

AI in Corporate Governance: Can Machines Recover Corporate Purpose?

Finance Working Paper N° 1048/2025

March 2025

Boris Nikolov

University of Lausanne, SFI and ECGI

Norman Schuerhoff

University of Lausanne, SFI and CEPR

Sam Wagner

University of Lausanne and SFI

© Boris Nikolov, Norman Schuerhoff and Sam Wagner 2025. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

This paper can be downloaded without charge from:
http://ssrn.com/abstract_id=5166191

www.ecgi.global/content/working-papers

ECGI Working Paper Series in Finance

AI in Corporate Governance: Can Machines Recover Corporate Purpose?

Working Paper N° 1048/2025

March 2025

Boris Nikolov
Norman Schuerhoff
Sam Wagner

Abstract

A key question in automating governance is whether machines can recover the corporate objective. We develop a corporate recovery theorem that establishes when this is possible and provide a practical framework for its application. Training a machine on a large dataset of firms' investment and financial decisions, we find that most neoclassical models fail to explain the data since the machine learns from managers to underestimate the shadow cost of capital. This bias persists even after accounting for financial frictions, intangible intensity, behavioral factors, and ESG. We develop an alignment measure that shows why managerial alignment with shareholder-value remains imperfect and how to debias managerial decisions.

Keywords: Corporate Purpose, Inverse Reinforcement Learning

JEL Classifications: D22, G30, L21

Boris Nikolov*
Professor of Finance
University of Lausanne
Extranef 228
1015 Lausanne, Switzerland
phone: +41 216 926 126
e-mail: boris.nikolov@unil.ch

Norman Schuerhoff
Professor of Finance
University of Lausanne
Extranef 239
1015 Lausanne, Switzerland
e-mail: norman.schuerhoff@unil.ch

Sam Wagner
Researcher
University of Lausanne
Extranef 252
1015 Lausanne, Switzerland
e-mail: sam.wagner@unil.ch

*Corresponding Author

Swiss Finance Institute

Research Paper Series

N°25-23

AI in Corporate Governance: Can Machines
Recover Corporate Purpose?



Boris Nikolov

University of Lausanne, Swiss Finance Institute, and ECGI

Norman Schürhoff

University of Lausanne, Swiss Finance Institute, and CEPR

Sam Wagner

University of Lausanne and Swiss Finance Institute

AI in Corporate Governance: Can Machines Recover Corporate Purpose?

Boris Nikolov*, Norman Schürhoff†, Sam Wagner‡

February, 2025

Abstract

A key question in automating governance is whether machines can recover the corporate objective. We develop a corporate recovery theorem that establishes when this is possible and provide a practical framework for its application. Training a machine on a large dataset of firms' investment and financial decisions, we find that most neoclassical models fail to explain the data since the machine learns from managers to underestimate the shadow cost of capital. This bias persists even after accounting for financial frictions, intangible intensity, behavioral factors, and ESG. We develop an alignment measure that shows why managerial alignment with shareholder-value remains imperfect and how to debias managerial decisions.

Keywords: Corporate Purpose, Inverse Reinforcement Learning

JEL Classification: D22, G30, L21

*Faculty of Business and Economics at the University of Lausanne, Swiss Finance Institute (SFI), and European Corporate Governance Institute (ECGI). E-mail: boris.nikolov@unil.ch. Address: Extranef 227, 1015 Lausanne, Switzerland.

†Faculty of Business and Economics at the University of Lausanne, Swiss Finance Institute (SFI), and Center for Economic and Policy Research (CEPR). E-mail: norman.schuerhoff@unil.ch. Address: Extranef 239, 1015 Lausanne, Switzerland.

‡University of Lausanne and Swiss Finance Institute (SFI). E-mail: sam.wagner@unil.ch. Address: Extranef 252, 1015 Lausanne, Switzerland.

Corporate purpose, or the objective function of the firm defines the goals and values that management aims to achieve. The doctrine for the last century has been that firms ought to maximize shareholder value, which means they invest and finance to increase the wealth of shareholders (Friedman, 1970). Shareholder-value maximization has been challenged by alternative theories focusing on misalignment of incentives, stakeholders, welfare, and non-standard preferences. A firm can and in some cases should have objectives beyond maximizing shareholder value for a number of reasons. CEOs and other corporate executives may face incentive problems leading to agency conflicts in which their personal interests do not align with the interests of shareholders and which can result in suboptimal decisions and outcomes for the corporation (Jensen and Meckling, 1976).¹ Corporations may also have environmental and social responsibility objectives, such as reducing carbon footprint or promoting diversity and inclusion; other objectives may relate to employee satisfaction and well-being, product quality, or customer satisfaction (Freeman, 1984). Ultimately, the objective function of a firm likely depends on a number of features, including its governance, social objectives, environmental impact and even organizational structure, ownership, industry, culture, and the values and goals of all its stakeholders.

