Towards Sustainable Investment Policies Informed by Opponent Shaping

Juan Agustin Duque*, Razvan Ciuca*, Ayoub Echchahed, Hugo Larochelle, Aaron Courville

firstname.lastname@umontreal.ca

University of Montreal & Mila

Abstract

Addressing climate change requires global coordination, yet rational economic actors often prioritize immediate gains over collective welfare, resulting in social dilemmas. InvestESG is a recently proposed multi-agent simulation that captures the dynamic interplay between investors and companies under climate risk. We provide a formal characterization of the conditions under which InvestESG exhibits an intertemporal social dilemma, deriving theoretical thresholds at which individual incentives diverge from collective welfare. Building on this, we apply Advantage Alignment, a scalable opponent shaping algorithm shown to be effective in general-sum games, to influence agent learning in InvestESG. We offer theoretical insights into why Advantage Alignment systematically favors socially beneficial equilibria by biasing learning dynamics toward cooperative outcomes. Our results demonstrate that strategically shaping the learning processes of economic agents can result in better outcomes that could inform policy mechanisms to better align market incentives with long-term sustainability goals.

1 Introduction

Climate change is one of the defining challenges of our time, which demands immediate, coordinated global action to mitigate catastrophic environmental and economic consequences (Pörtner et al., 2022). Despite international agreements such as the Paris Agreement, nations continually grapple with aligning individual economic incentives with collective environmental goals. This misalignment can be effectively modeled as a *social dilemma*: a scenario in which players, acting independently and rationally according to their own interest, produce a suboptimal collective outcome (Rapoport and Chammah, 1965). In climate negotiations, nations often prefer short-term economic benefits over the collective welfare of aggressive mitigation strategies, leading to Pareto-suboptimal equilibria.

Previous advances in reinforcement learning (RL) have demonstrated remarkable capabilities in optimizing agent behaviors in various complex tasks (Mnih et al., 2013; Schulman et al., 2017). However, when confronted with general-sum games such as climate change negotiations, traditional RL methods consistently converge to selfish strategies that maximize immediate individual rewards but result in globally suboptimal outcomes (Sandholm and Crites, 1996; Foerster et al., 2018). These greedy equilibria do not realize the potential gains achievable through cooperation, underscoring the critical need for methods that inherently incentivize collaborative behavior among rational players.

To address this fundamental limitation, *opponent shaping* (Foerster et al., 2018) has emerged as a promising approach, allowing players to strategically influence the learning dynamics of other participants. Advantage Alignment, a recent advancement within opponent shaping algorithms (Duque et al., 2025), operates by aligning the *advantages* of interacting players. Unlike earlier opponent

shaping methods that rely on computationally expensive imagined parameter updates (Foerster et al., 2018; Willi et al., 2022), Advantage Alignment modifies policy gradients directly, promoting actions that benefit both individual and collective outcomes efficiently.

Conventional MARL methods (IPPO (Lu et al., 2024)) optimize myopic returns, while recent opponent-shaping approaches (LOLA (Foerster et al., 2018), M-FOS (Lu et al., 2022), BRS Aghajohari et al. (2024a)) scale poorly beyond simple games or are limited to discrete action spaces (LOQA (Aghajohari et al., 2024b)). In this paper, we apply Advantage Alignment in the context of sustainable investment policies in the InvestESG environment (Hou et al., 2025), a realistic simulation of corporate and investor interactions under climate risk. Specifically, our contributions are as follows.

- We formally prove for a simplified InvestESG that the parameter α (the climate responsiveness to mitigation action) is critical for making the game a social dilemma, and empirically show that the full game becomes a dilemma with $\alpha = 70$.
- We apply Advantage Alignment to handle the complex dynamics characteristic of the InvestESG simulation, demonstrating its scalability and effectiveness.
- We provide a theoretical argument for why Advantage Alignment finds better equilibria than other baselines (including PPO with incentives and PPO summing the rewards).
- We provide insights into the policies discovered by Advantage Alignment, which contribute to mitigation more effectively than other algorithms.

2 Background

2.1 Markov Games

We consider the formalism of n-player, fully observable, general-sum Markov Games (Shapley, 1953) which are represented by a tuple: $\mathcal{M} = (N, \mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$. Here, the state space is denoted $\mathcal{S}$. The joint action space for all players in the game is written as $\mathcal{A} := \mathcal{A}^1 \times \ldots \times \mathcal{A}^n$. The transition function, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, maps from every state and joint action to a distribution over states. The reward structure is given by $\mathcal{R} = \{r^1, \ldots, r^n\}$, where each $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the reward received by player i. The discount factor, $\gamma \in [0, 1]$, determines the relative importance of future rewards.

2.2 Social Dilemmas

Social dilemmas, as introduced by Rapoport and Chammah (1965), describe multi-player scenarios in which individual rational play leads to an outcome with less utility than that which would have occurred if all players had cooperated. Formally, consider a normal-form game with n players, where each player i selects an action a^i from their action space $\mathcal{A}^i$, and receives utility $r^i(a^1, \ldots, a^n)$. Let $\mathbf{a}_\text{C} = (a^1_\text{C}, \ldots, a^n_\text{C})$ denote a joint cooperative strategy profile that maximizes the total social welfare.

Definition 1 (Nash Equilibrium (Nash, 1950))**.** *Consider an n-player game where each player $i \in \{1, \ldots, n\}$ has an action space $\mathcal{A}^i$ and a payoff function $r^i : \mathcal{A}^1 \times \cdots \times \mathcal{A}^n \to \mathbb{R}$. A strategy profile $\mathbf{a}_N = (a^1_N, \ldots, a^n_N)$ is a Nash equilibrium if for all players i:*

$$r^i(a^i_N, \mathbf{a}^{-i}_N) \geq r^i(a^i, \mathbf{a}^{-i}_N), \quad \forall a^i \in \mathcal{A}^i.$$

Here, $\mathbf{a}^{-i}_N$ denotes the strategies chosen by all players other than player i.

Definition 2 (Social Dilemma)**.** *Let $\mathbf{a}_N$ be a Nash equilibrium of the game. A game is said to be a social dilemma if it satisfies both of the following conditions:*

1. ***Collective Inefficiency:*** *The socially optimal outcome yields strictly higher total utility than the Nash equilibrium:*

$$\sum_{i=1}^{n} r^i(\mathbf{a}_C) > \sum_{i=1}^{n} r^i(\mathbf{a}_N).$$

2. ***Individual Incentive to Defect:*** *For each player i, there exists some joint strategy $\mathbf{a}^{-i}$ such that deviating from the cooperative strategy is individually beneficial:*

$$r^i(a_D^i, \mathbf{a}_C^{-i}) > r^i(a_C^i, \mathbf{a}_C^{-i}),$$

where a_D^i is a defection action and $\mathbf{a}_C^{-i}$ denotes the cooperative actions of all other players.

These two conditions capture the tension at the heart of social dilemmas: cooperation is globally optimal, but locally unstable due to unilateral incentives to defect (Capraro and Halpern, 2015). However, this definition makes strong assumptions that may not hold in the real world (e.g. uniqueness of Nash). For analyzing social dilemmas in Markov games, we define the *price of anarchy* (Nisan et al., 2007).

Definition 3 (Price of Anarchy (Markov Games))**.** *Let $\mathcal{M}$ be an n-player Markov game. Denote by Π the set of joint policies and by $\mathcal{N} \subseteq \Pi$ the subset that are Nash (Definition 1). For a joint policy $\pi \in \Pi$ let $\mathcal{W}(\pi;\mu) := \sum_i V_\pi^i(\mu)$ be the social welfare function. The price of anarchy is*

$$\mathcal{P}_a = \frac{\max_{\pi\in\Pi} \mathcal{W}(\pi;\mu)}{\min_{\pi\in\mathcal{N}} \mathcal{W}(\pi;\mu)}.$$

A Markov game exhibits a social dilemma whenever $\mathcal{P}_a > 1$, i.e. the worst equilibrium achieves strictly lower total return than some feasible, $\mathcal{W}$-maximizing, joint policy. Intuitively, the price of anarchy is the ratio between the welfare of coordinated and individual rational play.

