journal of Financial Economics*, *145*, 153–178.

AllianzGI. (2025a, May 12). *AllianzGI announces its intention to vote against the reelection of Adidas' Chair at its upcoming AGM — Allianz Global Investors*. Allianz Global Investors. Retrieved June 12, 2025, from https://www.allianzgi.com/en/press-centre/media/press-releases/20250512-allianzgi-announces-its-intention-to-vote-against-the-reelection-of-adidas-chair

AllianzGI. (2025b, May 19). *AllianzGI announces its support for two shareholder proposals at Meta Platforms' 2025 AGM — Allianz Global Investors*. Allianz Global Investors. Retrieved June 12, 2025, from https://www.allianzgi.com/en/press-centre/media/press-releases/20250519-allianzgi-announces-its-support-for-two-shareholder-proposals-at-meta-platforms-2025-agm

AllianzGI. (2025c, June 2). *AllianzGI announces its intention to vote for three resolutions at the Alphabet upcoming AGM — Allianz Global Investors*. Allianz Global Investors. Retrieved June 12, 2025, from https://www.allianzgi.com/en/press-centre/media/press-releases/20250602-allianzgi-announces-its-intention-to-vote-for-three-resolutions-at-the-alphabet-upcoming-agm

Amundi. (2024, April 11). *Amundi Voting Strategy 2024: Key figures from the 2023 voting season*. Amundi Investment Solutions. Retrieved June 12, 2025, from https://int.media.amundi.com/news/amundi-voting-strategy-2024-key-figures-from-the-2023-voting-season-f59a-b6afb.html

Angeletos, G.-M., Hellwig, C., & Pavan, A. (2007). Dynamic Global Games of Regime Change: Learning, Multiplicity, and the Timing of Attacks. *Econometrica*, *75*(3), 711–756.

Appel, I. R., Gormley, T. A., & Keim, D. B. (2016). Passive investors, not passive owners. *Journal of Financial Economics*, *121*(1), 111–141.

Bacharach, M., Gold, N., & Sugden, R. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.

Bainbridge, S. M. (2005, August 1). *The Case for Limited Shareholder Voting Rights*. 781429. https://doi.org/10.2139/ssrn.781429

Barko, T., Cremers, M., & Renneboog, L. (2017, May 31). *Shareholder Engagement on Environmental, Social, and Governance Performance* (SSRN Scholarly Paper No. ID 2977219). Social Science Research Network. Rochester, NY.

Baron, J., & Spranca, M. (1997). Protected Values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1–16.

Barrage, L., & Nordhaus, W. D. (2023, April). *Policies, Projections, and the Social Cost of Carbon: Results from the DICE-2023 Model*. 31112. https://doi.org/10.3386/w31112

Becht, M., Franks, J., Grant, J., & Wagner, H. F. (2017). Returns to Hedge Fund Activism: An International Study. *The Review of Financial Studies*, *30*(9), 2933–2971.

Becht, M., Franks, J., Mayer, C., & Rossi, S. (2009). Returns to Shareholder Activism: Evidence from a Clinical Study of the Hermes UK Focus Fund. *The Review of Financial Studies*, *22*(8), 3093–3129.

Becht, M., Franks, J. R., Miyajima, H., & Suzuki, K. (2023, July 21). *Does Paying Passive Managers to Engage Improve ESG Performance?* 4506415. https://doi.org/10.2139/ssrn.4506415

Becht, M., Franks, J. R., & Wagner, H. F. (2019, October 1). *Corporate Governance Through Voice and Exit*. 3456626. https://doi.org/10.2139/ssrn.3456626

Bénabou, R., Falk, A., Henkel, L., & Tirole, J. (2020). Eliciting Moral Preferences: Theory and Experiment. *Econometrica*.

Benny, J. (2017). BlackRock urges Exxon to disclose more about climate change-related risks [newspaper]. *Reuters: Business*.

Bentham, J. (2017, October 8). *An Introduction to the Principles of Morals and Legislation*. CreateSpace Independent Publishing Platform.

Berk, J. B., & van Binsbergen, J. H. (2025). The impact of impact investing. *Journal of Financial Economics*, *164*, 103972.

Bieber, E., & Klingsberg, E. (2016, February 24). *What the 2016 BlackRock Letter Means for Shareholder Engagement and Disclosure Practices*. The Harvard Law School Forum on Corporate Governance. Retrieved June 12, 2025, from https://corpgov.law.harvard.edu/2016/02/24/what-the-2016-blackrock-letter-means-for-shareholder-engagement-and-disclosure-practices/

Blanding, M. (2017, January 30). *Vanguard, Trian And The Problem With 'Passive' Index Funds — Working Knowledge*. Harvard Business School. Retrieved June 12, 2025, from https://www.library.hbs.edu/working-knowledge/passive-index-fund-leaders-push-for-shareholder-reforms

Bloomberg. (2024, September 17). *Vanguard investor program fails to rock proxy votes*. InvestmentNews. Retrieved June 12, 2025, from https://www.investmentnews.com/ria-news/vanguard-investor-program-fails-to-rock-proxy-votes/257209

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*.

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*, *91*(6), 3253–3285.

Braham, M., & Hees, M. van. (2020). Kantian Kantian Optimization. *Erasmus Journal for Philosophy and Economics*, *13*(2), 30–42.

Brav, A., Jiang, W., & Kim, H. (2015). The Real Effects of Hedge Fund Activism: Productivity, Asset Allocation, and Labor Outcomes. *The Review of Financial Studies*, *28*(10), 2723–2769.

Broccardo, E., Hart, O., & Zingales, L. (2022). Exit versus Voice. *Journal of Political Economy*, *130*(12), 3101–3145.

Bybel, O. (2025, June 2). *AllianzGI push for AI and child safety transparency at Alphabet AGM*. Selector. Retrieved June 12, 2025, from https://citywire.com/selector/news/allianzgi-push-for-ai-and-child-safety-transparency-at-alphabet-agm/a2467016

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Candriam. (2025). *Predeclaration of Voting Intentions*. Candriam. Retrieved June 12, 2025, from https://www.candriam.com/en/professional/insight-overview/publications/predeclaration-of-voting-intentions/

Carlsson, H., & van Damme, E. (1993). Global Games and Equilibrium Selection. *Econometrica*, *61*(5), 989–1018.

Carpenter, J. (2021). The shape of warm glow: Field experimental evidence from a fundraiser. *Journal of Economic Behavior & Organization*, *191*, 555–574.

Curry, P. A., & Roemer, J. E. (2012, April 22). *Evolutionary Stability of Kantian Optimization*. 2112098. https://doi.org/10.2139/ssrn.2112098

Denes, M. R., Karpoff, J. M., & McWilliams, V. B. (2017). Thirty years of shareholder activism: A survey of empirical research. *Journal of Corporate Finance*, *44*, 405–424.

Dimson, E., Karakaş, O., & Li, X. (2015). Active Ownership. *The Review of Financial Studies*, *28*(12), 3225–3268.

Dimson, E., Karakaş, O., & Li, X. (2025). Coordinated Engagements. *The Journal of Finance*.

Dyck, A., Lins, K. V., Roth, L., & Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? International evidence. *Journal of Financial Economics*, *131*(3), 693–714.

Edmans, A., Levit, D., & Schneemeier, J. (2023, July 12). *Socially Responsible Divestment*. 4093518. https://doi.org/10.2139/ssrn.4093518

Edmans, A., & Manso, G. (2011). Governance Through Trading and Intervention: A Theory of Multiple Blockholders. *The Review of Financial Studies*, *24*(7), 2395–2428.

Enke, B., Rodríguez-Padilla, R., & Zimmermann, F. (2020, July). *Moral Universalism and the Structure of Ideology*. 27511. https://doi.org/10.3386/w27511

Fahlenbrach, R., Rudolf, N., & Wegerich, A. (2023). *Leading by Example: Can One Universal Shareholder's Voting Pre-Disclosure Influence Voting Outcomes?* https://doi.org/10.2139/ssrn.4660355

Fisch, J. E., & Schwartz, J. (2023, May 30). *Corporate Democracy and the Intermediary Voting Dilemma*. 4360428. Retrieved November 2, 2025, from https://papers.ssrn.com/abstract=4360428

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79.

Goldstein, I., & Huang, C. (2016). Bayesian Persuasion in Coordination Games. *American Economic Review*, *106*(5), 592–596.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Goswami, R. (2024, April 2). *Disney's largest shareholder Vanguard reportedly backing management over Peltz in board fight*. CNBC. Retrieved June 12, 2025, from https://www.cnbc.com/2024/04/02/disney-proxy-fight-vanguard-voting-for-management-over-peltz.html

Gramitto Ricci, S. A., Greenwood, D. J. H., & Sautter, C. M. (2025, February 18). *The Shareholder Democracy Lie*. 5143857. https://doi.org/10.2139/ssrn.5143857

Hartzmark, S. M., & Sussman, A. B. (2019). Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows. *The Journal of Finance*, *74*(6), 2789–2837.

Heinkel, R., Kraus, A., & Zechner, J. (2001). The Effect of Green Investment on Corporate Behavior. *The Journal of Financial and Quantitative Analysis*, *36*(4), 431–449.

Hoepner, A. G. F., Oikonomou, I., Sautner, Z., Starks, L. T., & Zhou, X. Y. (2024). ESG shareholder engagement and downside risk. *Review of Finance*, *28*(2), 483–510.

Hong, H., & Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics*, *93*(1), 15–36.

Hooker, B. (1990). Rule-Consequentialism. *Mind*, *99*(393), 67–77.

Hume, D. (1740). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*. A M S Press, Incorporated.

Invesco. (2025, May). *Invesco's Policy Statement on Global Corporate Governance and Proxy Voting*. Invesco.

Jarvis, A., & Forster, P. M. (2024). Estimated human-induced warming from a linear temperature and atmospheric CO2 relationship. *Nature Geoscience*, *17*(12), 1222–1224.