In this paper we ask if a machine can recover the de-facto corporate objective from data on managers' observed behavior. This is a key issue in automating corporate governance. The firm's objective is not directly observable in the data. However, if the reward associated with a managerial action can be measured across different economic states, one can derive conditions under which the objective function of the firm and the features that determine it can be recovered. In addition, one can go beyond identifying the reward function and generalize to new economic environments by training a machine to perform the same task(s) as the CEO or other managers. This may ultimately allow to automate corporate decision making and other complex managerial tasks. After establishing conditions for recoverability and generalizability, we show empirically what de-facto objective function of the firm can be recovered from data on firms' observed investment, financing, and operational policies and what insights this provides about the most pertinent economic frictions faced by corporations. We also develop an alignment measure based on the firm's recovered value function that quantifies the extent of managerial alignment with shareholder value.

A recent strand of artificial intelligence techniques, Inverse Reinforcement Learning (IRL), recovers an agent's reward function or objective function based on observed behavior.² The basic idea of IRL is to

¹To address agency conflicts, corporations often implement governance mechanisms to better align the managerial incentives with those of the shareholders, including executive compensation, board of directors, ownership structure, auditing and monitoring, and shareholder activism.

²IRL learns the reward function by observing the optimal policy/behavior. IRL is the reverse of the reinforcement

observe the actions of a human expert, in our case the CEO managing the firm, and then infer the reward function that the expert is maximizing.³ Once the reward function is inferred, it is interpretable and can be used to train an artificial agent to perform the task in a new economic environment similar to the expert without the assistance of the original expert. IRL is a challenging problem and generally ill-defined. The majority of the existing IRL literature has therefore focused on a simpler task, recovering the agent’s policy function, instead of the agent’s objective or reward function. The advantage of recovering the policy is that it can directly be applied to train a new agent. The disadvantage is that policy-based IRL is not necessarily robust to exogenous factors and does not naturally generalize to new environments. Recently, [Cao et al. \(2021\)](#) and [Rolland et al. \(2022\)](#) have shown that the agent’s objective function and state-action contingent rewards can be recovered in a Markov decision problem (MDP) under some regularity conditions. While [Cao et al. \(2021\)](#) assume a condition that the agents’ MDPs are value-distinguishing, [Rolland et al. \(2022\)](#) only impose rank conditions on the state transition probabilities and illustrate how they can be checked empirically. They then show that the value function can be recovered state by state up to a constant if the agents’ policies are entropy regularized by a known penalty, such that they are stochastic and ergodic, and either the same agent acts in environments varying over time or different agents share the same reward. The limiting assumption is that the model imposes optimizing behavior.

Based on these recent insights in computer science, we develop a corporate recovery theorem that establishes when a machine can recover corporate purpose. We apply our IRL technology to financial data on firms’ investment and financial policies which allows us to recover the importance of various features affecting firms’ objectives, which we then link to competing theories of the firm. We start with an infinite-horizon Markov decision problem in which the agent (the manager of the firm) may have some incentive for experimentation. In this framework, we generalize the results in [Cao et al. \(2021\)](#) and [Rolland et al. \(2022\)](#) to the economically relevant case where the agents’ reward functions contain additively separable preference shocks and the relative importance of exploration vs. exploitation in the agent’s objective function (i.e., the entropy regularization weight) is unknown to the observer/machine, which is a scenario likely to be the case in economics and finance. We show that the agent’s instantaneous rewards and value function can still be recovered up to affine transformations for each state-action pair observed in the data.

learning (RL) approach, where the agent learns the optimal policy based on a given reward function ([Sutton and Barto, 2018](#)). RL is typically used in environments where the agent can experiment and learn by receiving feedback from rewards, such as in games, robotics, or financial trading. [Campello et al. \(2024\)](#) apply RL in corporate finance.

³We do not take a stance on whether the firm is run by the single CEO, a group of managers, or managers and employees jointly. We identify the reward structure of the effective decision maker(s).