2.3 Opponent Shaping

We assume familiarity with reinforcement learning; the necessary background is detailed in Appendix A. Opponent shaping, as introduced by Foerster et al. Foerster et al. (2018), changes the paradigm of reinforcement learning: instead of assuming stationarity of the environment, opponent shaping algorithms assume that other players are learning agents. By making assumptions about the learning algorithms (Foerster et al., 2018) or modeling the problem as a meta-game (Lu et al., 2022), opponent shaping algorithms attempt to influence the learning dynamics of other players. Duque et al. Duque et al. (2025), proved that controlling other players via the Q-values is equivalent to doing policy gradient with the following modified advantage:

$$A^{*,i}(s_t, \mathbf{a}_t) = \left(A^i(s_t, \mathbf{a}_t) + \beta\gamma \cdot \sum_{j\neq i} \left(\sum_{k<t} \gamma^{t-k} A^i(s_k, \mathbf{a}_k) \right) A^j(s_t, \mathbf{a}_t) \right).$$

This Advantage Alignment formula can be plugged into the PPO formulation for a scalable Opponent Shaping algorithm (Proximal Advantage Alignment) that works in complicated multi-agent settings with high dimensional state representations, partial observability and multiple (more than 2) agents.

3 InvestESG

InvestESG (Hou et al., 2025) is a multi-agent reinforcement learning (MARL) environment designed to study the long-term impact of corporate climate investments and ESG (Environmental, Social, and Governance) disclosure mandates in an intertemporal economy under climate risk. The environment simulates interactions between company agents (who allocate capital across mitigation, resilience, and greenwashing strategies) and investor agents (who reallocate capital across companies based on profitability and ESG scores). Climate risk evolves over a 100-year horizon, shaped by cumulative mitigation efforts, and materializes through climate events that inflict economic losses on companies. By default, InvestESG uses 5 companies, 3 investors and doesn't allow for resilience

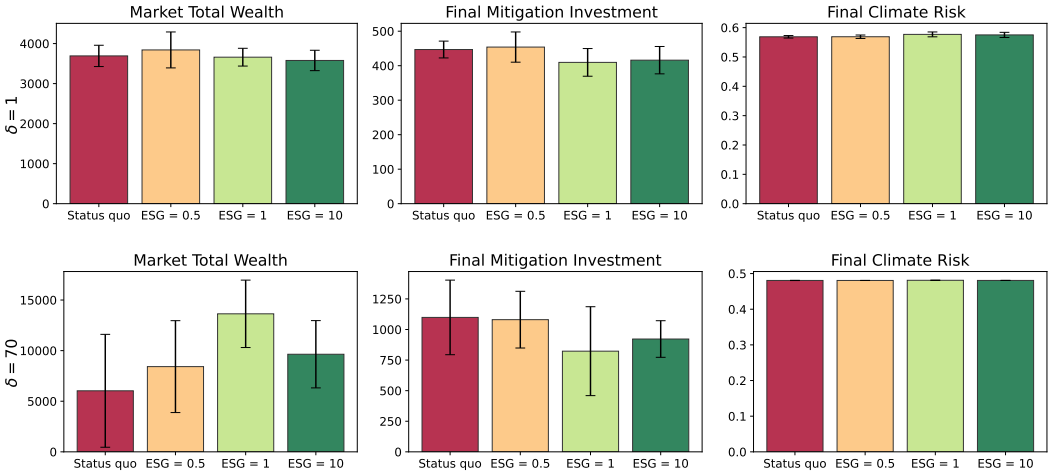


Figure 1: Comparison of final environment metrics at different α (introduced in equation B) values, for 10 seeds of PPO agents: With the default scaling ($\alpha = 1$), the differences in market total wealth and final mitigation amount are negligible between agents with different ESG incentives. *Status Quo* refers to PPO agents trained without incentives. Increasing the scaling factor ($\alpha = 70$) results in higher market total wealth for ESG-conscious investors, and higher price of anarchy. The whiskers indicate a 1-standard deviation confidence interval. It should be noted that the ESG is a multiplicative parameter, ESG $= 10$ represents an immense prosocial reward: the investors in that scenario effectively do not care at all about their own profits.

or greenwashing strategies, although these parameters can be modified. The main limitations of this simulator, as a result of this parameterization, are clearly stated in the InvestESG paper (Hou et al., 2025). We detail the dynamics of the simulator in the Appendix B.

4 A Formal Analysis of InvestESG as a Social Dilemma

The InvestESG environment was originally introduced as a scenario capturing essential features of social dilemmas in sustainable investing. Under its default calibration, however, private and collective pay-offs remain almost perfectly aligned, so the game fails to reproduce the free-rider tension that defines a genuine social dilemma: the empirical price of anarchy (Definition 3) is close to 1 (Figure 1). The observed equilibrium does not change even when adding extremely strong ESG incentives such that investors effectively only care about ESG (the ESG=10 case in the figure). In a social dilemma, we would expect to see that adding prosocial incentives changes the final behavior of agents, yet this is not what we see empirically. Here, we derive conditions under which InvestESG explicitly meets the definition of a social dilemma, clarifying both theoretically and empirically when this environment manifests the characteristic tensions between individual rationality and collective welfare. We find that the effectiveness of mitigation actions λ is the critical parameter which makes InvestESG a social dilemma. Identifying this regime is essential, because it mirrors the real-world climate-mitigation problem in which near-term incentives myopically outweigh the long-term benefits of coordinated action for self-interested agents (Barrett, 1994). We look at three metrics: (1) **Market Total Wealth** is the combined value of all companies' capital and investors' cash at the end of the simulation; (2) **Final Mitigation Investment** is the total amount all companies have cumulatively spent on mitigation by that point; and (3) **Final Climate Risk** is the model-estimated probability that at least one damaging climate event occurs at the final timestep.

4.1 Set-up and Notation

Consider the amount of mitigation that company i invests at time t, denoted u_t^i. Intuitively, a company cares about the marginal effect, or *benefit*, that this mitigation amount will have at the next time-step on their capital, K_{t+1}^i. Since K_{t+1}^i is a random variable that depends on the realized num-

ber of climate events X_t, we take the marginal effect to be the partial derivative of the mitigation amount w.r.t. the expected value of the capital at the next time-step. See Appendix C for details.

Definition 4 (Private Marginal Gradient)**.** *Let u_t^i be the mitigation amount of company i at time t, then the private marginal gradient is*

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} + \mathbb{E}\left[\frac{(K_{t+1}^i)^2}{(1-X_tL_i)^2(1-u_t^i)(1+\gamma)}\sum_e \frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right]$$

namely, the marginal effect in the capital of company i by investing in mitigation at time-step t.

This gradient indicates whether a company has a selfish, myopic incentive to invest in mitigation: if the gradient is positive, then the capital at the next time-step increases for each dollar invested mitigating. Similarly, we define the social marginal gradient.

Definition 5 (Social Marginal Gradient)**.** *Let u_t^i be the mitigation amount of company i at time t, then the social marginal gradient is*

$$\frac{d}{du_t^i}\mathbb{E}\left[\sum_k K_{t+1}^k\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} + \sum_k \mathbb{E}\left[\frac{(K_{t+1}^k)^2}{(1-X_tL_k)^2(1-u_t^k)(1+\gamma)}\sum_e \frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right]$$

namely, the marginal increase in the capital of all companies by company i investing in mitigation at the next time-step.

Our argument builds from the intuition that a social dilemma arises if there is a fundamental tension between these two gradients: the social marginal gradient pushes companies to mitigate, whereas the private gradient does not. The final proof, however, is more involved, as we analyze the effect of *all* previous mitigation actions on the final private and social capitals.

4.2 When is InvestESG a Social Dilemma?

The full stochastic dynamics of InvestESG are too complicated for exact analytic analysis so we simplify the environment. To isolate the *social-dilemma* component of climate-mitigation decisions, we work under the following assumptions for the remainder of this section.