Johnson, L. (2024, August 22). *BlackRock's support for E+S proposals whittles away in 2024 proxy season — ESG Dive*. Retrieved June 12, 2025, from https://www.esgdive.com/news/blackrock-2024-investment-stewardship-report-declining-environmental-social-support/725010/

Johnson, T. L., & Swem, N. (2021). Reputation and investor activism: A structural approach. *Journal of Financial Economics*, *139*(1), 29–56.

Kant, I. (1788). *Critique of Practical Reason* (M. Gregor, Trans.; 2nd ed.). Cambridge University Press.

Kant, I. (1797). *Groundwork of the metaphysics of morals* (M. J. Gregor, Trans.). Cambridge University Press.
OCLC: 37361734.

Kubany, E. S., & Watson, S. B. (2003). Guilt: Elaboration of a multidimensional model. *The Psychological Record*, *53*(1), 51–90.

Ledyard, J. O. (1994). Public Goods: A Survey of Experimental Research. *Public Economics*.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.

McGowan, J. (2024, August 26). *BlackRock Didn't Pull Support Of ESG, Just 'Poor Quality' Shareholder Proposals*. Forbes. Retrieved June 12, 2025, from https://www.forbes.com/sites/jonmcgowan/2024/08/26/blackrock-didnt-pull-support-of-esg-just-poor-quality-shareholder-proposals/

Michl, P., Meindl, T., Meister, F., Born, C., Engel, R. R., Reiser, M., & Hennig-Fast, K. (2014). Neurobiological underpinnings of shame and guilt: A pilot fMRI study. *Social Cognitive and Affective Neuroscience*, *9*(2), 150–157.

Morris, S., & Shin, H. S. (1998). Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks. *The American Economic Review*, *88*(3), 587–597.

Morris, S., & Shin, H. S. (2002). Social Value of Public Information. *American Economic Review*, *92*(5), 1521–1534.

Niszczota, P., Conway, P., & Białek, M. (2024). Moral decay in investment. *Journal of Experimental Social Psychology*, *115*, 104664.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science (New York, N.y.)*, *314*(5805), 1560–1563.

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

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

Pedersen, L. H. (2025, September 30). *Can Sustainable Finance Save the Planet?* 5549819. Retrieved October 3, 2025, from https://papers.ssrn.com/abstract=5549819

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2020). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics*.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, *83*(5), 1281–1302.

Riedl, A., & Smeets, P. (2017). Why Do Investors Hold Socially Responsible Mutual Funds? *The Journal of Finance*, *72*(6), 2505–2550.

Roemer, J. E. (2010). Kantian Equilibrium. *The Scandinavian Journal of Economics*, *112*(1), 1–24.

Roemer, J. E. (2012, March 1). *Kantian Optimization, Social Ethos, and Pareto Efficiency*. 2021366. Retrieved May 9, 2023, from https://papers.ssrn.com/abstract=2021366

Roemer, J. E. (2013, October 25). *Kantian Optimization: An Approach to Cooperative Behavior*. 2345335. Retrieved May 9, 2023, from https://papers.ssrn.com/abstract=2345335

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, *127*, 45–57.

Roemer, J. E., & Silvestre, J. (2023). Kant and Lindahl. *The Scandinavian Journal of Economics*, *1*(32).

Ross, W. D. (2002). *The Right and the Good*. Clarendon Press.

Roy, M., & Binnie, I. (2024, May 20). *CalPERS to vote against Exxon board members — Reuters*. Reuters. Retrieved June 12, 2025, from https://www.reuters.com/business/calpers-vote-against-exxon-board-members-2024-05-20/

Solsvik, T. (2024). Norway's wealth fund asks Shell for more climate policy details [newspaper]. *Reuters: Sustainability*.

Sugden, R. (2003). The Logic of Team Reasoning. *Philosophical Explorations*, *6*(3), 165–181.

Sun, L., & Shapiro, J. M. (2022). A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure. *Journal of Economic Perspectives*, *36*(4), 193–204.

Sustainalytics, M. (2024). *Voting on ESG: A Gap Becomes a Gulf*. Sustainalytics. Retrieved June 12, 2025, from https://connect.sustainalytics.com/voting-on-esg-a-gap-becomes-a-gulf

Tenenbaum, S. (2017). Action, Deontology, and Risk: Against the Multiplicative Model. *Ethics*, *127*(3), 674–707.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853–870.

White, A. (2024, January 22). *Asset managers inconsistent on labour rights: 2023 proxy voting results - Top1000funds.com*. Top1000 Funds. Retrieved June 12, 2025, from https://www.top1000funds.com/2024/01/asset-managers-inconsistent-on-labour-rights-2023-proxy-voting-results/

Wilson, J. (2025, May 9). *2024 Stewardship Report — Morgan Stanley*. Morgan Stanley Investment Management. Retrieved June 12, 2025, from https://www.calvert.com/insights/blog/calvert-2024-stewardship-report.html

Yang, K., & Yasuda, A. (2023, November 19). *Decoding Sustainable Investment Strategies: Bridging Intentions and Outcomes*. 4637264. https://doi.org/10.2139/ssrn.4637264

Zingales, L., Hart, O., & Landemore, H. E. (2025, July 31). *How To Implement Shareholder Democracy*. 5039736. https://doi.org/10.2139/ssrn.5039736

A Proofs and Derivations

A.1 Generalized Offset Rule

As an alternative to the linear model in the main text (Section 2.3), I propose here a generalized offset rule that captures nonlinear coordination dynamics.

The quantity of externality the firm is obligated to offset is modeled as the outcome of a coordination game among utilitarian investors, since deontologists do not engage in activism. Reflecting institutional mechanics, offset purchases become more likely as support for a proposal increases. The firm's offset obligation is modeled as a smooth, rescaled sigmoid function of investor support: initially convex, then concave, capturing coordination dynamics with increasing and diminishing marginal efficacy of activism.

Let $\tau \in [0,1]$ denote the fraction of votes cast in favor of the proposal, and let $\overline{\tau} \in (0,1)$ be the notional threshold beyond which firms become responsive to activism. Then the offset quantity is given by:

$$\Delta\lambda(\tau) = \bar{\lambda} \cdot \left(\tfrac{2}{1+\exp(-\kappa(\tau-\overline{\tau}))} - 1\right), \tag{34}$$

where $\bar{\lambda} < \lambda$ is the maximum feasible offset and $\kappa > 0$ governs the curvature of the response. The function satisfies $\Delta\lambda(\overline{\tau}) = \frac{1}{2}\bar{\lambda}$ and asymptotes toward 0 as $\tau \to 0$ and toward $\bar{\lambda}$ as $\tau \to 1$. For sufficiently large κ, the offset response approximates a threshold-triggered step function, while remaining smooth and differentiable.

Incorporating both share-based and headcount-based coordination, the perceived offset for agent i is given by the following generalized offset rule.

Definition 5 (Sigmoid Offset Rule)**.** *The externality offset perceived by agent i is given by:*

$$\Delta\lambda_i = \mu \cdot \Delta\lambda(\tau_i^s) + (1-\mu) \cdot \Delta\lambda(\tau_i^n), \tag{35}$$

$$\tau_i^s = \tfrac{\sum_j I_j^U w_j \tilde{\phi}_j^i}{P}, \tag{36}$$

$$\tau_i^n = \tfrac{\sum_j I_j^U \mathbf{1}\{\tilde{\phi}_j^i > 0\}}{n}, \tag{37}$$

$$\tilde{\phi}_j^i = \begin{cases} \phi_j & \textit{if } I_i^N = 1 \\ \phi_i & \textit{otherwise} \end{cases}, \tag{38}$$

where $\mu \in [0,1]$ determines the relative weight assigned to share-based versus headcount-based coordination, and all other variables are as previously defined.

The resulting offset rule is smooth, bounded, and behaviorally plausible. It captures both institutional inertia (via $\overline{\tau}$), increasing coordination efficacy (via κ), and the tension between wealth-driven influence and egalitarian norms of moral agency. However, the added complexity complicates the derivation of closed-form solutions for optimal portfolio weights and market-clearing prices. Therefore, I retain the linear offset rule in the main text for analytical tractability, while noting that the generalized sigmoid rule can be implemented numerically to explore robustness and richer coordination dynamics.

A.2 Proof of Proposition 1: Marginal Moral Costs by Type

First, consider the marginal effect of an investor's portfolio weight ϕ_i on their perceived reduction in externality $\Delta\lambda_i$.

Lemma 1 (Ordering of Marginal Effects on externality by Investor Type)**.** *The marginal effect of an investor's portfolio weight ϕ_i on their perceived reduction in externality $\Delta\lambda_i$ is ordered as follows for any given investor i and $\mu > 0$:*

$$\left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{\textit{Kant-Util}} > \left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{\textit{Kant-Deon}} > \left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{\textit{Nash-Util}} > \left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{\textit{Nash-Deon}}. \tag{39}$$

For $\mu = 0$, all marginal effects equal zero as headcount-weighted coordination implies that the investor's portfolio weight does not affect externality.

Proof. Differentiating $\Delta\lambda_i$ with respect to ϕ_i gives

$$\frac{\partial\Delta\lambda_i}{\partial\phi_i} = \begin{cases} \frac{\mu w_i \lambda}{P} & \text{if } i \in \text{Nash-Util}, \\ 0 & \text{if } i \in \text{Nash-Deon}, \\ \frac{\mu\left(w_i + \sum_{j\neq i} I_j^U w_j\right)\lambda}{P} & \text{if } i \in \text{Kant-Util}, \\ \frac{\mu\left(\sum_{j\neq i} I_j^U w_j\right)\lambda}{P} & \text{if } i \in \text{Kant-Deon}. \end{cases} \tag{40}$$

Given $\mu, w_i, P > 0$, the ordering follows directly from the nesting of wealth terms across types.[5] □

Next, consider the first order condition for the investor's optimization problem.