Based on these results, we provide a practical framework for its application in economics and finance, which in its simplest form is similar to augmented multivariate regressions specifying a set of features (or independent variables) and recovering the coefficients on these features.

Applying the corporate recovery framework to a large dataset on investment and financial policies of publicly traded U.S. firms and features from various data sources, we document a number of stylized facts about corporate purpose. Our first main result concerns the relative importance of current profits to investment. Neoclassical models postulate that firms’ investment is determined by expected profitability and the cost of capital, that is, the reward function is of the form $Profitability - p \times InvestmentRate$, reflecting a price of capital equal to $p = 1$. However, the machine learns that managers act as if they underestimate the cost of investment by up to 35-55%, or the manager’s shadow price of capital p ranges from 0.45 to 0.65 depending on the specification, significantly less than one. This finding is robust to alternative definitions of investment and capital, the inclusion of additional features such as proxies for financing frictions, control features, managerial compensation and ESG scores, as well as different discount rates. These results suggest *ad faciem* that machine-run firms would over-invest and under-explore.

The over-investment is pervasive across profitability levels, with the relative increase the largest for low levels of profitability. In addition, managers appear to do too little exploration, i.e., experiment by varying investment less than optimal. This *corporate reward puzzle* is inconsistent with the predictions from the foundational q -theory model and calls for alternative theories.⁴ While both over- and under-investment have been documented empirically (Blanchard et al., 1994; Cho, 1998; Richardson, 2006; Ferreira and Matos, 2008; Cronqvist and Fahlenbrach, 2009), under-exploration and hence too little experimentation over investment opportunities have not been documented before. We link these distortions in the observed corporate behavior to the manager’s shadow cost of capital, which is too low. Consistent with our findings, Ben-David and Chinco (2024) demonstrate that a parametric model where managers are EPS maximizers can account for observed capital budgeting behavior, but they do not explicitly consider investment rates nor exploration incentives. Jha et al. (2024) use ChatGPT to compute a firm-level investment score and show that high-investment-score firms experience negative future abnormal returns, consistent with

⁴Over-investment is consistent with models of asymmetric information, agency costs, and managerial overconfidence. The free cash flow hypothesis of Jensen (1986) (see also Richardson (2006)) suggests that firms with substantial free cash flow and limited growth opportunities are prone to overinvestment, as managers prefer to invest excess cash rather than return it to shareholders. Agency cost models (Myers and Majluf, 1984; Shleifer and Vishny, 1989; Stulz, 1990) highlight conflicts between shareholders and managers. In situations where managerial incentives are misaligned with shareholder interests, managers may overinvest in projects that do not maximize shareholder value. Managerial overconfidence models (Malmendier and Tate, 2005) incorporate behavioral biases to predict that managers will overestimate the returns on investment projects, leading to overinvestment.

value-destructing over-investment but ChatGPT is silent about exploration/experimentation.

To better understand the reasons for this observed investment behavior, we explore whether financial factors, intangible intensity, compensation policy, managerial risk aversion, and ESG concerns enter the corporate objective. We find that financial conditions, including cash holdings and (net) book leverage, contribute to the corporate reward function. Intangible intensity (following Peters and Taylor (2017); Ward (2023); Chen et al. (2024)) significantly improves the model fit, but it does not resolve the corporate reward puzzle that the manager’s shadow price of capital is less than one. Accounting for managers’ and employees’ ownership stake and deferred compensation, consistent with dynamic contracting theories (DeMarzo and Fishman, 2007; Bolton et al., 2011; DeMarzo et al., 2012), improves model fit and ameliorates the corporate reward puzzle, while managerial traits and ESG have little incremental impact beyond their materiality potentially due to the difficulty of accurate measurement.

We construct an alignment measure, *Alignment*, defined as the correlation between the recovered value function—representing the reward that managers maximize—and observed market values across different economic states. Our findings reveal that managerial alignment with shareholder value is imperfect. In line with agency theory, which predicts that managers, absent effective governance mechanisms, pursue their own rewards rather than that of shareholders (i.e., $Alignment < 1$), we quantify this misalignment relative to both Tobin’s q and Total q . Although compensation policies including managerial ownership, bonuses, and deferred compensation significantly influence the recovered reward function, only employee deferred compensation yields improvements in our alignment measure, reaching 80% in the recent decade.