Assumption 1. (Static Investor Portfolios). *Investor holdings are time–invariant. Formally, $H_{t,i}^j = H_{0,i}^j \; \forall\, t \geq 0,\; i \in [M],\; j \in [N]$. Thus investors never rebalance their capital allocations.*

Assumption 2. (Single-Company Allocation). *Each investor supplies capital to exactly one firm.*

Assumption 3. (No Historical Mitigation). *Prior to the period of interest, firms have undertaken no mitigation expenditure: $u_s^i = 0 \; \forall\, s < t,\; i \in [M]$.*

We are interested in computing the incentives on companies to start mitigating climate change in a world where no one has ever mitigated before and where the investor dynamics are simplified. In this simplified world, we show that the existence of a social dilemma critically depends on the parameter λ_e of the simulation, which is the responsiveness of climate change to mitigation investment. We give here an outline of the argument, the full derivation is in Appendix C:

1. There exists $\lambda_{\text{low}} > 0$ small enough such that in both the individual and social case, none of the derivatives $d\mathbb{E}[K_{t+1}^i]/du_{t-k}^i$ are positive.
2. In the individual case, there exists λ such that at least some of the derivatives are positive, i.e. that it is possible to mitigate for self-interested reasons.
3. By intermediate-value theorem there is a $\lambda_{\text{critical}}$ where at least one of $d\mathbb{E}[K_{t+1}^i]/du_{t-k}^i$ is 0.
4. For the social case, the derivatives are *strictly* greater than those in the individual case i.e. company i's mitigation efforts on all other companies is positive.
5. Therefore, there exists a λ such that self-interested agents will not mitigate despite mitigation being beneficial for social welfare, thus we have a social dilemma.

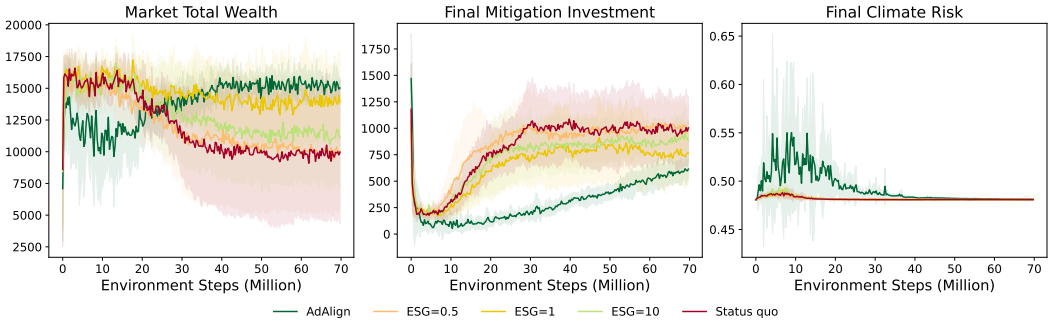


Figure 2: Training curves of PPO agents with different ESG values and Advantage Alignment (AdAlign) with $\alpha = 70$ for different metrics: Advantage Alignment agents achieve highest market total wealth by being more strategic about their mitigation investments compared to PPO agents. With significant lower final mitigation investment, AdAlign agents are able to achieve the same final climate risk and increase their capital returns. The shaded areas indicate a 1-standard deviation confidence interval. We note that 0.48 is the best achievable climate risk in the environment, as it corresponds to $1 - \prod_e (1 - P_0^e)$, the floor of the probabilities of each event.

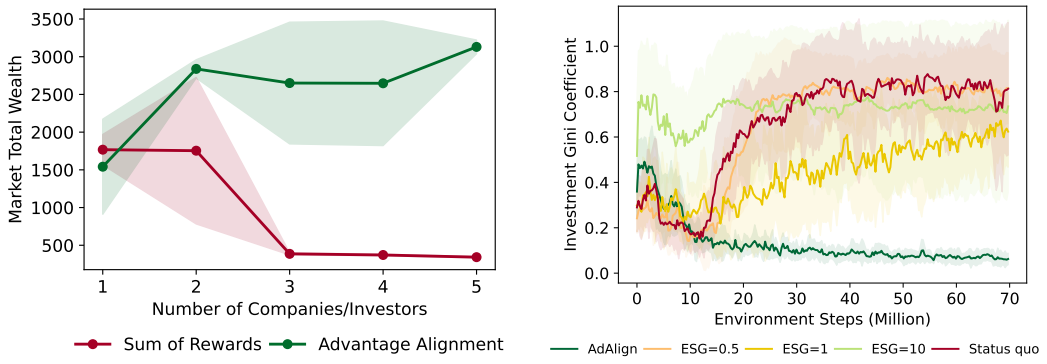
This suffices to establish the existence of 3 zones in parameter space, defined by the two parameters λ_{low} and $\lambda_{\text{critical}}$. For $\lambda < \lambda_{\text{low}}$, there is no dilemma because there is agreement between the signs of the individual and social gradients: mitigation is always net negative. If $\lambda_{\text{low}} \leq \lambda \leq \lambda_{\text{critical}}$,the gradient signs of the individual and social returns disagree, and we have a social dilemma. If $\lambda > \lambda_{\text{critical}}$, then there is again no strong dilemma and self-interested agents start mitigating myopically.

For proofs and formal theorems for each of the steps in the outline see Appendix C. We empirically validate that this threshold exists by conducting an experiment in which we train three seeds of PPO agents, keeping the number of companies and investors fixed to one, but sweeping over the scaling parameter α (Figure 4). We find the threshold to be $\alpha = 70$. After α increases beyond this threshold, PPO agents with different ESG incentives start to show different behaviors (Figure 1). We refer to this specific parameterization of InvestESG, with $\alpha = 70$, as α-InvestESG.

5 Applying Opponent Shaping to InvestESG

We apply Advantage Alignment (Duque et al., 2025) without ESG incentives to α-InvestESG (found empirically in the previous section). Crucially, our Advantage Alignment implementation differs from the PPO implementation in two aspects. First, we substitute the advantage in the policy gradient with the Advantage Alignment expression (equation 2.3). Second, following the original paper we use *self-play*: one set of parameters for the company and another set of parameters for investor players. We found that self-play makes training more stable for Advantage Alignment (AdAlign) agents but hurts PPO agents, thus we compare the best version of each in Figure 2 (AdAlign with self-play, PPO without). Advantage Alignment agents outperform all PPO agents, even those with ESG incentives, while simultaneously having the lowest amount of final mitigation investment. For details about the policy parameterization and hyper-parameters used for these experiments refer to the Appendix E.

Intuitively if we care about social welfare, then we should try to maximize it explicitly. However, this does not yield good results in practice. We run experiments with summed rewards with the default InvestESG settings (5 companies, 3 investors), however the PPO agents always converged to sub-optimal equilibria. Figure 3(a) shows that naively summing the rewards scales poorly in the multi-agent setting. We run 10 seeds of PPO agents that sum their rewards, scaling the number of agents in the environment, and plot the final market total wealth while keeping the initial capital fixed. Beyond 2 companies and 2 investors (4 agents total), PPO players summing the rewards are unable to find policies that maximize the social welfare. We hypothesize that the credit assignment problem becomes significantly more difficult as the number of agents increases, giving noisy tem-



(a) (b)

Figure 3: (a) Final market total wealth of 10 seeds of Advantage Alignment without ESG incentives, and PPO agents trained using summed rewards in InvestESG ($\alpha = 70$). On the x-axis we increase the number of companies and investors (1 company and 1 investor, 2 companies and 2 investors, etc.) while keeping the initial capital the same. Sum rewards is unable to find the action profile that maximizes social welfare once the number of players grows beyond a threshold (> 2), whereas Advantage Alignment consistently finds the same solution once the number of agents is large enough (> 1) going up to 10 agents. (b) Gini coefficient measuring inequality among company investments for different algorithms. Lower Gini indicates lower inequality and better distribution of resources. The shaded areas indicate a 1-standard deviation confidence interval.

poral difference signals to centralized critic methods. In contrast, Advantage Alignment maintains high social welfare across all experiments. We analyze the reasons why this happens in section 6.