Lemma 2. *First order condition for the investor's problem in Definition 3 is given by*

$$\mathbb{E}\left[\frac{\partial u(w_i')}{\partial w_i'}\frac{\partial w_i'}{\partial\phi_i}\right] = I_i^U \frac{\partial v(\lambda - \Delta\lambda_i)}{\partial\Delta\lambda_i}\left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right) + (1 - I_i^U)(1 - I_i^N)\frac{\partial\chi(\phi_i, \lambda)}{\partial\phi_i}$$
$$+ (1 - I_i^U)I_i^N\left(\frac{\partial\chi(\phi_i, \lambda - \Delta\lambda_i)}{\partial\phi_i} + \frac{\partial\chi(\phi_i, \lambda - \Delta\lambda_i)}{\partial\Delta\lambda_i}\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right), \tag{41}$$

or equivalently,

$$\mathbb{E}\left[\frac{\partial u(w_i')}{\partial w_i'}\frac{\partial w_i'}{\partial\phi_i}\right] = \begin{cases} \frac{\partial v(\lambda - \Delta\lambda_i)}{\partial\Delta\lambda_i}\left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{NU} & \textit{if } i \in \textit{Nash-Util}, \\ \frac{\partial\chi(\phi_i, \lambda - \Delta\lambda_i)}{\partial\phi_i} & \textit{if } i \in \textit{Nash-Deon}, \\ \frac{\partial v(\lambda - \Delta\lambda_i)}{\partial\Delta\lambda_i}\left(\frac{\partial\Delta\lambda_i}{\partial\phi_i}\right)_{KU} & \textit{if } i \in \textit{Kant-Util}, \\ \frac{\partial\chi(\phi_i, \lambda)}{\partial\phi_i} & \textit{if } i \in \textit{Kant-Deon}. \end{cases} \tag{42}$$

where the right hand side is the marginal moral cost of increasing portfolio weight ϕ_i for each type of investor, and $\frac{\partial\Delta\lambda_i}{\partial\phi_i}$ is the marginal effect of increasing portfolio weight on the investor's perceived reduction in externality.

[5] The summation operator appears for Kant investors as $\tilde{\phi}_j = \phi_i \ \forall i \in \text{Kant}$. The perceived offset function for Kant investors is differentiable almost everywhere. At interior solutions where $\phi_i > 0$, both the share-weighted and headcount-weighted terms are smooth in ϕ_i, enabling marginal effect comparisons. At the boundary $\phi_i = 0$, the headcount based coordination game introduces a discontinuity; however, equilibrium analysis focuses on interior solutions, and smoothing approximations (e.g., sigmoid developed in Appendix Section A.1) can recover differentiability if needed. This ensures tractability of the Kant FOC without loss of generality.

Proof. The first order condition follows from differentiating the Lagrangian of the investor's problem in Definition 3 with respect to ϕ_i and setting it equal to zero. The equivalence between Eq. (41) and Eq. (42) follows from substituting the indicator functions I_i^U and I_i^N with their respective values for each investor type. □

The ordering of marginal moral costs by type follows directly. Let C_τ be the marginal moral cost of increasing portfolio weight for an agent of type $\tau \in \{NU, ND, KU, KD\}$. The ordering of C_τ by type for any given level of ϕ_i is

$$C_{\text{KU}}(\phi_i) < C_{\text{NU}}(\phi_i) < 0 < C_{\text{ND}}(\phi_i) < C_{\text{KD}}(\phi_i). \tag{43}$$

The ordering of $C_\tau(\phi_i)$ is equivalent to the ordering of the right-hand side of the first order condition in Lemma 2, which is in turn determined by the following properties that hold by construction:

$$\frac{\partial v}{\partial \Delta\lambda_i} < 0, \quad \frac{\partial \chi}{\partial \phi} > 0 > \frac{\partial \chi}{\partial \Delta\lambda_i}, \quad \text{and } \frac{\partial \chi}{\partial \phi} > \left| \frac{\partial \chi}{\partial \Delta\lambda_i} \right|. \tag{44}$$

Together with the ordering of marginal effects on externality stated in Lemma 1, the stated ordering of marginal moral cost follows.

As a corollary, the optimal portfolio weights are ordered by type as follows, provided wealth and risk aversion are identical across types. This follows from the first order condition equating marginal benefit and marginal cost of increasing portfolio weight. With pecuniary marginal benefit identical across types, the ordering of marginal moral costs implies the ordering of optimal portfolio weights.

$$\phi^*_{KU} < \phi^*_{NU} < \phi^*_{ND} < \phi^*_{KD}. \tag{45}$$

□

A.3 Closed Form Optimal Portfolio Weights with CRRA Utility

Let utility from wealth be given by a constant relative risk aversion (CRRA) utility function, $u(w) = \frac{w^{1-\gamma}}{1-\gamma}$ where γ is the coefficient of relative risk aversion. Recall that the moral disutility functions are specified as

$$\begin{aligned} v(\lambda - \Delta\lambda_i) &= \frac{1}{2}a(\lambda - \Delta\lambda_i)^2, \\ \chi(\phi_i, \lambda) &= \frac{1}{2}b_1\lambda\phi_i^2 - I_i^N b_2 \frac{\Delta\lambda_i}{\lambda}\phi_i \end{aligned} \tag{46}$$

where $a, b_1, b_2 > 0$ and $b_1 > b_2$.

The marginal moral cost of increasing portfolio weight for each type is given by

$$C_{NU}(\phi_{NU}) = -\frac{2a\lambda^2\mu w_{NU}\left(P - \left(w_{NU}\phi_{NU} + w_{KU}\phi_{KU}\right)\right)}{P^2}, \tag{47}$$

$$C_{ND}(\phi_{ND}) = \frac{b_1 P\lambda\mu\phi_{ND} - b_2\mu\left(w_{NU}\phi_{NU} + w_{KU}\phi_{KU}\right)}{P}, \tag{48}$$

$$C_{KU}(\phi_{KU}) = -\frac{2a\lambda^2\mu\left(w_{NU} + w_{KU}\right)\left(P - \left(w_{NU}\phi_{KU} + w_{KU}\phi_{KU}\right)\right)}{P^2}, \tag{49}$$

$$C_{KD}(\phi_{KD}) = b_1\lambda\mu\phi_{KD}. \tag{50}$$

The left-hand side of the first order condition in Lemma 2 represents the pecuniary marginal benefit of increasing portfolio weight, and is given by

$$\mathbb{E}\left[(w'_\tau)^{-\gamma} w_\tau \left(\frac{\theta - q\Delta\lambda}{P} - \frac{q}{P}\frac{\partial\Delta\lambda}{\partial\phi_\tau} - R_f\right)\right] \tag{51}$$

$$= w_\tau^{1-\gamma}\mathbb{E}\left[\left(R_f + \phi_\tau\left(\frac{\theta - q\Delta\lambda}{P} - R_f\right)\right)^{-\gamma}\left(\frac{\theta - q\Delta\lambda}{P} - \frac{q}{P}\frac{\partial\Delta\lambda}{\partial\phi_\tau} - R_f\right)\right] \tag{52}$$

$$\approx w_\tau^{1-\gamma} R_f^{-\gamma}\left(\mathbb{E}\left(\frac{\theta - q\Delta\lambda}{P} - \frac{q}{P}\frac{\partial\Delta\lambda}{\partial\phi_\tau} - R_f\right) - \gamma R_f^{-1}\mathbb{E}\left(\frac{\theta - q\Delta\lambda}{P} - R_f\right)^2\phi_\tau\right), \tag{53}$$

with $\phi_\tau \in \{\phi_{NU}, \phi_{ND}, \phi_{KU}, \phi_{KD}\}$. The approximation is valid for small ϕ_τ. Note that the marginal moral cost $C_\tau(\phi^*_\tau)$ is linear in the investors' portfolio weights, and this allows us to solve for the optimal portfolio weight ϕ^*_τ in closed form as a function of price P for each investor type.

The system of equations determining optimal portfolio weights is given by setting Eq. (53) equal to the respective marginal moral costs in Eqs. (47) to (50) for each investor type. Rearranging terms yields the following system of four linear equations in four unknowns:

$$\phi^*_{NU} = \frac{P^2 R_f}{\gamma \mathbb{V}(\theta)} \Bigg[\frac{\mathbb{E}(\theta) a \lambda \mu R_f^\gamma w_{NU}^\gamma \left(-\frac{1}{2}\lambda(1-\mu)+\lambda-\frac{\lambda\mu(w_{NU}\phi^*_{NU}+w_{KU}\phi^*_{KU})}{P} \right)}{P} + \frac{\mathbb{E}(\theta)-q\left(\frac{1}{2}\lambda(1-\mu)+\frac{\lambda\mu(w_{NU}\phi^*_{NU}+w_{KU}\phi^*_{KU})}{P}\right)-\frac{\lambda\mu q w_{NU}}{P}}{P} - R_f \Bigg], \tag{54}$$

$$\phi^*_{ND} = \frac{P^2 R_f}{\gamma \mathbb{V}(\theta)} \Bigg[-R_f^\gamma w_{ND}^{\gamma-1} \left(\mathbb{E}(\theta) b_1 \lambda \phi^*_{ND} - \frac{b_2\left(\frac{1}{2}\lambda(1-\mu)+\frac{\lambda\mu(w_{NU}\phi^*_{NU}+w_{KU}\phi^*_{KU})}{P}\right)}{\lambda} \right) + \frac{\mathbb{E}(\theta)-q\left(\frac{1}{2}\lambda(1-\mu)+\frac{\lambda\mu(w_{NU}\phi^*_{NU}+w_{KU}\phi^*_{KU})}{P}\right)}{P} - R_f \Bigg], \tag{55}$$

$$\phi^*_{KU} = \frac{P^2 R_f}{\gamma \mathbb{V}(\theta)} \Bigg[-R_f^\gamma w_{KU}^{\gamma-1} \left(-\frac{\mathbb{E}(\theta) a \lambda \mu (w_{NU}+w_{KU})\left(-\frac{1}{2}\lambda(1-\mu)+\lambda-\frac{\lambda\mu(w_{NU}\phi^*_{KU}+w_{KU}\phi^*_{KU})}{P}\right)}{P} \right) + \frac{\mathbb{E}(\theta)-q\left(\frac{1}{2}\lambda(1-\mu)+\frac{\lambda\mu(w_{NU}\phi^*_{KU}+w_{KU}\phi^*_{KU})}{P}\right)-\frac{\lambda\mu q(w_{NU}+w_{KU})}{P}}{P} - R_f \Bigg], \tag{56}$$

$$\phi^*_{KD} = \frac{P^2 R_f}{\gamma \mathbb{V}(\theta)} \Bigg[-R_f^\gamma w_{KD}^{\gamma-1} \left(\mathbb{E}(\theta) b_1 \lambda \phi^*_{KD} \right) + \frac{\mathbb{E}(\theta)-q\left(\frac{1}{2}\lambda(1-\mu)+\frac{\lambda\mu(w_{NU}\phi^*_{KD}+w_{KU}\phi^*_{KD})}{P}\right)-\frac{\lambda\mu q(w_{NU}+w_{KU})}{P}}{P} - R_f \Bigg]. \tag{57}$$

Given a price P, this is a system of four linear equations in as many unknowns, and can be solved using standard linear algebra techniques. The resulting demand curves, $\phi_i^*(P)$, are continuous and differentiable functions of the market clearing price P. The system is well-posed under the assumptions of the model, and the solutions are unique for each agent type.