Although both the manager’s value and firm value increase with profitability, firm value exhibits an inverse U-shaped pattern in firm size whereas the recovered manager’s value is U-shaped and reaches its maximum in small, highly profitable firms. This divergence suggests unresolved misalignment issues, e.g., agency frictions due to illiquidity and distress, intangibles and E/S matters, that are not fully mitigated by compensation structures. Managers in smaller firms, operating under higher financial frictions, liquidity constraints, weaker governance and greater social engagement, likely enjoy greater discretion and are able to extract private benefits through compensation and control. In contrast, in mid-sized and large firms, stronger institutional oversight constrains managerial behavior despite high firm value.

From a stakeholder perspective, social and environmental performance should positively affect the reward function. Indeed, managers’ survey responses and shareholder communications suggest that, de jure, firms operate with stakeholder interests in mind (Graham, 2022; Rajan et al., 2024). When we

add social and environmental performance scores to our model, we observe a modest improvement in model fit and no improvement in alignment. Strong governance and social performance improve the manager’s period-by-period reward. In contrast, environmental performance has a negative impact. This outcome could imply either that environmental concerns have a significant material impact thus lowering the reward or that managers are less inclined to prioritize environmental performance, a finding that aligns with survey evidence showing that only about 20% of CFOs manage the company with the environment as a key priority.⁵ Even after incorporating compensation policies and ESG considerations the alignment measure remains imperfect, particularly in distressed low firm size states.

Literature review. The debate over corporate objectives is long-standing and centers on two primary perspectives: shareholder primacy and stakeholder theory. The shareholder primacy view—championed by scholars such as [Friedman \(1970\)](#)—holds that a corporation’s sole responsibility is to maximize shareholder wealth by generating a stream of discounted cash flows. Under competitive markets and in the absence of monopolistic distortions, this approach not only maximizes shareholder wealth but also promotes social welfare, as shareholders are the residual claimants to a firm’s profits ([Jensen and Meckling, 1976](#)). In contrast, stakeholder theory, as advanced by [Freeman \(1984\)](#), argues that firms should also consider the interests of employees, customers, suppliers, and communities. This broader perspective implies that a firm’s reward function may include non-pecuniary elements—such as environmental, social, and governance (ESG) concerns, corporate social responsibility (CSR), and sustainable development goals (SDGs)—alongside the profit motive ([Elhauge, 2005](#); [Graff Zivin and Small, 2005](#); [Baron, 2007](#); [Bénabou and Tirole, 2010](#); [Hart and Zingales, 2017, 2022](#)). The separation of ownership and control gives rise to the well-known principal-agent problem ([Jensen and Meckling, 1976](#)). Managers may pursue their own interests—driven by a different reward function—rather than those of shareholders, and the resultant (mis)alignment can be mitigated or amplified by financing frictions, intangibles, compensation structures, and ESG constraints. This misalignment results in what we term the “corporate reward puzzle,” where observed investment decisions deviate from what standard firm value maximization would predict because the shadow cost of capital is too low. Several alternative theories that we revisit have been proposed to explain these deviations.⁶

⁵Alternatively, the aggregate reward function we recover might reflect a mix of managers, some of whom value environmental performance while others do not ([Hart and Zingales, 2017](#)).

⁶The free cash-flow hypothesis ([Jensen, 1986](#); [Richardson, 2006](#)) posits that firms with excess internal funds and limited growth opportunities are prone to over-investment. Without effective external monitoring, managers invest

Our approach contributes to this rich literature by employing inverse reinforcement learning to empirically recover managers’ reward functions from observed behavior and debias their decisions. IRL, which is closely related to revealed preference theory (Samuelson, 1938), allows us to infer the underlying preferences that drive decision-making without imposing strong structural assumptions. In doing so, we not only account for the traditional financial frictions and agency conflicts but also incorporate the trade-off between exploration and exploitation—capturing how managers balance learning about uncertain payoffs with the drive to maximize immediate rewards. This novel application of IRL in corporate finance provides novel insights into the corporate reward puzzle and helps reconcile disparate findings in the literature.

More broadly, we demonstrate how machine learning techniques, such as IRL, can be used in economics and finance to analyze settings where agents make intertemporal decisions. Chen et al. (2023) show how to incorporate economic restrictions from structural models into a machine learning model through transfer learning. Campello et al. (2024) employ RL in corporate finance to learn managers’ policy functions, explaining and predicting firm outcomes and guiding policy recommendations. Complementing this innovative work, we establish conditions when RL can be used to learn the manager’s objective function. Our IRL approach infers managers’ underlying incentives and rewards based on their actions across various states, which allows to address incentive alignment and explore how optimal policies must adapt as the economic environment evolves. The advantage of our approach is that it is interpretable and generalizable and, hence, holds significant potential for applications in finance where learning about agents’ reward and value functions can provide economic insights without relying on specific parametric models.