6 On the Effectiveness of Advantage Alignment

Advantage Alignment is particularly effective at finding social welfare maximizing strategies in social dilemmas at least in part because it has a strong initial bias towards cooperative strategies. When advantages are estimated with Generalized Advantage Estimation (GAE; Schulman et al., 2018), Advantage Alignment nudges agents toward cooperation. For agent i the transformed advantage is

$$A_t^{*,i} \;=\; A_t^i \;+\; \beta\gamma \sum_{j\neq i}\Big(\sum_{k<t}\gamma^{t-k}A_k^i\Big)A_t^j,$$

with $A_t^i \equiv A^i(s_t, \mathbf{a}_t)$. Define the on-policy mean $b^i := \mathbb{E}\big[\sum_{k<t}\gamma^{t-k}A_k^i\big]$. Decomposing the expectation gives

$$A_t^{*,i} \;=\; \underbrace{A_t^i + \beta\gamma b^i \sum_{j\neq i} A_t^j}_{\text{cooperative bias}} + \beta\gamma \sum_{j\neq i}\underbrace{\Big(\sum_{k<t}\gamma^{t-k}A_k^i - b^i\Big)}_{\text{zero-mean}} A_t^j.$$

The first term favors joint gains; if $\beta\gamma b^i = 1$, the cooperative term is exactly that encountered in summed-reward learning. With perfect critics, $b^i = 0$, this bias vanishes. However, during actor-critic training, the critic lags the improving policy, so the one-step TD error

$$\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$$

is positively biased (see Kearns and Singh (2000)) for a formal argument based on contraction mappings): $\mathbb{E}[\delta_t] > 0 \implies \mathbb{E}[A_t^{\text{GAE}}] > 0 \implies b^i > 0$. Hence early training amplifies cooperative updates. As the critic catches up and TD errors become unbiased, $b^i \to 0$ and the cooperative bias fades. This strong initial bias prevents the algorithm from falling into defective equilibria early in training.

7 Related Work

Opponent shaping treats co-learners as adaptive parts of the environment and deliberately steers their updates. LOLA (Foerster et al., 2018) introduced the idea by differentiating through a single naïve policy-gradient step of other players. Follow-up methods refined the update rule (Letcher et al., 2021; Zhao et al., 2022; Willi et al., 2022) or approximated best-response dynamics (Aghajohari et al., 2024a; Lu et al., 2022). LOQA replaced higher-order differentiation with REINFORCE-based control of opponents' Q-values, enabling scalability to larger games (Aghajohari et al., 2024b). Previous work on *Advantage Alignment* showed that many of these algorithms implicitly multiply agents' advantages and offered a simpler, computationally cheaper formulation that preserves Nash equilibria in general-sum games (Duque et al., 2025). The present paper leverages that insight to shape investor–company interactions in a realistic climate-finance simulator.

A growing line of research embeds MARL agents inside climate models to study emergent behavior, test policy instruments, or optimize control strategies. RICE-N extends Nordhaus' regional integrated assessment model into a multi-agent game in which learning agents negotiate emission targets and investment pathways (Zhang et al., 2022). The *AI for Global Climate Cooperation* competition has since used this platform to study how AI negotiators can sustain long-run cooperation under enforcement uncertainty. Our work differs in scope: rather than modeling inter-government bargaining, we focus on the micro-level interaction between profit-seeking firms and ESG-motivated investors captured by InvestESG (Hou et al., 2025). Inspired by the *AI Economist* (Zheng et al., 2021), Wang et al. simulate a cap-and-trade market with enterprise and government agents; the government agent learns dynamic permit-allocation rules that balance output, equity, and emissions (Wang et al., 2024). Our study is complementary: we assume no central planner and instead demonstrate that shaping firm–investor learning alone can steer the economy toward socially desirable equilibria.

8 Conclusion

In this paper, we presented a rigorous theoretical and empirical examination of economic simulations under climate risk, using game theory. By formally characterizing the InvestESG environment as an intertemporal social dilemma, we established the precise conditions under which individual incentives diverge from collective sustainability goals. This allowed us to fine-tune InvestESG to capture the social dilemma conditions observed in the real world. Leveraging the scalable power of Advantage Alignment, we demonstrated that strategically shaping agent learning dynamics can effectively mitigate selfish incentives, systematically driving the equilibrium toward socially beneficial outcomes, even in the absence of government mandates.

Our empirical results underscore the practical efficacy of opponent-shaping algorithms in complex, high-dimensional economic scenarios, addressing a critical gap where previous multi-agent RL methods have failed. By elucidating why Advantage Alignment naturally improves the credit assignment problem and finds more cooperative equilibria, we provided key insights that pave the way for robust, policy-relevant solutions capable of maximizing the social welfare more effectively. Beyond theoretical advancements, this work signifies a tangible step toward harnessing artificial intelligence to tackle real-world climate policy challenges and inform government decision-making.

References

H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, editors. *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2022. doi: 10.1017/9781009325844. URL https://doi.org/10.1017/9781009325844.

Anatol Rapoport and Albert Chammah. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press, 1965.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Tuomas Sandholm and Robert Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Bio Systems*, 37(1-2):147–166, 1996.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness, 2018.

Juan A. Duque, Milad Aghajohari, Tim Cooijmans, Razvan Ciuca, Tianyu Zhang, Gauthier Gidel, and Aaron Courville. Advantage alignment algorithms, 2025. URL https://arxiv.org/abs/2406.14662.

Timon Willi, Alistair Letcher, Johannes Treutlein, and Jakob Foerster. Cola: Consistent learning with opponent-learning awareness, 2022.

Zongqing Lu et al. An empirical study on google research football multi-agent reinforcement learning. *Machine Intelligence Research*, 21(2):123–135, 2024. doi: 10.1007/s11633-023-1426-8. URL https://arxiv.org/abs/2305.09458.

Chris Lu, Timon Willi, Christian Schroeder de Witt, and Jakob Foerster. Model-free opponent shaping, 2022.

Milad Aghajohari, Tim Cooijmans, Juan A. Duque, Shunichi Akatsuka, and Aaron Courville. Best response shaping, 2024a.

Milad Aghajohari, Juan A. Duque, Tim Cooijmans, and Aaron Courville. Loqa: Learning with opponent q-learning awareness, 2024b. URL https://arxiv.org/abs/2405.01035.

Xiaoxuan Hou, Jiayi Yuan, Joel Z. Leibo, and Natasha Jaques. Investesg: A multi-agent reinforcement learning benchmark for studying climate investment as a social dilemma, 2025. URL https://arxiv.org/abs/2411.09856.

Lloyd Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950. doi: 10.1073/pnas.36.1.48. URL https://www.pnas.org/doi/10.1073/pnas.36.1.48.

Valerio Capraro and Joseph Y. Halpern. Translucent players: Explaining cooperative behavior in social dilemmas. In R. Ramanujam, editor, *Proceedings Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2015, Carnegie Mellon University, Pittsburgh, USA, June 4-6, 2015*, volume 215 of *EPTCS*, pages 114–126, 2015. doi: 10.4204/EPTCS.215.9. URL https://doi.org/10.4204/EPTCS.215.9.

Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007. ISBN 978-0-521-87282-9.

Scott Barrett. Self-enforcing international environmental agreements. *Oxford Economic Papers*, 46 (Supplement):878–894, 1994.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

Michael J. Kearns and Satinder Singh. Bias–variance error bounds for temporal difference updates. In Nicolò Cesa-Bianchi and Sally A. Goldman, editors, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (COLT 2000), June 28 – July 1 2000, Palo Alto, California, USA*, pages 142–147. Morgan Kaufmann, 2000. URL https://dblp.org/rec/conf/colt/KearnsS00.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games, 2021.

Stephen Zhao, Chris Lu, Roger Baker Grosse, and Jakob Nicolaus Foerster. Proximal learning with opponent-learning awareness, 2022.

Tianyu Zhang, Andrew Williams, Soham R. Phade, Sunil Srinivasa, Yang Zhang, Prateek Gupta, Y. Bengio, and Stephan Zheng. Ai for global climate cooperation: Modeling global climate negotiations, agreements, and long-term cooperation in rice-n. *Social Science Research Network*, 2022. doi: 10.48550/arXiv.2208.07004.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The ai economist: Optimal economic policy design via two-level deep reinforcement learning, 2021. URL https://arxiv.org/abs/2108.02755.

Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Carbon market simulation with adaptive mechanism design. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), Demo Track*, 2024.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

Alekh Agarwal, Nan Jiang, Sham Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. Preprint, 2021.