A.4 Proof of Proposition 3: Equilibrium Uniqueness

With CRRA pecuniary utility and quadratic moral disutility functions be specified as in Eq. (46), a unique Nash-Kant equilibrium exists, and is globally unique under a sufficient condition derived below.

To establish global uniqueness, consider the sources of multiplicity:

1. Multiple market-clearing prices P^* for a given strategy profile Φ^*,
2. Multiple optimal strategy profiles Φ^* for a given price P^*,
3. Self-confirming beliefs and recursive feedback loops, and
4. Global fixed-point structure of the price–strategy mapping.

The proof proceeds by considering each of these sources of multiplicity in turn.

Price given strategy profile. Market clearing price is a linear function of portfolio weights. Given any sequence of portfolio weights Φ^*, there exists a unique market clearing price.

Strategy profile given price. Each agent's optimal allocation is linear in the allocations of other agents, given price P^*. Linear algebra guarantees a unique solution for the system of equations as shown in Section A.3. Therefore, given price P^*, there exists a unique portfolio allocation rule for each agent type $\phi_i^*(P^*)$.

Belief structure. Given a price P, Kant portfolio weights are uniquely determined by their first order conditions, as Kant agents do not adjust their portfolio weights based on the actions of others. However, the fixity of Kant agents creates an asymmetry that may allow the Nash agents to coordinate on multiple consistent beliefs—each believing the other will play a certain way, and best-responding accordingly.

Here, note that best responses of Nash agents are linear in the portfolio weights of other agents. The representative agent model implies there is no room for within-type multiplicity of beliefs, so the feedback loop if any comes from mutual reinforcement between Nash utilitarians and Nash deontologists. Nash utilitarians exhibit strategic substitutability, while Nash deontologists exhibit strategic complementarity. As a result, any feedback loop between Nash utilitarians and deontologists is, directionally speeaking, mutually cancelling rather than reinforcing. Therefore, the mapping from beliefs to best responses among Nash agents is a contraction mapping under any reasonable parameterization, and Banach's fixed point theorem guarantees convergence to a unique fixed point of beliefs given price P^*.

Global fixed-point structure of the price–strategy mapping. To establish global uniqueness, it suffices to show that the aggregate excess demand function intersects the horizontal axis at most once in the economically feasible region S, where $S = \{P \in \mathbb{R}^+ : P \in (0, \sum_\tau w_\tau], \phi_\tau(P) \in [0,1] \;\; \forall \tau\}$. In words, S is the set of all prices at which every agent demands a non-negative quantity of the risky asset, and does not demand more than their total wealth.

The demand curves for each agent type, and the aggregate excess demand function, are shown in Fig. 9 under the baseline parameterization summarized in Table 2. As noted earlier, the ranking of demand curves by type is invariant to parameter values, and follows from the ordering of marginal moral costs established in Proposition 1. Constraints on price can therefore be derived with reference to the Kant Deontologist's demand curve alone, as the Kant Deontologist is the key marginal agent determining the boundaries of S.

For the CRRA-quadratic utility combination, substitute $\phi_{KD}(P_0) = 0$ in the Kant Deontologist's optimal portfolio weight equation in Eq. (57). This yields a quadratic equation in P. Let P_0 be its lower root, and note that P_0 is uniquely determined given model parameters. The uniqueness condition can then be expressed as $\Phi(P_0) > P_0$. Although the closed form expression of this condition

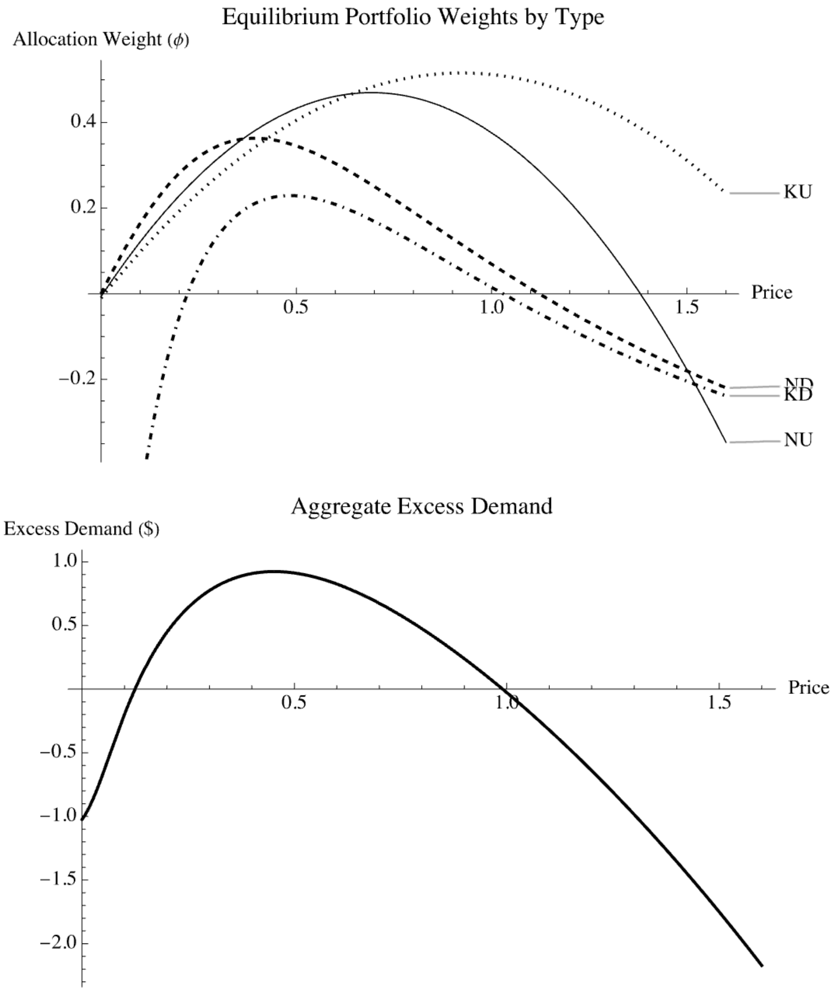


Figure 9: $\phi^*_\tau(P)$ for each investor type τ and aggregate excess demand $\Phi(P^*) - P^*$ under CRRA utility from wealth and quadratic moral costs.

is algebraically cumbersome, it can be numerically computed easily for any given calibration.

Lemma 3 (Kant Deontologist determines boundaries of S)**.** *Let P_0 solve $\phi_{KD}(P) = 0, \quad \phi'_{KD}(P) > 0$, P_1 solve $\phi_{KD}(P) = 1, \quad \phi'_{KD}(P) < 0$, and $S = \{P \in \mathbb{R}^+ : P \in (0, \sum_\tau w_\tau], \phi_\tau(P) \in [0,1] \ \forall \tau\}$, the economically feasible region. Then, it must be that market clearing price $P^* \in [P_0, P_1]$.*

Proof. By definition of P_0, for any $P < P_0$, it follows that $\phi_{KD}(P) < 0$, so any $P < P_0$ cannot lie in S. Similarly, for any $P > P_1$, it follows that $\phi_{KD}(P) < 0$, so any $P > P_1$ cannot lie in S. Therefore, $S = [P_0, P_1]$. If a market clearing equilibrium price P^* exists, it must lie in S, so $P^* \in [P_0, P_1]$. □

Given the feasible region $[P_0, P_1]$ determined above, a sufficient condition for uniqueness follows.

Proposition 5 (Sufficient Condition for Uniqueness of Nash-Kant Equilibrium)**.** *Let P_0 and P_1 be as defined in Lemma 3, and let $\Phi(P)$ denote the aggregate invested capital across all agents as a function of price P. A sufficient condition for uniqueness of the mixed Nash-Kant equilibrium is that*

$$\Phi(P_0) > P_0, \qquad \Phi(P_1) < P_1 \tag{58}$$

Proof. From Lemma 3, any market clearing price P^* must lie in the feasible region $S = [P_0, P_1]$. The aggregate excess demand function is parabolic in shape and intersects the horizontal axis at most twice in $[0, \sum_\tau w_\tau]$. Therefore, if $\Phi(P_0) > P_0$ and $\Phi(P_1) < P_1$, it follows that the aggregate excess demand function intersects the horizontal axis exactly once in the feasible region S. Hence, there exists a unique market clearing price $P^* \in S$. Given P^*, there exists a unique portfolio allocation rule for each agent type $\phi_i^*(P^*)$. Therefore, there exists a unique Nash-Kant equilibrium. □

B Kant Equilibrium: Definitions and Properties

B.1 Theoretical Foundations

Philosophical foundations clarify the two orthogonal axes used in this paper to characterize ethics. Bentham (2017)'s utilitarianism grounds moral imperatives that minimize aggregate harm (consistent with engagement to reduce externalities), while deontological strands **kantCritiquePracticalReason1883**; Ross (2002) and Tenenbaum (2017) stress complicity-avoidance and side-constraints. Hume (1740)'s Treatise is canonical background for this paper's treatment of belief, motivation, and norm internalization. Empirically oriented cognitive work shows that universalization is not just philosophy: humans' moral judgments systematically track "what if everyone did that?", lending behavioral plausibility to Kant-type operators embedded in economic models (Enke et al., 2020; Levine et al., 2020; Michl et al., 2014).