Our paper also contributes to the burgeoning literature on the biases from ML/AI and the trade-off in projects that increase firm size or secure personal benefits rather than maximizing shareholder value. According to trade-off theory, firms balance the benefits of debt against the costs associated with financial distress. Underlevered firms often have excess cash, which can lead managers to over-invest, whereas over-levered firms may face risk-shifting problems and invest excessively in high-risk projects. Empirical studies have confirmed that capital structure dynamics significantly affect investment behavior (Tobin, 1969; Hayashi, 1982; Abel and Eberly, 1994). Recent research by Ward (2023) and Chen et al. (2024) highlights that the increasing role of intangible capital—such as intellectual property and brand value—can exacerbate agency conflicts. The opacity and measurement challenges associated with intangible assets often lead to higher investment levels, although deferred compensation mechanisms may help mitigate these conflicts. Moral hazard models underscore how asymmetric information between managers and shareholders can result in inefficient investment decisions (Myers and Majluf, 1984; Shleifer and Vishny, 1989; Stulz, 1990). Performance-based compensation is frequently proposed as a means to align managerial incentives with shareholder interests, although its effectiveness may vary depending on the design of these incentives. The dynamic contracting literature examines how long-term incentive constraints affect investment decisions over time (DeMarzo and Fishman, 2007; Bolton et al., 2011; DeMarzo et al., 2012). These models often predict underinvestment as managers face liquidity constraints and career concerns that limit their willingness to commit to risky projects. Finally, behavioral theories challenge the assumption of fully rational managers, showing that managerial biases such as overconfidence and miscalibration can lead to systematic over-investment as managers overestimate returns and underestimate risks (Malmendier and Tate, 2005; Ben-David et al., 2013).

offs between improved predictions and distortions from applying advanced data analytics. These biases stem from various sources, including the training data, algorithm design, and the broader socio-economic context in which these technologies are deployed.⁷ Fuster et al. (2022) investigate the adoption of ML/AI in finance and show that while advanced analytics enhance predictive accuracy, they may also exacerbate disparities. Similarly, our paper shows that a machine-run corporation trained on behavioral data will likely over-invest and under-explore investment opportunities, since the machine learns that managerial objectives diverge from shareholder interests. However, interpretable ML/AI, such as IRL, could also be used to correct these biases.

1 The Recovery Framework

One of the goals in economics and finance is to understand agents' objective function and explain their behavior based on the rewards they receive from their actions. In a corporate setting, corporate purpose captures the objective function of the firm when investment and financial decisions are delegated to a CEO or other manager. This section derives conditions for when and how IRL can be used to recover the reward function that corporate managers use in making their intertemporal investment and financing decisions. Most applied models of corporate decision-making posit a set of economic assumptions and infer the model parameters from data on, for instance, firms' investment and financing decisions, including profitability, size, investment, depreciation, leverage, and other features.⁸ We explore conditions for the recovery of an agent's reward and value functions, assuming optimality but no parametric restrictions.

1.1 Model setup

The agent's (a CEO or other manager) decision problem can be formulated as follows. Given a set of states $s \in \mathcal{S}$ and a set of potential actions $a \in \mathcal{A}$, agent $i = 1, \dots, I$ receives instantaneous reward $u_i(s, a)$ and has discount factor γ_i , which determines the importance of future rewards in the agent's intertemporal optimization. Let $T_i(s'|s, a)_{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ denote the state transition probability for agent i , that is, the probability of arriving in state s' when taking action a in state s which is allowed to vary by

⁷The [Financial Conduct Authority](#) highlights that biases often stem from historical data reflecting past discriminatory practices, which, if uncorrected, perpetuate existing inequalities. [ML/AI algorithms](#) have been shown to exhibit biases even when sensitive attributes are excluded from the input data. [ML/AI systems](#) can inadvertently reinforce systemic biases, emphasizing the need for alternative methods such as IRL to mitigate such effects.

⁸A similar logic applies to other fields of finance when trying to explain, e.g., consumer financial decisions.

agent. The agent must trade off exploitation (maximizing immediate rewards) and exploration (gathering information about the environment), which leads to stochastic policies $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ where $\mathcal{P}(\cdot)$ denotes the set of all probability distributions over all actions in $\mathcal{A}$ at every state.