Ronald Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Appendix

A Reinforcement Learning Background

Consider n players interacting in an environment with policies π^i, with $i \in [1, \ldots, n]$, each parameterized by θ^i. Here, the result for each player is determined not only by their own actions but also by those of other players (von Neumann and Morgenstern, 1944). Following the convention of Agarwal et al. (2021), we can write the probability of a trajectory τ under the policies of all players as follows: $\mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}(\tau) = \mu(s_0) \cdot \prod_{i=1}^{n} \pi^i(a_0^i|s_0) \cdot P(s_1|s_0, \mathbf{a}_0) \ldots$ Where μ denotes the initial distribution over states, and $P(\cdot|s, \mathbf{a})$ is the transition distribution over states, which depends on the current state, s, and the joint action profile of all players, $\mathbf{a}$. In reinforcement learning the goal is to maximize the expected sum of future discounted reward, i.e. the value function:

$$V^i(\mu) \equiv \mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}} \left[R^i(\tau) \right] = \mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}} \left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, \mathbf{a}_t) \right].$$

We also define the action-value function, also know as the Q-value, to define the advantage function:

$$Q^i(s, \mathbf{a}) = \mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}} \left[R^i(\tau) | s_0 = s, \mathbf{a}_0 = \mathbf{a} \right], \quad A^i(s, \mathbf{a}) = Q^i(s, \mathbf{a}) - V^i(s).$$

To optimize the value function, policy gradient methods perform gradient ascent on the policy parameters following a REINFORCE estimator (Williams, 1992) of the form:

$$\nabla_{\theta^i} V^i(\mu) = \mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}} \left[\sum_{t=0}^{\infty} \gamma^t A^i(s_t, \mathbf{a}_t) \nabla_{\theta^i} \log \pi^i(a_t^i|s_t) \right].$$

The most popular variant of policy gradient algorithms, Proximal Policy Optimization (PPO) (Schulman et al., 2017), makes conservative updates to the policy using a clipping heuristic:

$$\mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi^i,\pi^{-i}}} \left[\min \left\{ \frac{\pi^i(a_t^i|s_t;\theta^i)}{\pi^i(a_t^i|s_t;\theta_l^i)} A^i(s_t, \mathbf{a}_t),\ \mathrm{clip}\left(\frac{\pi^i(a_t^i|s_t;\theta^i)}{\pi^i(a_t^i|s_t;\theta_l^i)}; 1-\epsilon, 1+\epsilon \right) A^i(s_t, \mathbf{a}_t) \right\} \right],$$

Where θ_l^i denotes the policy parameters of player i after l gradient updates. Intuitively, PPO attempts to keep the updated policy close to the original policy in terms of total variation distance.

B Economic and Environmental Dynamics of InvestESG

For notational convenience, we use the index $i \in [1, \ldots, M]$ to refer to companies and index $j \in [1, \ldots, N]$ to refer to investors. In InvestESG, each investor's state is represented by their cash level C_t^j and investment portfolio $\mathbf{S}_t^j = (H_{t,1}^j, \ldots, H_{t,M}^j)$, where $H_{t,i}^j$ denotes the investment that investor j holds in company i at timestep t. At each timestep, investors pick a binary portfolio vector $\mathbf{a}_t^j = [a_{t,1}^j, \ldots, a_{t,M}^j]$ representing which companies to invest in and companies choose to mitigate a percentage of its capital, $u_t^i \in [0, \bar{u}]$. The interim capital of each company at time $t+1$ is given by:

$$K_{t+1,\text{interim}}^i = K_t^i - \underbrace{\sum_{j=1}^{N} H_{t,i}^j}_{\text{cash returned to investors}} + \underbrace{\sum_{j=1}^{N} a_{t,i}^j \frac{\mathcal{K}_t^j}{||\mathbf{a}_t^j||_1}}_{\text{new equity investments}},$$

Where K_t^i denotes the capital of company i and $\mathcal{K}_t^j$ the capital of investor j at time t. The reward at each timestep for company i is given by the profit margin, $r_t^i = K_{t+1}^i - K_{t+1,\text{interim}}^i$. Climate dynamics, namely the risk that a particular type of harmful climate event occurs P_t^e (where e can be either an extreme heat, h, heavy precipitation, p, or drought, d, event) grows linearly w.r.t. initial event probability P_0^e following:

$$P_t^e = \frac{\mu_e t}{1 + \lambda_e U_t} + P_0^e, \qquad \lambda_e = \alpha \times \tilde{\lambda}_e,$$

Where μ_e, $\hat{\lambda}_e$, are calibration constants set to ensure that an annual mitigation investment of 2.3 trillion is required to achieve IPCC's 1.5°C scenario Hou et al. (2025). We add an additional calibration constant, α, that scales the $\hat{\lambda}$ parameters in order to make InvestESG a social dilemma (for details see section 4). The total risk is $P_t = 1 - (1 - P_t^h)(1 - P_t^p)(1 - P_t^d)$, and the cumulative mitigation U_t is:

$$U_t = U_{t-1} + \sum_{i=1}^{M} u_t^i K_{t+1,\text{interim}}^i.$$

Finally, the capital of company i is updated as follows:

$$K_{t+1}^i = (1 + \rho_t^i) K_{t+1,\text{interim}}^i, \quad \rho_t^i = (1 - u_t^i)(1 + \gamma)(1 - X_t L_i) - 1.$$

Where, ρ_t^i is the profit margin of company i. Each type of harmful climate event (extreme heat, heavy precipitation or drought) is modeled independently by Bernoulli(P_t^e), where $X_t \in \{0, 1, 2, 3\}$ is the total number of climate events realized at time t. L_i above denotes the loss coefficient of company i and γ (which differs from the discount factor) is the market performance baseline.

C Mathematical Derivations

In this section we prove the main result of this work, namely, that InvestESG is a social dilemma for the correct choice of the climate mitigation effectiveness parameter λ. We begin by stating our assumptions about the environment:

Assumption 1. (Static Investor Portfolios). *Investor holdings are time–invariant. Formally,* $H^j_{t,i} = H^j_{0,i}\ \forall\, t \geq 0,\ i \in [M],\ j \in [N]$. *Thus investors never rebalance their capital allocations.*

Assumption 2. (Single-Company Allocation). *Each investor supplies capital to exactly one firm.*

Assumption 3. (No Historical Mitigation). *Prior to the period of interest, firms have undertaken no mitigation expenditure:* $u^i_s = 0\ \forall\, s < t,\ i \in [M]$.

C.1 Individual Case Analysis

The capital of a company is:

$$K^i_{t+1} = (1+\rho^i_t)K^i_{t+1,\text{interim}}.$$
$$\rho^i_t = (1-u^i_t)(1+\gamma)(1-X_tL_i) - 1.$$
$$K^i_{t+1,\text{interim}} = K^i_t - \sum_{j=1}^{N} H^j_{t,i} + \sum_{j=1}^{N} a^j_{t,i}\frac{\mathcal{K}^j_t}{||\mathbf{a}^j_t||_1}.$$

where $\mathcal{K}^j_t$ is the capital of the j-th investor. In this argument we will isolate the companies' behavior by assuming that the investors do not change their capital allocations over time, and that each investor invests in a single company. This means that the money returned to investors is the same as the money investors put in at the current time-step. Giving us

$$K^i_{t+1,\text{interim}} = K^i_t,$$

Plugging this into the equation for K^i_{t+1} gives us:

$$K^i_{t+1} = \Big((1-u^i_t)(1+\gamma)(1-X_tL_i)\Big)K^i_t.$$

Here X_t is the number of climate events at time t, it can go up to 3, if extreme heat, precipitation and drought all happen. The probability of each of those events in X_t depends on the mitigation up to time t, called U_t:

$$P^e_t \;=\; \frac{\mu_e t}{1+\lambda_e U_t} \;+\; P^e_0, \qquad \lambda^e = \alpha \times \tilde{\lambda}^e,$$

Let's consider a company which has never done any mitigation in the past, so $U_{t-1} = 0$, but is considering doing so at this time-step, we can differentiate $\mathbb{E}[K^i_{t+1}]$ with respect to u^i_t to investigate if it has an incentive to start mitigating. If this derivative is positive, then the company has an immediate, myopic incentive to take mitigation actions.