Methodologically, this paper builds on Roemer's foundational work on Kant optimization, which formalizes cooperation by positing agents who act as if their strategies were universal laws (Roemer, 2010, 2012, 2013, 2015). While Roemer focuses on Kant equilibrium in stylized, homogeneous settings, real markets are populated by a mix of cooperative and self-interested actors. The current work operationalizes Kantian reasoning in a heterogeneous-agent game, embedding moral types in a standard Bayesian environment and showing how their presence shifts outcomes without necessitating universal adoption of Kant logic. The result is a hybrid equilibrium concept that quantifies the impact of norm-driven reasoning in a market context, addressing critiques that Kant logic is too idealistic or fragile for realistic economic environments.

B.2 Definitions

I begin with a general formulation of Kant equilibrium in strategic games, adapted from Roemer (2015) and Braham and Hees (2020). Let N denote the set of players, S the strategy space, and $V^i : S^n \to \mathbb{R}$ the payoff function of player $i \in N$.

Definition 6 (Kant Equilibrium)**.** *A strategy profile* $\mathbf{s}^* = (s_1^*, \ldots, s_n^*) \in S^n$ *is a* Kant equilibrium *if for every player* $i \in N$*,*

$$V^i(s_i^*, \ldots, s_i^*) \geq V^i(s, \ldots, s) \quad \text{for all } s \in S.$$

This definition captures the core idea: the evaluation of any deviation is made under the counterfactual that *everyone* would deviate in the same way. In other words, a Kant equilibrium is a fixed point of moral reasoning, where each player's strategy is one they would endorse as a universal law.

For comparison, recall the definition of Nash equilibrium:

Definition 7 (Nash Equilibrium)**.** *A strategy profile* $\mathbf{s}^* = (s_1^*, \ldots, s_n^*) \in S^n$ *is a* Nash equilibrium *if for every player* $i \in N$*,*

$$V^i(s_i^*, s_{-i}^*) \geq V^i(s_i, s_{-i}^*) \quad \textit{for all } s_i \in S.$$

In this sense, Kant equilibrium is a fixed point of moral reasoning rather than strategic expectation. It requires no assumption about the actual behavior of others, nor any probabilistic belief about likely responses. What matters is whether a given strategy can be morally endorsed as a universal law. This makes the Kant solution concept uniquely suited to modeling forms of ethical behavior that are resistant to incentives, non-reciprocal in structure, and grounded in moral principle rather than empirical expectations.

The abstract definition of Kant equilibrium has been articulated in several intuitive and behaviorally grounded ways in the literature. These interpretations help clarify the underlying reasoning structure that distinguishes Kant agents from Nash optimizers:

Roemer (2015). *"If I want to deviate from a contemplated action profile of all agents, I may do so only if I would have all others deviate in like manner."*

Roemer and Silvestre (2023). *"I choose the amount of contribution that I would like everybody to make," or equivalently, no agent prefers that everybody alter their strategy by the same amount.*

Roemer (2013). *"If I were to deviate from my stipulated action, and all others were to deviate in like manner from theirs, I would not prefer the consequences of the new action profile to the original."*

Curry and Roemer (2012). *"I have a right to increase (or decrease) my action by a factor if and only if I would be happy if all others changed their action by the same factor in the same direction."*

These articulations share a common structural feature: the permissibility of deviation is evaluated not individually, but universally. Deviation is only allowed if one is willing to see it generalized to all. Roemer defines the deviation in additive and multiplicative terms, allowing for a flexible formulation based on context.

Braham and Hees (2020) calls such an equilibrium a subjective Kant equilibrium, arguing that the Kant agent is not truly cooperative in the sense that she solves for a strategy that, if chosen by all, would be optimal *for her*, but not necessarily *for all*, and further, the subjective Kant equilibrium imposes no requirement that the agents' beliefs be correct. They contrast the solution concept with the more restrictively defined Simple Kant equilibrium in which all players are required to choose the same universalizable strategy, which causes any Kant equilibrium to be self-fulfilling.

Definition 8 (Simple Kant equilibrium)**.** *Consider a game with a set of players* N*, strategy space* S *and payoff function* $V^i : S^n \to \mathbb{R}$ *for each player* $i \in N$*. A strategy profile* $\mathbf{s}^* \equiv (s_1^*, \ldots, s_n^*) \in S^n$ *is a Kant equilibrium if* $V^i(s_i^*, \ldots, s_i^*) \geq V^i(s, \ldots, s) \quad \forall i \in N, \ \forall s \in S$ *and every player weakly prefers the same universal strategy, i.e.* $s_i^* = s^* \ \forall i$

The authors argue, rightly, that Kant's categorical imperative is a test of maxims—the justifying principles behind actions—rather than the actions themselves. They caution against conflating the universalizability of strategies with that of reasons, especially in environments where actions may encode multiple or conflicting maxims. This critique is both philosophically rigorous and substantively important.

That said, the application in this paper—portfolio allocation in an externality environment—features a tightly constrained and ethically transparent action space, where the mapping from moral principle to strategy is effectively one-to-one. In this context, the universalizability of an action (portfolio weight) faithfully represents the universalizability of its underlying maxim (participation in, or abstention from, externality-generating investment). The goal here is not to reproduce the full normative scope of Kant ethics, but to translate its core reasoning logic into a tractable formal device. This approach builds directly on the intellectual foundation laid by Braham, Hees, Roemer, and others, and any departure from their frameworks should be read as context-specific rather than conceptual dissent.

B.3 A Note on Strategic Interdependence

While the Kant agent acts instrumentally within a normatively constrained strategy set, in the current implementation, the universalizability test constrains the agent's strategy set to a singleton. This invites the charge that the Kant agent is behaviorally static, and the model is strategically vacuous, as strategic interdependence is lost in behavior, even if it survives in logic. This paper rescues strategic interdependence without sacrificing normative integrity using three extensions, discussed in detail in Section 5. introducing non-triviality in the admissible test through fuzzy, imperfect or context-dependent universalizability, allowing for dynamic Kantian reasoning where the admissible strategy set evolves over time, and modeling a meta-game of reasoning-type selection where agents choose between acting as Kant or Nash players based on perceived norm uptake. These extensions restore the richness of strategic interaction while preserving the Kant logic of moral reasoning. The core idea remains: Kant agents evaluate strategies based on their universalizability, not empirical beliefs, and this normative constraint shapes their actions in a way that is distinct from Nash equilibrium.

Empirically, Kant reasoning is observed among institutional investors who pre-commit to specific behaviors, thus precluding themselves from reacting to observed actions of other market participants. Such pre-commitment can take the form of public declarations of voting intentions or adherence to specific ethical guidelines which are not contingent on others' actions. Norges Bank Investment Management (NBIM), the manager of Norway's sovereign wealth fund, publicly announced in 2023 that it would vote against a climate resolution at BP's AGM five days before the vote (Adomaitis, 2023). Likewise, in 2024, NBIM disclosed it would not support a climate resolution at Shell's AGM, stating Shell's updated strategy was already 'sufficiently' Paris-aligned (Solsvik, 2024). These commitments are not contingent best responses. They are public declarations of norm-based strategies that, crucially, do not depend on what other voters do.

B.4 Kant and Nash Equilibria in The Prisoner's Dilemma

Consider the two-player Prisoner's Dilemma.

	C	D
C	(R, R)	(S, T)
D	(T, S)	(P, P)

where C and D are the choices of the players, R is the reward for mutual cooperation, T is the temptation to defect, S is the sucker's payoff, and P is the punishment for mutual defection. The payoffs are ordered as follows: $T > R > P > S$.

Note that under Kant reasoning, the equilibrium strategy profile is immediately determined as (C, C), since cooperating is the only action universally acceptable to both players without contradiction. In this example, the Kant equilibrium is also Pareto efficient, although this need not hold in general.

To recover Kant equilibrium from a modified Nash game, consider the following structure:

Moral Disutility. Introduce a moral utility term, $M(\cdot)$, which captures the moral discomfort associated with the choices made by the players.

$$U_i(a_i, a_j; \omega) = \pi_i(a_i, a_j) + \omega M(a_i, a_j) \tag{59}$$

where $\pi_i(a_i, a_j) \in \{R, S, T, P\}$ is the material payoff from the game, $\omega > 0$ is the strength of Kant moral weight, and $M(a_i, a_j)$ is twice differentiable, concave, and increases in both arguments.

Belief Structure. Under the psychological-game belief that others will mirror the agent's choice, the Kant's problem is to choose $a_i \in \{C, D\}$ to maximise $U_i(a_i, a_i) = \pi_i(a_i, a_i) + \omega M(a_i, a_i)$.

Equilibrium. The equilibrium condition satisfies

$$U_i(a^*, a^*) \geq U_i(a', a') \quad \forall a' \in \{C, D\} \tag{60}$$

The agent chooses C over D iff

$$\omega \geq \omega^* \tag{61}$$

where

$$\omega^* = \tfrac{\pi_i(D,D) - \pi_i(C,C)}{M(C,C) - M(D,D)} = \tfrac{P - R}{M(C,C) - M(D,D)} \tag{62}$$

Existence is guaranteed by the finite action space and uniqueness is guaranteed by the strict ordering of the payoffs and concavity of the moral utility function. The equilibrium is Pareto efficient, as the players are better off cooperating than defecting.

Comparative Statics. The equilibrium condition is non-degenerate, as the agent's moral weight ω is a function of the payoff structure. Consider how the agent's moral weight threshold, ω^* varies with the payoff structure.

$$\frac{\partial\omega^*}{\partial(P-R)} > 0, \quad \frac{\partial\omega^*}{\partial(\Delta M)} < 0 \tag{63}$$

where ΔM is the moral gain from universalized cooperation. Stronger moral weight and bigger moral gains from cooperation lead to a higher threshold for cooperation.

The model predicts cooperation when $\omega > \omega^*$. While ω^* can be estimated from experimental variation in cooperation rates, as in Bénabou et al. (2020) and Carpenter (2021), the moral weight ω is not as easily identified.