To model this, we assume the agent solves an infinite-horizon Markov decision problem with instantaneous reward $u_i(s, a)$ and entropy regularization $H(\pi_i)$ which captures the manager's incentive for exploration. At each time t , the agent observes the current state s_t and takes action $a \in \mathcal{A}$ with probability $\pi_i(a|s_t)$:

$$V_i^{\pi^*}(s) = \sup_{\pi} \mathbb{E}_i \left[\sum_{t=0}^{\infty} (\gamma_i^t (u_i(s_t, a_t) + \lambda H(\pi_i(\cdot|s_t)))) \right], \quad (1)$$

where $V_i^{\pi^*}(s)$ is the value function under optimal policy π^* and initial state $s_0 = s$, $\mathbb{E}_i$ denotes the expectation over trajectories $\{(s_t, a_t)\}_{t \geq 0}$ under $T_i(s_{t+1}|s_t, a_t)$ and following policy $\pi_i(\cdot|s_t)$ and $H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a) \log \pi(a)$ is the entropy of π . The parameter $\lambda > 0$ controls the trade-off between exploration (high entropy, more diverse action choices) and exploitation (low entropy, more deterministic action choices) and, hence, the dispersion in π which ensures that all actions occur with non-zero probability in each state. The assumption of a shared common reward across the agents ensures that the instantaneous rewards $u_i(s, a)$ can be recovered from data on the agents' behavior.

Assumption 1. *Agents $i = 1, \dots, I$ share the same instantaneous reward up to an additively separable preference shock ϵ_{it} : $u_i(s_t, a_t) = u(s_t, a_t) + \epsilon_{it}$, with $\mathbb{E}[\epsilon_{it}|T_j(\cdot|s_t, a_t)] = 0$ for all (i, j) .*

Here we allow each agent to receive different instantaneous rewards from the same actions, while [Cao et al. \(2021\)](#) and [Rolland et al. \(2022\)](#) assume $\epsilon_{it} = 0$ which may not hold in financial data. The solution to (1) is an entropy-regularized optimal policy π_i^* that satisfies for all $a \in \mathcal{A}$:

$$\pi_i^*(a|s) = \frac{e^{\frac{1}{\lambda} Q_i^{\pi_i^*}(s, a)}}{\sum_{a' \in \mathcal{A}} e^{\frac{1}{\lambda} Q_i^{\pi_i^*}(s, a')}}, \quad (2)$$

where the Q_i^{π} -function is the agent-specific state-action value of π at $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$Q_i^{\pi}(s, a) \equiv u_i(s, a) + \gamma_i \sum_{a' \in \mathcal{A}} T_i(s'|s, a) V_i^{\pi}(s'), \quad (3)$$

and $V_i^{\pi}(s')$ is the continuation value given π . The entropy regularization $\lambda > 0$ can be arbitrarily small and ensures that the agent's policies are stochastic and ergodic, such that $\pi_i^*(a|s) \in (0, 1)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

The policy function (2) generates for each agent a set of observed state-action trajectories $\{s_0, a_0, s_1, a_1, s_2, a_2, \dots, s_\infty, a_\infty\}$ starting from initial state $s_0 = s$ that allow to estimate $\pi_i = \pi_i^*(a|s)$ and $T_i = T_i(s'|s, a)$ for any i and (s, a) . We now show that recovery of u_i from these trajectories is impossible so long as $I = 1$.

1.2 (Im)possibility of reward recovery

For any i with fixed policy $\pi_i(a|s) \in (0, 1)$ and an arbitrary choice of function $v_i : \mathcal{S} \rightarrow \mathbb{R}$, Cao et al. (2021) show there is a unique corresponding reward function $u_i(s, a)$ satisfying

$$u_i(s, a) = \lambda \ln \pi_i(a|s) - \gamma_i \sum_{s' \in \mathcal{S}} T_i(s'|s, a) v_i(s') + v_i(s). \quad (4)$$

Condition (4) with $\pi_i^*(a|s)$ yields agent i 's reward function iff $v_i(s) = V_i^{\pi^*}(s)$ for all $s \in \mathcal{S}$.