Lemma 1. *For every time-step t, company i and scalar $\alpha_i > 0$, we have that the social marginal gradient is strictly greater than the private marginal gradient.*

$$\frac{d}{du^i_t}\sum_j \alpha_j \mathbb{E}\Big[K^j_{t+1}\Big] > \alpha_i \frac{d}{du^i_t}\mathbb{E}\Big[K^i_{t+1}\Big].$$

Proof.

$$\frac{d}{du^i_t}\mathbb{E}\Big[K^i_{t+1}\Big] = \frac{d}{du^i_t}\mathbb{E}\Big[\Big((1-u^i_t)(1+\gamma)(1-X_tL_i)\Big)K^i_t\Big].$$

The only parts of this equation that depend on u_t^i are the factor of $(1-u_t^i)$ and through $(1-X_tL_i)$. Now for small values of climate risk events, we can approximate X_t as a Bernoulli variable with probability $P_t = \sum_e P_t^e$, hence the expectation above can be written as a sum over $X_t = \{0,1\}$, the algebraic trick used here comes from $f(x,y,z) = xyz \implies df/dx = yz = f(x,y,z)/x$, giving us:

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} - \mathbb{E}\left[\frac{K_{t+1}^i}{1-X_tL_i}\frac{dP_t}{du_t^i}\right].$$

And now for $\frac{dP_t}{du_t^i}$, referring back to equation C.1:

$$\frac{dP_t}{du_t^i} = \frac{dU_t}{du_t^i}\frac{d}{dU_t}\sum_e\left(\frac{\mu_e t}{1+\lambda_e U_t} \;+\; P_0^e\right) = K_{t+1,\text{interim}}^i \sum_e \frac{-\lambda_e\mu_e t}{(1+\lambda_e U_t)^2},$$

Where the factor of $K_{t+1,\text{interim}}^i$ comes from the fact that $U_t \equiv \sum_{s<t} u_s^i K_{s+1,\text{interim}}^i$. And finally:

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} + \mathbb{E}\left[\frac{K_{t+1}^i}{1-X_tL_i}K_{t+1,\text{interim}}^i\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right].$$

This equation places a lower bound on how incentivized a selfish company is to invest in mitigation, it does not include the long-term effects of mitigation at all. If λ_e is set high enough that this derivative is positive, we do not have a social dilemma at all. Noticing that $K_{t+1,\text{interim}}^i = K_{t+1}^i/(1+\rho_t^i)$ (from the equations at the beginning of the section), we can replace:

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} + \mathbb{E}\left[\frac{(K_{t+1}^i)^2}{(1-X_tL_i)^2(1-u_t^i)(1+\gamma)}\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right].$$

We can also trivially run the same steps for $\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^j\right]$ with $j \neq i$, i.e. computing the derivative of the return of agent j w.r.t. the mitigation of agent i, which will remove the first term in the previous equation. The cumulative mitigation U_t includes terms from all agents, so the contribution of the $(1-X_tL_j)$ term to the gradient remains:

$$\text{for } i \neq j: \frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^j\right] = \mathbb{E}\left[\frac{K_{t+1}^jK_{t+1}^i}{(1-X_tL_j)(1-X_tL_i)(1-u_t^i)(1+\gamma)}\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right].$$

We notice that this is strictly non-negative, as expected (since all K_t^i are non-negative). All agents would want someone else to mitigate for them, they get all the benefits of mitigation without any of the costs. Together with the non-negativity of K_t^i, we get the following inequality:

$$\forall i \qquad \frac{d}{du_t^i}\sum_j\mathbb{E}\left[K_{t+1}^j\right] > \frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right],$$

which falls straight out of equation C.1 being strictly non-negative. This means that the "social derivative" is always greater than the selfish derivative for mitigation actions u_t^i. In particular, we also get the extra inequality:

$$\forall t\forall i \in \{0,...,n\},\ \forall\alpha_i > 0 \qquad \frac{d}{du_t^i}\sum_j\alpha_j\mathbb{E}\left[K_{t+1}^j\right] > \alpha_i\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right].$$

which we will need later. □

C.2 Dependence on all Previous Mitigation Actions

Lemma 2. *For every **previous** time-step $t-k,\ 0<k<t$, company i and scalar $\alpha_i>0$, we have that the social gradient is strictly greater than the private gradient.*

$$\frac{d}{du^i_{t-k}}\sum_l \alpha_l \mathbb{E}\Big[K^l_{t+1}\Big] > \frac{d}{du^i_{t-k}}\alpha_i \mathbb{E}\Big[K^i_{t+1}\Big].$$

Proof. Writing the main equation again for visibility:

$$K^i_{t+1} = \Big((1-u^i_t)(1+\gamma)(1-X_tL_i)\Big)K^i_t.$$

To compute the dependence of K^i_{t+1} on u^i_{t-k} for a general $k\geq 1$, we note that this enters in two ways in the equation above, first through the effect on P_i, which affects K^j_{t+1} through $(1-u^i_t)(1+\gamma)(1-X_tL_i)$, and this will give us a factor like the second one in equation C.1. The second way is through effects on K^i_t, which were not present in the previous section:

$$\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_{t+1}\Big] = -\mathbb{E}\Big[\frac{K^i_{t+1}}{1-X_tL_i}\frac{dP_t}{du^i_{t-k}}\Big] + \Big((1-u^i_t)(1+\gamma)(1-P_tL_i)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_t\Big].$$

We have used the independence of X_t and K_t to push in the expectation values. These two variables are independent if the past mitigation actions are kept constant, which is the case for us, since we are analyzing the case where all companies have never invested in mitigation (by construction). dP_t/du_{t-k} is computed as before:

$$\frac{dP_t}{du^i_{t-k}} = \frac{dU_t}{du^i_{t-k}}\frac{d}{dU_t}\sum_e\Big(\frac{\mu_e t}{1+\lambda_e U_t}\ +\ P^e_0\Big) = K^i_{t+1-k,\text{interim}}\sum_e\frac{-\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}.$$

We can now use equation C.2 to induce a recurrence relation between the derivatives of all factors $\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_{t+1}\Big]$. All put together, the recurrence relation is

$$\begin{aligned}\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_{t+1}\Big] =\ & \mathbb{E}\Big[\frac{(K^i_{t+1})^2}{(1-X_tL_i)^2(1-u^i_t)(1+\gamma)}\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\Big]\\ &+\Big((1-u^i_t)(1+\gamma)(1-P_tL_i)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_t\Big].\end{aligned}$$

And for the derivative with a different agent $l\neq i$:

$$\begin{aligned}\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^l_{t+1}\Big] =&\mathbb{E}\Big[\frac{K^i_{t+1}K^l_{t+1}}{(1-X_tL_i)(1-X_tL_l)(1-u^i_t)(1+\gamma)}\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\Big]\\ &+\Big((1-u^l_t)(1+\gamma)(1-P_tL_l)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^l_t\Big].\end{aligned}$$

We now wish to get the equivalent of lemma 1 for mitigations u^i_{t-k} at any previous time-step. We use a proof by induction on the index k from the base case of $k=0$ in equation C.1. Let $\alpha_i>0$ be strictly positive real numbers:

$$\begin{aligned}\frac{d}{du^i_{t-k}}\sum_l\alpha_l\mathbb{E}\Big[K^l_{t+1}\Big] =&\\ &\sum_l\alpha_l\mathbb{E}\Big[\frac{K^i_{t+1}K^l_{t+1}}{(1-X_tL_i)(1-X_tL_l)(1-u^i_t)(1+\gamma)}\sum_e\frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\Big]\\ &+\sum_l\alpha_l\Big((1-u^l_t)(1+\gamma)(1-P_tL_l)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^l_t\Big].\end{aligned}$$

We now invoke the induction assumption for the term in the last line, which matches the induction case with $\alpha'_l = \alpha_l\big((1-u^l_t)(1+\gamma)(1-P_tL_l)\big)$, this is why we needed the assumption with a positive linear combination. The assumption gives us the following:

$$\sum_l \alpha_l \Big((1-u^l_t)(1+\gamma)(1-P_tL_l)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^l_t\Big] > \alpha_i\Big((1-u^i_t)(1+\gamma)(1-P_tL_i)\Big)\frac{d}{du^i_{t-k}}\mathbb{E}\Big[K^i_t\Big],$$