Therefore, while the above exercise demonstrates the relationship between Kant and modified Nash equilibria in a canonical setting, it is important to note that this equivalence relies on specific assumptions about the moral utility function and belief structure. In the above construction, the belief structure and moral payoff function are both ad hoc impositions designed to recover Kant logic, and do not arise naturally from first principles, nor do they generalize easily to more complex environments. As such, while Kant equilibrium can be seen as a limit case of psychological Nash, it cannot be recovered by any pure payoff modification in a standard Nash framework without also altering the belief formation.

B.5 Kant Equilibrium Compared to Existing Models

Levine et al. (2020) provide compelling empirical evidence that adults use the principle of universalizability in moral reasoning, and that this principle is not reducible to either outcome-based or rule-based reasoning. These insights are echoed in experimental public goods games (Ledyard, 1994), where many players continue to contribute to the public good even when informed that others are defecting. When asked why, they often report 'wanting to do the right thing' or cite discomfort with violating their own standards—reflecting principled action decoupled from beliefs about others' play. Such agents persist in ethical strategies that violate the best-response condition.

Some motivations for this behavior—such as self-image, guilt, or warm glow—can be modeled by modifying utility functions within standard frameworks. However, agents who act based on the principle of universalizability require a different representational structure: one where beliefs are bracketed and the moral permissibility of action is assessed under the hypothetical that all do the same. Such reasoning cannot be rationalized within any belief-consistent best-response framework unless one allows for distorted, fixed-point, or non-empirical belief structures. The Kant equilibrium provides a tractable solution concept that does away with belief consistency but respects alternative conceptions of rationality rooted in principled commitment.

While the Kant solution concept departs sharply from traditional belief-consistent frameworks, it is not without precedents or relatives. I illustrate below Kant equilibrium's relationship to important modeling approaches in the literature on ethical behavior.

Rule Consequentialism The Kant approach has theoretical precedent in rule consequentialism (Hooker, 1990), which directs agents to act by rules that, if generally adopted, would lead to the best outcomes. Despite subtle differences (e.g., rule consequentialism model is consequentialist in nature, whereas Kant reasoning accommodates both consequentialist and deontological motives), the frameworks share a core structure: both ask agents to justify actions by their hypothetical universalizability, not by belief-updated predictions. This rule-based logic is not reducible to Nash without infinite regress.

Fairness and Social Preferences Rabin (1993) augments standard utility functions with social preferences, where agents receive bonuses for kindness and penalties for unkindness. This formulation allows equilibria to deviate from purely self-interested Nash outcomes by incorporating beliefs about others' intentions and fairness. To explore the connection more precisely, suppose we augment each Kant's utility with a payoff externality from the aggregate level of moral cooperation:

$$U_i^K(\phi;\Phi) \;=\; \underbrace{\mathbb{E}\,U_i(\phi,\Phi)}_{\text{pecuniary payoff}} \;+\kappa\,(\Phi-\Phi_{\text{fair}}) \;-\; \eta\,[\Phi_{\text{fair}}-\Phi]_+ \tag{64}$$

where $\Phi \equiv \phi$ in a Kant equilibrium (by universalizability), Φ_{fair} is a normatively "just" benchmark for cooperation—perhaps the utilitarian social optimum—and $\kappa, \eta > 0$ are parameters capturing kindness gains and under-cooperation penalties, respectively.

This formulation reveals a deep structural relationship between the Rabin fairness model and Kant optimization: both favor actions that promote socially beneficial outcomes, and both discourage self-serving deviation. However, the mechanisms differ. Rabin's agents optimize over beliefs about others' actions, using correct beliefs in a modified utility framework. In contrast, Kant agents do not form probabilistic expectations. They evaluate the strategy *as if* it were to be adopted universally, regardless of what others actually do. This is a structural belief distortion: agents do not just care about fairness—they apply a normative test to their own strategy, assuming it becomes law.

Despite differences, both frameworks yield a common prediction: ethical preferences can sustain cooperative outcomes even in the absence of external enforcement. Just as the kindness-adjusted utility in Rabin's game shifts equilibrium away from the competitive Nash equilibrium towards the cooperative (and Pareto improving) one, the Kant universalizability rule shifts behavior toward emissions-mitigating activism—provided the moral gains from cooperation outweigh pecuniary risks.

Psychological Games In the framework developed by Geanakoplos et al. (1989), utilities depend not only on strategies but also on beliefs—potentially of higher order. Rabin's fairness model is a Nash equilibrium in such a psychological game: payoffs are contingent on beliefs about others' intentions.

Kant equilibrium can also be cast as a psychological game: one in which utility depends on the agent's counterfactual belief that all others will match their own action. Unlike GPS models that encode increasingly accurate higher-order beliefs, the Kant model inserts a fixed-point belief independent of empirical inference. The agent does not expect to be mimicked but chooses as if their action sets the norm. In this sense, the fixed-point structure of Kant beliefs cannot be captured by any finite order of belief updating—no amount of higher-order reasoning in standard frameworks can generate the counterfactual universalizability logic.

Team Reasoning In team reasoning models (Bacharach et al., 2006; Sugden, 2003), players adopt a shared identity and ask "what should *we* do?" rather than "what should *I* do?" Such agents endogenize the group perspective and select strategies that optimize the team's outcome.

While conceptually close, Kant agents differ in one crucial respect: they do not necessarily identify with a group, nor require common knowledge of team reasoning. They simply apply a counterfactual universalizability test to their own strategy. In this sense, team reasoning and Kant reasoning both depart from individualistic best-response logic, but via different routes. Thus, Kant reasoning need not rely on team identification or group salience; it suffices for agents to apply a universalizability operator to their own strategy, regardless of group boundaries.

B.6 Self-Enforcement

A perennial objection to models of ethical reasoning is whether principled behavior can persist amid opportunistic defection. The standard answer, self-confirming equilibrium (SCE), requires that each agent's strategy is optimal given beliefs that are correct about on-path play.

Definition 9 (Self-Confirming Equilibrium (SCE))**.** *A self-confirming equilibrium (SCE) is an outcome in which every agent's strategy is optimal, given their (possibly incorrect) beliefs about the play of others—provided these beliefs are correct along the path of play actually realized. Mathematically, a profile Φ^* is an SCE if, for each agent i,*

1. *ϕ_i^* solves $\max_{\phi_i} U_i(\phi_i; \mathbb{P}_i)$, where $\mathbb{P}_i$ is agent i's belief over others' play;*
2. *$\mathbb{P}_i(\phi_j = \phi_j^*) = 1$ for all j whose actions are observed by i on path.*

SCE is weaker than Nash equilibrium, as it does not require correct beliefs about unobserved, off-path play. In many environments, SCE is the natural limit of adaptive learning when agents observe only realized outcomes.

However, SCE is ill-suited for Kant equilibrium (or Kant-Nash hybrid equilibrium), because universalizability is a structural rule—agents' beliefs are fixed by ethical principle, not revised in light of empirical evidence. Kant logic is immune to belief updating; the equilibrium is a fixed point of rules, not of learning dynamics.

This necessitates the development of an appropriate analog of self-confirmation for the current context, so that we can assess the local stability of Kant equilibria without requiring agents to hold empirically correct beliefs about others' strategies. I propose self-enforceability at the level of decision rules as a suitable alternative to self-confirmation of beliefs.

Definition 10 (Equilibrium with Self-Enforcing Decision Rules)**.** *An equilibrium is self-enforcing at the level of decision rules if, for each agent i, given her adopted decision rule (Kant, Nash, or hybrid), no agent has an incentive to deviate in action, holding her rule fixed.*

This definition captures the essence of Kant equilibrium: agents act according to their adopted decision rule, and no agent can profitably deviate in action while holding her rule fixed.

Yet this criterion remains shallow unless we interrogate rule-level mutations, not just action deviations. In the current model, deviation from Kant reasoning is not a mere change in portfolio choice, but a change in *type*—the agent becomes Nash. The deeper stability question is evolutionary: can ethical rules persist under selection, mutation, or social learning?

The analysis and predictions of the model are thus conditional on the exogenous existence of rule heterogeneity. Explaining the endogenous stability or emergence of Kant logic is a distinct, and richer, research question, and a beginning is attempted in Section 5.

B.7 Off-Equilibrium Plausibility and Refinements

In the context of Kant equilibrium, off-equilibrium plausibility refers to the behavior of agents when they deviate from the Kant strategy profile. Specifically, we are interested in whether small deviations from the Kant equilibrium lead to a return to the original equilibrium or trigger further deviations. Are small deviations punished or rewarded? Do they lead to unraveling of the Kant equilibrium?

Classical refinements such as off-equilibrium plausibility (OEP) rely on the credibility of beliefs following off-path play, ensuring that equilibria are not supported by implausible or non-credible beliefs about how agents would respond to unexpected deviations. In Kant equilibrium, the logic of universalizability fixes agents' reasoning ex ante and precludes belief-updating in response to deviations; as such, the classical OEP condition is vacuously satisfied or inapplicable.

Small deviations from Kant equilibrium are not punished in the way that subgame-perfect or belief-based equilibria would prescribe, because the Kant agent's decision rule is fixed by principle, not by expectation of punishment or reward. If an agent "deviates" from Kant reasoning, she becomes a Nash type by definition—the deviation is not marginal, but a switch in behavioral type. There is no endogenous discipline for small deviations within the Kant type—either one is Kant, or not. The only stability test is at the type level, not at the margin.

Nevertheless, a Kant-friendly analog to OEP would seek to ensure that equilibrium is robust not only to type-level mutations but also to small, perhaps inadvertent perturbations to the universalizability mandate. For example, such a refinement would ask:

- Are Kant strategies robust to unexpected deviations—not just type-level mutations, but possible mistakes or partial lapses in universalizability?
- Do agents punish or reward deviations in a way that would support the stability of Kant logic?
- Is there a social or institutional structure that restores Kant play if small numbers deviate?

Defining such a refinement would require specifying how the agent's decision rule—and possibly the social or institutional context—responds to minor deviations: does the system restore universalizability-based decision-making, or does deviation cascade into unraveling?