The problem of reward recovery using state-action data is to find the function $u_i(s, a)$ that assigns a reward to each state $s \in \mathcal{S}$ and potential action $a \in \mathcal{A}$ and maximizes the log-likelihood of observing the agent's behavior, subject to the constraint that the agent's policies are optimal with respect to $u_i(s, a)$, that is, (2) holds.⁹ An assumption we will use when recovering the agents' reward is that λ is the same across i . Even then, as already observed in Russell (1998) the recovery problem using state-action data is fundamentally ill-posed in the special case $I = 1$. For instance, all actions are optimal if we assume $u_i(s, a)$ is a constant. More generally, Ng et al. (1999) show that, for any reward transformation $h : \mathcal{S} \rightarrow \mathbb{R}$, called "shaping potential," the reward $\tilde{u}_i = u_i + \gamma h(S_{t+1}) - h(S_t)$ yields the same optimal policies for every unregularized Markov decision problem.

Theorem 1 (Impossibility). *In the special case $I = 1$, the agents' instantaneous reward functions $u_i(s, a)$ and their value functions $V_i^{\pi^*}(s)$, $i = 1, \dots, I$, cannot be recovered using state-action data.*

Figure 1 in Panel (a) illustrates Theorem 1, though we defer the discussion to Section 1.4. In the remainder, we require $I > 1$. Once the individual reward function $u_i(s, a)$ is found, it can be used to back out the agent's value function $V_i^{\pi^*}(s)$. Alternatively, condition (4) shows we can also proceed in reverse order, backing out u_i from $V_i^{\pi^*}$. The recovered reward function $u(s, a)$, that is, the component

⁹The optimization problem can be formulated as follows (Rolland et al., 2022): $\max \ln \Pr(\cdot|u)$, the log-likelihood of observing the agent's behavior given the reward function u subject to: $\Pr(s_{t+1}, a_{t+1}|s_t, a_t, u) = \arg \max_{a'} \sum_{s'} \Pr(s_{t+1}|s_t, a_t) u(s_t)$, which represents the expert's policy where $\Pr(s_{t+1}|s_t, a_t)$ is the transition function which defines the probability of transitioning from state s_t to state s_{t+1} when action a_t is taken, and $\Pr(s_{t+1}, a_{t+1}|s_t, a_t, u)$ is the policy, which defines the probability of taking action a_{t+1} in state s_{t+1} given u .

shared among all experts can then be used to train (using reinforcement learning) agents to perform the same task(s) in a new environment. These new agents learn their own optimal policy by using the reward function u and the transition function $T(s'|s, a)$ to determine the optimal actions to take in each state.

1.3 Strong reward recovery

In an environment with $I > 1$ laid out in Section 1.1, we aim to recover the instantaneous reward functions $u_i(s, a)$ that best explain the agents' behavior. It will be convenient to work with the log probability of action a by agent i in state s , $\ln \pi_i^*(a|s)$, and collect them for each agent and state-action pair in the vector

$$y_i = (\ln \pi_i^*(a|s))_{|\mathcal{S}||\mathcal{A}|\times 1}. \quad (5)$$

The transition probabilities $T_i = T_i(s'|s, a)$ in (3) are endogenously determined by π_i^* and known given (s, a) which allows to define the matrix

$$X_i = -(I - \gamma_i T_i)_{|\mathcal{S}||\mathcal{A}|\times |\mathcal{S}|}. \quad (6)$$

We define the residual makers for projections on the space spanned by X_i as $\mathbf{M}_i = I - X_i(X_i^\top X_i)^{-1}X_i^\top$.

Theorem 2 (Strong Recovery). *The agents' instantaneous reward functions $u_i(s, a)$ and their value functions $V_i^{\pi^*}(s)$, $i = 1, \dots, I$, can be recovered up to affine transformations if and only if $\text{rank}([X_1 X_2 \dots X_I]) = I|\mathcal{S}| - 1$.*

The corporate reward recovery Theorem 2 is weaker but more widely applicable than Theorem 3 in Rolland et al. (2022) as it does not require observing λ . The optimality conditions (2) and (4) show that the agent's reward is the unexplained part in a projection of the log choice probabilities on the transition probabilities. Using the definitions (5) and (6) and rescaling v_i and u_i by λ , the optimality conditions allow to decompose the log-probabilities y_i into the projection on (a transformation of) the transition probabilities and residuals that capture the agent's reward: $y_i = X_i v_i + u_i$. For any two agents i and j ,

$$y_i - y_j = \begin{pmatrix} X_i & -X_j \end{pmatrix} \cdot \begin{pmatrix} v_i \\ v_j \end{pmatrix} + \varepsilon_{ij}, \quad (7)$$

with iid noise ε_{ij} is independent of the instantaneous reward u .