But now we note that the elements in the first sum above are all positive, and so

$$\sum_l \alpha_l \mathbb{E}\Big[\frac{K^i_{t+1}K^l_{t+1}}{(1-X_tL_i)(1-X_tL_l)(1-u^i_t)(1+\gamma)}\sum_e \frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\Big] > \alpha_i\mathbb{E}\Big[\frac{(K^i_{t+1})^2}{(1-X_tL_i)^2(1-u^i_t)(1+\gamma)}\sum_e \frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\Big],$$

This suffices to establish the desired inequality

$$\frac{d}{du^i_{t-k}}\sum_l \alpha_l\mathbb{E}\Big[K^l_{t+1}\Big] > \frac{d}{du^i_{t-k}}\alpha_i\mathbb{E}\Big[K^i_{t+1}\Big],$$

Which as a special case ($\alpha_l = 1$ for all l) tells us that the gradient of the social welfare with respect to the mitigation actions of any agent at any point in the past is higher than the selfish gradient. □

C.3 Induction Argument Re-stated

Theorem 1 (Social-mitigation)**.** *For all* $\lambda > 0$ *,* $\forall t, i, k$,

$$\frac{d}{du^i_{t-k}}\mathbb{E}\left[\sum_j K^j_{t+1}\right] > \frac{d}{du^i_{t-k}}\mathbb{E}\left[K^i_{t+1}\right].$$

Proof. In order to prove theorem 1, we need to prove the following equation for all i, t, k:

$$\frac{d}{du^i_{t-k}}\sum_l \alpha_l\mathbb{E}\Big[K^l_{t+1}\Big] > \frac{d}{du^i_{t-k}}\alpha_i\mathbb{E}\Big[K^i_{t+1}\Big],$$

we proceed by induction on k, using $k = 0$ as base case:

Induction base case $k = 0$**.** The base case is trivial from lemma 1.

Induction recursion. For the induction, we assume that equation C.3 holds for all $k' < k$, and show that this implies that the equation is true for k. This is what equations C.2 through C.2 are showing.

By setting $\alpha_l = 1$ in the equation, we finish the proof. □

C.4 All Selfish Gradients are Negative at Low Enough λ

Theorem 2 (Low mitigation effectiveness)**.** *There exists* $\lambda_{low} > 0$ *such that for all* $\lambda < \lambda_{low}$*, we have:*

$$\frac{d}{du^i_{t-k}}\mathbb{E}\big[K^i_{t+1}\big] < 0 \quad \text{and} \quad \frac{d}{du^i_{t-k}}\sum_j \mathbb{E}\big[K^j_{t+1}\big] < 0 \quad \forall i, t, k.$$

Proof. We proceed by simple recurrence to show that all gradients are negative at $\lambda_e = 0$. We note that equation C.1 at $\lambda_e = 0$ reduces to

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i},$$

thus this gradient is negative. Using equation C.2, we see that the first term vanished in the $\lambda_e = 0$, giving us a term which only depends on $\frac{d}{du_{t-k}^i}\mathbb{E}\left[K_t^i\right]$, which we know is negative:

$$\frac{d}{du_{t-k}^i}\mathbb{E}\left[K_{t+1}^i\right] = \Big((1-u_t^i)(1+\gamma)(1-P_tL_i)\Big)\frac{d}{du_{t-k}^i}\mathbb{E}\left[K_t^i\right] < 0.$$

Thus there exists a λ_e such that all the selfish gradients are negative, completing the proof. □

C.5 Some Selfish Gradients are Positive

Theorem 3 (Self-interested mitigation)**.** *There exists $\lambda > 0$ such that:*

$$\forall t, i, \exists k, \quad \frac{d}{du_{t-k}^i}\mathbb{E}\big[K_{t+1}^i\big] > 0.$$

Proof. Again from equation C.1:

$$\frac{d}{du_t^i}\mathbb{E}\left[K_{t+1}^i\right] = -\frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i} + \mathbb{E}\left[\frac{(K_{t+1}^i)^2}{(1-X_tL_i)^2(1-u_t^i)(1+\gamma)}\sum_e \frac{\lambda_e\mu_e t}{(1+\lambda_e U_t)^2}\right].$$

We evaluate the gradient at the $U_t = 0$ policy, hence we can isolate the λ_e where the sign flips:

$$\lambda_e^{\text{sign flip}} \times \mathbb{E}\left[\frac{(K_{t+1}^i)^2}{(1-X_tL_i)^2(1-u_t^i)(1+\gamma)}\sum_e \mu_e t\right] = \frac{\mathbb{E}[K_{t+1}^i]}{1-u_t^i},$$

which implies that at λ sufficiently high, at least some selfish gradients become positive, and the theorem is proved. □

C.6 Main Result

We can now state the main social dilemma theorem, which says that there is a λ where all the individual gradients are strictly negative, but at least some social gradients are positive.

Corollary 1 (InvestESG is a Social Dilemma)**.** *There exists some mitigation effectiveness parameter $\lambda > 0$ such that*

$$\forall t, i, k, \quad \frac{d}{du_{t-k}^i}\mathbb{E}\big[K_{t+1}^i\big] < 0 \quad \text{and} \quad \forall t\ \exists i, k \quad \frac{d}{du_{t-k}^i}\sum_j \mathbb{E}\big[K_{t+1}^j\big] > 0.$$

Proof. From the intermediate value theorem and theorems 2 and 3, we conclude that there exists $\lambda_{\text{critical}}$ for which $\exists t, i, k \frac{d}{du_{t-k}^i}\mathbb{E}\big[K_{t+1}^i\big] = 0$, and from continuity of $\frac{d}{du_{t-k}^i}\mathbb{E}\big[K_{t+1}^i\big]$ with respect to λ we infer that there exists $\epsilon > 0$ such that for $\lambda_{\text{critical}} \geq \lambda > \lambda_{\text{critical}} - \epsilon$ we have $\forall t, i, k \frac{d}{du_{t-k}^i}\mathbb{E}\big[K_{t+1}^i\big] < 0$. Now theorem 1 applied at $\lambda_{\text{critical}}$ tells us $\exists t, i, k \quad \frac{d}{du_{t-k}^i}\mathbb{E}\big[\sum_j K_{t+1}^j\big] > 0$, and again by continuity this quantity remains positive at some smaller $\lambda_{\text{critical}} - \epsilon^*$. Thus $\lambda = \lambda_{\text{critical}} - \min(\epsilon, \epsilon^*)$ ensures that all individual gradients are negative while there exists a social gradient which is positive. □

D On the Cooperative Bias of Advantage Alignment

Advantage Alignment has a bias towards promoting cooperation between agents when the advantages are estimated using generalized advantage estimation (GAE) (Schulman et al., 2018). The main insight is that TD-1 errors have a positive bias for on-policy trajectories because the critic lags behind the actor; advantages are systematically underestimated, producing a cooperation bias in the algorithm. The advantage alignment updated advantages are:

$$A^{*,i}(s_t, \mathbf{a}_t) = \left(A^i(s_t, \mathbf{a}_t) + \beta\gamma \cdot \sum_{j \neq i} \left(\sum_{k<t} \gamma^{t-k} A^i(s_k, \mathbf{a}_k) \right) A^j(s_t, \mathbf{a}_t) \right)$$

The term $\sum_{k<t} \gamma^{t-k} A^i(s_k, \mathbf{a}_k)$ acts as a multiplicative factor for the opponent advantages. If we let $b^i = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi^i, \pi^{-i}}} \left[\sum_{k<t} \gamma^{t-k} A^i(s_k, \mathbf{a}_k) \right]$, we can split the expression into its biased and unbiased parts:

$$A^{*,i}_{\text{biased part}}(s_t, \mathbf{a}_t) = \left(A^i(s_t, \mathbf{a}_t) + b^i \times \beta\gamma \sum_{j \neq i} A^j(s_t, \mathbf{a}_t) \right)$$

$$A^{*,i}_{\text{unbiased part}}(s_t, \mathbf{a}_t) = \beta\gamma \cdot \sum_{j \neq i} \left(-b^i + \sum_{k<t} \gamma^{t-k} A^i(s_k, \mathbf{a}_k) \right) A^j(s_t, \mathbf{a}_t)$$

It is this biased part of the new advantages which produce the cooperation bias. In the extreme case where $b \times \beta\gamma = 1$, the learning process collapses to simply doing summed-reward training.