Developing a general theory of universalizability-consistent plausibility is a promising direction for future research. Here, I note that the current model is robust only to action-level and type-level deviations as modeled, and more sophisticated robustness criteria remain an open problem.

C Empirical Appendix

C.1 Empirical Markers of Investor Types

Large institutional investors vary widely in their approach to proxy voting on high-profile Environmental and Social shareholder proposals. Below, I summarize representative cases of all investor types considered in this paper. Tables 14 and 15 summarize the core identifying features of utilitarian and deontological investors used in this paper to provide empirical evidence, alongside behaviors that appear similar but fall short of the classification.

Screening Procedure. To make the classification procedure transparent and readily replicable, the first pass uses a rule-based, case-insensitive regular-expression (regex) screen over fund disclosures (prospectuses, SAI/policies, stewardship reports). The scanner operates on sliding text windows and records hits when any pattern in a category is matched within the window. Each category has a minimum-hit threshold and a match aggregation rule. Documents that clear a threshold are flagged for manual validation; funds that do not clear any threshold default to the neutral type unless subsequent review provides contrary evidence. The screen focuses on exclusionary and engagement language, with auxiliary categories (voting mechanics; collective action) used to refine or contextualize classifications. Flagged documents are reviewed to confirm that (i) exclusionary language reflects normative screening (not purely financial risk screens), and (ii) engagement relates to environmental/social externalities (not only governance mechanics). Representative (non-exhaustive) keywords used in the scan are listed below for illustration. Full regex expressions, including pluralization and common variants, are retained in the replication code.

Exclusion (normative screens). abortion; adult entertainment; alcohol / liquor; tobacco / cigarettes; gambling / casino / sports betting; weapons / armaments / “controversial weapons”; fossil fuels (coal, oil, natural gas); Sharia / Shari’ah; faith-based / religious; “controversial sectors.”

Engagement (E/S objectives). engagement strategy / policy / plan; engage with companies / issuers / portfolio companies; corporate engagement; engagement campaign; active engagement; investor–company dialogue; report engagement outcomes.

Voting (auxiliary). proxy voting policy / guidelines; voting framework; vote for / against shareholder proposals; vote with / against management; voting decisions.

Collective action (auxiliary). Principles for Responsible Investment (PRI); Net Zero Asset Managers (NZAM); Net Zero Asset Owners (NZAO); Climate Action 100+ (CA100+); coordinated stewardship / engagement; public campaigns.

Funds flagged as *exclusion* (normative prohibitions) are provisionally deontological; funds flagged as *engagement* (explicit environmental or social influence objectives) are provisionally utilitarian. Voting and collective-action hits are informative but not determinative. In cases with mixed signals, the manual check assigns the *dominant orientation* (see examples below). Neutral is the residual for funds with neither exclusions nor E/S engagement aims.

Table 14: Identifying Utilitarian Investors Empirically

Identifying Behavior of a Utilitarian	Related but Inadequate Behavior for Utilitarian Classification
States commitment to engage with firms on environmental or social goals	Makes broad ESG commitments without reference to engagement or firm influence
Describes policy or history of direct engagement on E/S issues (e.g., climate risk, supply chain labor standards)	Describes engagement practices focused exclusively on corporate governance or executive compensation
Participates in collaborative engagements focused on E/S issues (e.g., CA100+, ICCR campaigns)	Participates in collaborative forums without specifying engagement topics or outcomes
Signals willingness to escalate (e.g., vote against directors, file shareholder proposals) in pursuit of environmental or social objectives	Mentions escalation or voting behavior solely around governance mechanics
Discloses stewardship actions tied to E/S themes (e.g., targeted engagement outcomes, priority sectors)	Publishes generic ESG or stewardship reports without actionables or topical specificity
Indicates intent to shape real-world firm behavior through voting, engagement, or public pressure	Frames ESG engagement as a tool for understanding firms, not influencing them

Consider the following excerpts from fund prospectuses and stewardship reports, illustrating the classification criteria.

Deontological funds. Funds apply categorical exclusions, such as bans on tobacco, fossil fuels, or controversial weapons. Typical exclusionary language includes:

> ...the fund's criteria do not permit investment in companies whose businesses rely significantly on alcohol, firearms, gambling, tobacco, nuclear power, or production of weaponry. Furthermore, companies deemed to have poor labor/employee relations or environmental records are also screened out of the investable universe...

> ...the fund will critically evaluate companies that significantly support governments that are under U.S. or international sanction for grave human rights abuses such as genocide...

> ...Sharia investing style utilizes negative screening against companies which are engaged in socially injurious activities like alcohol, gambling, interest or weapons. The Sharia [fund] also filters out the companies with higher leverage...

Utilitarian funds. Funds state an explicit objective to influence portfolio firms through engagement, voting, or collaborative action, with the aim of improving aggregate outcomes such as exter-

Table 15: Identifying Deontological Investors Empirically

Identifying Behavior of a Deontologist	Related but Inadequate Behavior for Deontological Classification
Applies normative exclusions based on sector, behavior, or legal/moral violations (e.g., tobacco, controversial weapons, child labor)	Conducts risk-based exclusions based on ESG controversies likely to affect firm value (i.e., financially motivated screening)
Implements static exclusion lists or policies tied to ethical codes, religious values, or treaties (e.g., UN Global Compact, Sharia law)	Excludes firms only after ESG incidents if there's clear reputational/financial damage—indicative of risk aversion, not deontology
Refers to ethical principles, fiduciary duties to clients' values, or moral beliefs in justifying portfolio constraints	Uses ESG terminology vaguely, e.g., "responsible investing," without referencing duties, ethics, or categorical norms
Avoids exposure to morally tainted sectors or jurisdictions, even when financial performance is not at risk	Divests only from firms after underperformance or due to regulatory pressure
Frames ESG in rights-based, duty-based, or ethical language (e.g., "violation of human dignity," "right to life")	Uses ESG as a label for branding or generic long-term risk mitigation
Refuses to invest in firms regardless of their progress or improvement on ESG issues (i.e., rejects best-in-class if in a "bad" sector)	Selects "ESG leaders" even in controversial sectors, indicating willingness to trade off moral concerns for relative improvement
Maintains rigid ex-ante ethical screens, with little to no variation over time or across jurisdictions	Adjusts ESG implementation based on local standards, evolving norms, or materiality thresholds—suggestive of instrumentalism

nality reduction. Typical engagement-oriented language includes:

> *...[the fund] seeks to identify opportunities for a company to improve its ESG practices, and will endeavor to work collaboratively with company management to establish concrete objectives and to develop a plan for meeting these objectives...*

> *...we engage proactively with issuers to encourage them to improve their ESG practices...[activities] include, but are not limited to, direct dialogue with company management, such as through in-person meetings, phone calls, electronic communications, and letters...*

> *...our current and ongoing [engagement] activities can be viewed through regular publication of case studies and thematic papers...*

Neutral funds. Funds do not disclose any exclusionary criteria or engagement objectives. These funds typically emphasize financial returns without reference to ethical considerations.

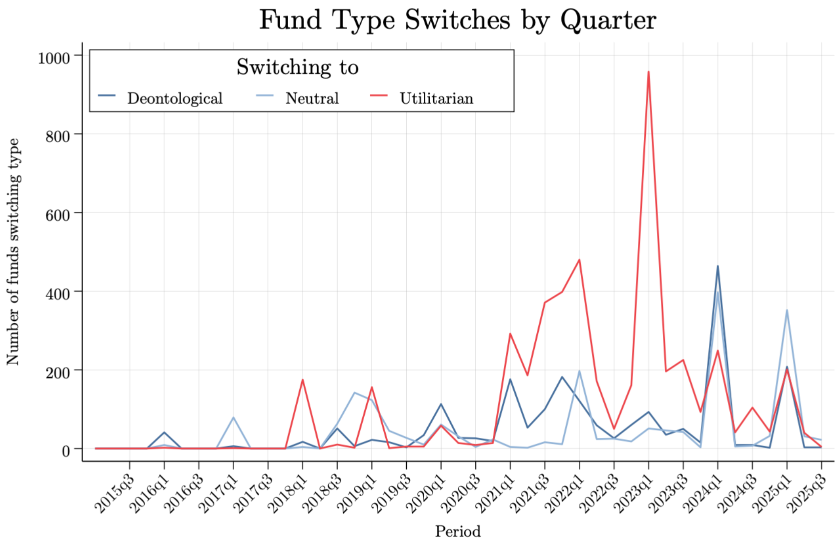
In practice, some funds disclose both exclusions and engagement, or use ESG terminology without a clear ethical logic. These are classified according to manually determined *dominant orientation*. For example, the following excerpt illustrates a fund that applies exclusions but also emphasizes engagement, and is classified as utilitarian based on overall emphasis on engagement rather than exclusion:

> *...we are active shareholders...we look for ways to engage with companies. Experience has taught us that most companies do not set out to do harm and, in most cases, they are not aware. By engaging with them, we provide them with information about any conflicting activity as well as suggest a course of action to remedy it. This is a balancing act with screening...*

While annually filed documents allow for a time-varying classification, for the Climate Action 100+ analysis used in this paper, I use a static classification based on filings up until 2017, and keep types fixed over the 2018–2025 period. This is to avoid endogeneity concerns where funds might switch types in response to observed engagement outcomes.

Separately, to assess classification stability, I track type switches over time using annual disclosures from 2015 to 2025. Fig. 10 shows the number of funds changing moral type classification each year. The data reveal that type switches spiked between 2021 and 2023, coinciding with regulatory turbulence around the EU Sustainable Finance Disclosure Regulation (SFDR) and related re-labelling of funds. These switches do not appear to reflect fundamental changes in investment philosophy, but rather administrative reclassifications in response to evolving regulatory definitions. These switches do not affect my empirical work, and are documented here for completeness.

Figure 10: Number of funds switching moral type classification over time, 2015–2025. The spike in switches around 2021–2023 coincides with regulatory turbulence around SFDR and related relabelling.