For advantage functions $A^j(s_t, \mathbf{a})$ which perfectly evaluate the behavior of the policies π_j, the terms above have $b^i = 0$, yielding no collaborative bias for advantage alignment. However, Generalized Advantage Estimation computes the advantages as:

$$\delta_t^V := r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$$

$$\hat{A}_t^{\text{GAE}(\lambda,\gamma)} := \sum_{l=0}^{\infty} (\gamma\lambda)^t \delta_{l+t}^V$$

Where the value networks V_ϕ are trained to evaluate policies π_θ which are themselves learning and improving. As π improves, the value networks perpetually lag behind, computing the value of slightly time-delayed policies. This implies that $V_\phi(s_t)$ will on average underestimate the next-step rewards see Kearns and Singh (2000), and this will give us a positive bias to the TD-errors seen on-policy:

$$\mathbb{E}_{\tau \sim \Pr_\mu^{\pi^i, \pi^{-i}}} [\delta_t^V] > 0$$

$$\implies \mathbb{E}_{\tau \sim \Pr_\mu^{\pi^i, \pi^{-i}}} \left[\hat{A}_t^{\text{GAE}(\lambda,\gamma)} \right] > 0$$

$$\implies \hat{b}^i = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi^i, \pi^{-i}}} \left[\sum_{k<t} \gamma^{t-k} \hat{A}^{i,\text{GAE}(\lambda,\gamma)}(s_k, \mathbf{a}_k) \right] > 0$$

This bias towards cooperation only manifests at the beginning of training, before the critic has fully caught up with the actor. At convergence of the actor, the critic does catch up, and the TD-errors become unbiased, removing the cooperative bias.

E Experimental Details

Throughout our experiments, we trained using the Adam optimizer (Kingma and Ba, 2017) and trained for 70 million environment steps (approximately 20000 gradient steps), as anything beyond that resulted in lower performance. Our experiments ran in approximately 4 hours on a NVIDIA L40s gpu.

E.1 PPO Policy Parameterization and Hyper-Parameters

As in the InvestESG paper (Hou et al., 2025), we use a two-layer fully-connected MLP with hidden size of 256 neurons and ReLU activations to parameterize the mean of a multivariate normal distribution. We use an additional parameter to parameterize the log standard deviation. For the critic, we use the same architecture (two-layer MLP, ReLU activations, with hidden size 256) to estimate the value directly. We swept over hyper-parameters to find the best for maximizing market total wealth. For all PPO experiments we used:

Hyperparameter	**Value**
Algorithm	IPPO
Total Environment Steps	70M
Number of Environments	64
Episode Length	100
Discount Factor (γ)	0.99
GAE λ	0.95
Policy Learning Rate	1e-4
Value Function Learning Rate	1e-4
Entropy Coefficient	0.05
Clip Ratio (ϵ)	0.2
Value Function Clip Range	10
Number of PPO Epochs	4
Number of Minibatches	20
Gradient Clipping	10.0
MLP Hidden Size	64
MLP Hidden Layers	2
Self-Play	False

Table 1: Key PPO hyperparameters used for training in the InvestESG environment.

E.2 Advantage Alignment Policy Parameterization and Hyper-Parameters

As in theprevious section, we use a two-layer fully-connected MLP with hidden size of 256 neurons and ReLU activations to parameterize the mean of a multivariate normal distribution. We use an additional parameter to parameterize the log standard deviation. For the critic, we use the same architecture (two-layer MLP, ReLU activations, with hidden size 256) to estimate the value directly. We swept over Advantage-Alignment specific hyper-parameters (keeping the PPO hyper-parameters fixed) to find the best for maximizing market total wealth. For all Advantage Alignment experiments we used:

Hyperparameter	**Value**
Algorithm	AdAlign
Total Environment Steps	70M
Number of Environments	64
Episode Length	100
Discount Factor (γ)	0.99
GAE λ	0.95
Policy Learning Rate	1e-4
Value Function Learning Rate	1e-4
Entropy Coefficient	0.05
Clip Ratio (ϵ)	0.2
Value Function Clip Range	10
Number of PPO Epochs	1
Number of Minibatches	20
Gradient Clipping	10.0
Use RNN	False
MLP Hidden Layers	2
Self-Play	True
Advantage Alignment β	0.2
Advantage Alignment γ	0.9

Table 2: Key Advantage Alignment hyperparameters used for training in the InvestESG environment.

F Additional Figures

F.1 Sweep over Lambda

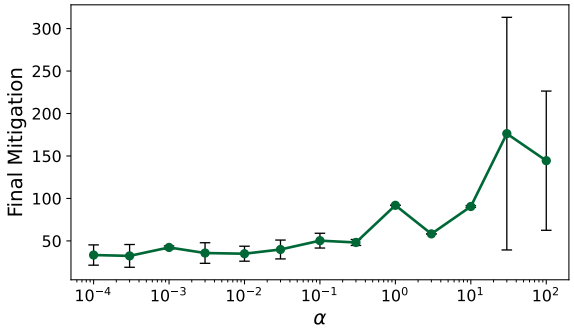


Figure 4: Final mitigation amount of 3 seeds of PPO agents trained in a single company and single investor environment for different values of α. Here the public and private gradients are the same and we clearly see a threshold $\alpha \approx 30$ at which the final mitigation amount significantly increases. This threshold empirically marks the change of sign of the gradient. The whiskers indicate a 1-standard deviation confidence interval.

F.2 Interpreting Neural Network Policies

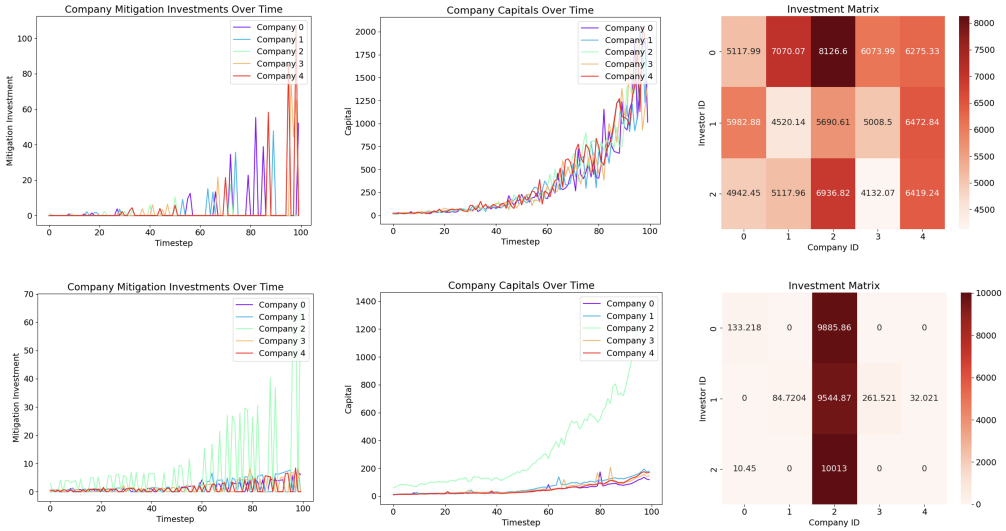


Figure 5: Policy dynamics in the ESG = 0 social dilemma. Top: Advantage Alignment (AA); bottom: PPO baseline. (i) Mitigation over time: AA invests only at climate-risk peaks and then returns to zero, whereas PPO agents invest erratically throughout. (ii) Company capital: AA keeps wealth roughly uniform across firms, while PPO gradually concentrates it. (iii) Final investment matrix: AA yields an almost symmetric allocation, whereas PPO shows fragmented, concentrated investments.

To assess how opponent shaping alters investment behaviors in the social-dilemma setting, we run simulations of Advantage Alignment (AA) and the standard PPO baseline with ESG score $= 0$, which means that there is no external incentive to mitigate. The policies obtained are summarized in Fig. 5. Three patterns emerge. (i) AA learns to mitigate just enough and only at the critical moments when climate risk increases, whereas PPO invests excessively. (ii) The capital allocations show that AA maintains an approximately uniform investment distribution across companies, while

PPO concentrates wealth, replicating the disparities reported by Hou et al. (2025) (Figure 3(b)). (iii) AA agents learn to coordinate themselves to share mitigation costs, in contrast to PPO agents whose independent updates yield uncoordinated, non-cooperative mitigation investments.