C.2 Identifying Nash and Kant Voting Patterns

Kant Style Transparency: Funds Preannouncing Votes

AllianzGI. AllianzGI has publicly pre-declared support for specific ESG shareholder proposals. For example, prior to the 2025 Alphabet AGM, AllianzGI issued a press release stating its intention to vote *in favor* of three social-oriented shareholder resolutions covering an AI human-rights impact assessment, lobbying alignment with child safety, and online child safety (AllianzGI, 2025c; Bybel, 2025). Similarly, AllianzGI pre-declared support for governance proposals at Meta in 2025, explicitly noting the decisions reflect the high importance they place on social issues materially affecting the business (AllianzGI, 2025a, 2025b).

Norges Bank Investment Management (NBIM). NBIM often reveals its voting stance on key climate resolutions ahead of time. For instance, NBIM publicly announced it would vote against a 2023 climate resolution at BP's AGM five days before the vote (Adomaitis, 2023). Likewise, in 2024, NBIM disclosed it would not support a climate resolution at Shell's AGM, stating Shell's updated strategy was already 'sufficiently' Paris-aligned (Solsvik, 2024).

Candriam. Candriam explicitly publishes select 'Predeclaration of Voting Intentions' on its website for upcoming shareholder meetings and resolutions, typically highlighting high-impact ESG items. These disclosures, though selective, indicate a willingness to signal principles and leadership (Candriam, 2025).

CalPERS. CalPERS has, on occasion, pre-announced votes to send a message. In May 2024, CalPERS declared it would vote against the entire board of ExxonMobil at the upcoming AGM, citing Exxon's hostility toward climate shareholder proposals as the rationale for this public rebuke (Roy & Binnie, 2024).

Funds that preannounce votes often cite stewardship benefits: by being transparent and principled, they aim to lead by example and influence both companies and fellow shareholders. There is a collective action angle—if large investors all took a firm public stance, management would feel greater pressure to heed shareholder concerns. Pre-declaring can also enhance an investor's reputation for integrity and satisfy beneficiaries who expect alignment with stated ESG values.

Nash Style Caution: Funds Keeping Votes Secret

Invesco. Invesco's proxy voting policy states the firm 'does not publicly disclose voting intentions in advance of shareholder meetings' (Invesco, 2025). The rationale is operational and legal: pre-declaring could invite lobbying or regulatory scrutiny. This preference to vote in private is typical among traditional asset managers.

Vanguard. Vanguard rarely announces specific votes beforehand. For example, in the high-profile 2023 Disney proxy contest, Vanguard declined to comment on its voting intentions, only revealing its position at the meeting itself (Goswami, 2024). In 2024, Vanguard funds supported zero E&S shareholder proposals by outsiders, with all voting decisions disclosed only after the fact (Blanding, 2017; Bloomberg, 2024).

BlackRock. BlackRock typically keeps silent before key votes. Even for the historic 2017 ExxonMobil climate-risk disclosure resolution—which passed with unexpected support from BlackRock and others—the firm's intentions were not made public prior to the AGM (Benny, 2017). BlackRock issues post-vote bulletins but avoids pre-announcing. Despite its CEO's public letters urging companies to disclose climate risks (Bieber & Klingsberg, 2016), the firm's support for ESG proposals has been on a declining trend (L. Johnson, 2024), although some argue that proposal quality might explain the trend better than a declining ESG mandate (McGowan, 2024).

Other U.S. Giants. The largest U.S. asset managers (BlackRock, Vanguard, State Street, Fidelity, JPMorgan) are generally tight-lipped, have low support rates for E&S proposals, and do not pre-announce intentions (White, 2024).

Funds cite several reasons for keeping votes secret until cast: fiduciary flexibility (allowing new information or negotiation with management), avoiding external pressure, and minimizing business or regulatory risks. From a game-theoretic perspective, this "vote on the day" approach is a classic Nash equilibrium—each fund waits to see the lay of the land, avoiding unilateral commitments and collective action traps. To be sure, reticence does not imply absence of principle; however, if an overarching normative commitment to sustainability existed, voting patterns would be predictable ex-ante, and the choice of when to disclose votes would be less consequential.

Dedicated sustainable funds tend to support a high proportion of E&S resolutions and are more willing to pre-declare their votes. For example, Amundi reports voting in favor of 88% of climate-related and 83% of social/human rights-related resolutions in 2023 (Amundi, 2024). Calvert likewise consistently backs sustainability proposals, continuing to support climate and diversity measures while mainstream managers cut back support (Wilson, 2025). In contrast, generalist mega-funds are more reticent, with some like Vanguard supporting no Environmental or Social shareholder proposals in 2024 (Sustainalytics, 2024). Such differences underscore the spectrum of behaviors, from funds that publicly exhibit ethical leadership in their voting to those that behave as self-interested players in a game.

These disparate approaches underscore the strange dynamics of asset management when it comes to shareholder activism on sustainability. While data is imperfect, even anecdotal evidence demonstrates that some funds willingly break industry norms to exhibit leadership, while others hang back, preserving maximum optionality and minimizing risk. Not all investors are the same: some publicly champion climate or social causes, while others remain silent—a divergence with real-world consequences for the success of ESG activism. To empirically classify funds as Kant or Nash, such nuances must be considered. For instance, a fund might vote consistently on certain issue-types while acting strategically on others. Rigid voting policies are therefore not sufficient for identifying Kantian strategies. Instead, evidence of intention-invariance—where the vote reflects what the fund would do regardless of its effect, and regardless of others' choices, may serve as a more reliable indicator. Table 16 provides a summary of criteria that may be used to identify an investor's strategic type.

C.3 Stacked Difference-in-Differences Event Study

Limitations of fixed effect models with staggered timing. With staggered treatment and heterogeneous effects, two-way fixed effects (TWFE) DiD can be contaminated by *treated-as-control* comparisons and non-convex weighting of cohort-time effects, leading to attenuation or sign reversals in dynamic event studies. Because CA100+ targets enter in waves and responses plausibly evolve by cohort and over time, a design that respects cohort timing and avoids treated controls is necessary to validate the findings. I estimate a stacked difference-in-differences event study following Callaway and Sant'Anna (2021) to compute an alternative non-parametric measure of the dynamic response to the shock by fund type.

Objects of interest, comparison group, and notation. Let i index funds, j firms, and t quarters. Let g_j be the first quarter in which firm j is designated a target (the *cohort*), and define the (absorbing) treatment status $D_{jt} = \mathbf{1}\{t \geq g_j\}$. The outcome is the percentage of firm j's equity owned by fund i at t, w_{ijt}. For a fund-type group $H \in \{\text{Util}, \text{Deon}, \text{Neut}\}$, the group-time average treatment effect is

$$\text{ATT}_H(g,t) \ = \ \mathbb{E}[w_{ijt}(1) - w_{ijt}(0) \mid g_j = g, \ i \in H, \ t] \,, \tag{65}$$

estimated by comparing treated dyads to an explicitly constructed *never-treated* control group at time t. We aggregate to event time $e = t - g$ via transparent cohort-size weights:

$$\text{ATT}_H(e) \ = \ \sum_g \omega_{H,g,e} \, \text{ATT}_H(g, g{+}e), \qquad e \in \mathcal{E}. \tag{66}$$

The event study stacks cohorts and computes $\text{ATT}_H(g,t)$ by contrasting the mean change in outcomes for the treated cohort (g) with the mean change for controls in the sample. This avoids treated-as-control contamination entirely and is particularly suitable here because CA100+ designations are absorbing treatments (once treated, the firm remains treated). Specifically, the fund-firm dyad (i,j) enters the risk set if fund i holds firm j pre-event; for each (g,t), treated dyads are those where j belongs to cohort g and $t \geq g$, and controls are dyads linked to firms never designated as targets.

Table 16: Identifying Strategic Investor Types Empirically

Dimension	Kant	Nash
Decision Principle	Votes as if their action were to be universalized; guided by principle	Votes to maximize expected influence; guided by outcome likelihood
Responsiveness to Others	Invariant to anticipated behavior of others	Strategically responsive to anticipated support or opposition
Voting Consistency	High consistency across similar proposals, justified by duty or norm	Variability across contexts; votes adjusted based on firm, timing, or campaign
Pre-declaration of Vote	May publicly declare voting intentions ahead of vote	Typically avoids pre-declaring; retains flexibility to respond to new information
Rationale Language	Emphasizes duties, fairness, universal standards (e.g., "we believe all companies should. . . ")	Emphasizes effectiveness, influence, or conditionality (e.g., "where appropriate," "on a case-by-case basis")
Treatment of Low-Pass Likelihood Proposals	May vote in favor regardless of low success probability	More likely to abstain or vote against if unlikely to pass
Internal Policy Design	Adopts rules that would be defensible regardless of influence	Adopts flexible policies tailored to context and outcomes
Nature of Collaboration and Responsiveness	Participates in collaborative initiatives or consensus-building efforts *without making action conditional on others' behavior*; articulates intent to act based on principle or universal norms, regardless of peer support (e.g., pre-declares votes, supports engagement even without majority backing)	Engages in collaboration or coordination *explicitly conditional* on others' participation or expected support; actions are framed as contingent upon majority alignment, investor consensus, or likelihood of success (e.g., "will act if sufficient investor support is secured")

Unlike TWFE, this method does not rely on high-dimensional fixed effects for identification. Parallel trends are defined *for each cohort* relative to its chosen controls (here, never-treated), and pre-trend diagnostics are meaningful at that level. While fixed effects may improve precision in auxiliary OLS representations, the core estimator is based on cohort-specific differences and weighting, not on a single global regression with α_{it} or κ_{jt}. Standard errors are clustered at the fund-firm dyad level, consistent with the sampling structure.

Identifying assumptions and diagnostics. This model with never-treated controls requires: (i) *cohort-specific parallel trends* for treated dyads relative to never-treated dyads in the absence of treatment; (ii) *no (or limited) anticipation*—handled by trimming early leads; (iii) *overlap*—sufficient support of never-treated comparators for each cohort-time cell; and (iv) *SUTVA/no interference* across untreated firms. The method avoids treated-as-control contamination; yields cohort-clean dynamics $\mathrm{ATT}_H(g,t)$ and transparent event-time aggregation; enables pre-trends to be interpretable; and does not rely on TWFE saturation. However, spillovers from coordination to non-target peers may bias estimates toward zero.