The following proposition provides closed-form expressions for recovering $V_i^{\pi^*}(s)$ and $u_i(s, a)$.

Proposition 1. *Let $X = (X_i - X_j) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 2|\mathcal{S}|}$ for any $i \neq j$ with $\text{rank}(X) = 2|\mathcal{S}| - 1$ be the stacked transition probabilities, $y = y_i - y_j \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 1}$ be the difference in log probabilities of actions taken across agents, and state values $v_i(s) \equiv V_i^{\pi^*}(s)$. Then:*

(i) *Up to affine transformations represented by constants $(a_v, b_v) \in \mathbb{R}$, the value functions can be recovered by*

$$v_i = a_v + b_v (X_i^\top \mathbf{M}_j X_i)^{-1} X_i^\top \mathbf{M}_j y. \quad (8)$$

These expressions correspond to weighted or generalized least squares.

(ii) *The agents' individual reward functions can be recovered for all (s, a) from (10) and any i by $u_i = y_i - X_i \hat{v}_i$. Finally, the shared common reward can be recovered for all (s, a) by $u = \lim_{I \rightarrow \infty} \frac{1}{I} \sum_{i=1}^I u_i$.*

The proposition shows that the agents' reward function is recovered for all (s, a) and any i up to an affine transformation represented by constants $(a_u, b_u) \in \mathbb{R}$ from

$$u_i(s, a) = a_u + b_u (\ln \pi_i^*(a|s) - \gamma_i \sum_{s' \in \mathcal{S}} T_i(s'|s, a) v_i(s') + v_i(s)), \quad (9)$$

and $u(s, a)$ can be approximated by $\frac{1}{I} \sum_i u_i(s, a)$. Expression (9) shows that the observed log probabilities $\ln \pi_i^*(a|s)$ and the observed transition probabilities $T_i(s'|s, a)$ can be used to back out the agents' reward $u(s, a)$ state-by-state and action-by-action once the value functions v_i have been recovered from (10). With the stacked vector $v = (v_i^\top \ v_j^\top)^\top \in \mathbb{R}^{2|\mathcal{S}| \times 1}$ and constants $(a_v, b_v) \in \mathbb{R}^2$, (8) yields

$$v = a_v \mathbf{1} + b_v (X^\top X)^{-1} X^\top y. \quad (10)$$

The intuition for the proposition is simple. The value functions are recovered using weighted least squares. The individual reward functions u_i are then recovered based on the estimated value functions and observed log probabilities $\ln \pi_i^*(a|s)$ and transition probabilities $T_i(s'|s, a)$. This recovery process determines the reward functions for each agent and each state-action pair. Additionally, the shared common reward $u(s, a)$ can be recovered by averaging the individual reward functions across all agents.

1.4 Weak reward recovery

We now make the additional assumption that the reward function u is linear in features $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, as this makes the model more amenable to use with corporate financial data.

Assumption 2. For $\theta \in \mathbb{R}^d$, $u(s, a) = f(s, a) \cdot \theta$.

We introduce the following notation. With X_i from (6), define for agents $(i, j) \in \{1, \dots, I\}^2$ the matrix

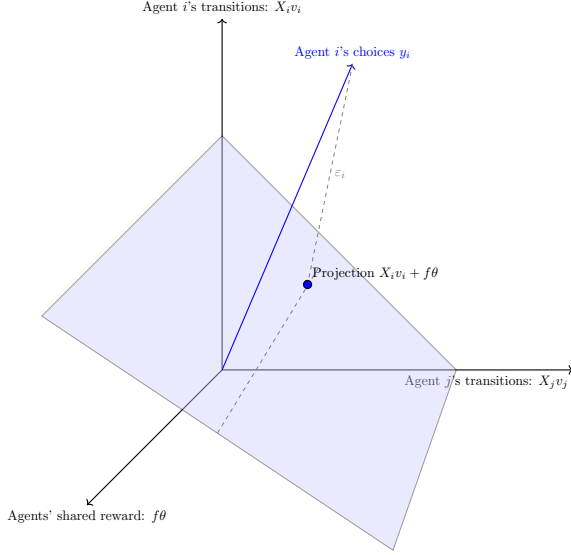
$$X = \begin{pmatrix} X_i & -X_j & 0 \\ X_i & 0 & f \end{pmatrix}. \quad (11)$$

With this definition at hand, we obtain the following recovery result:

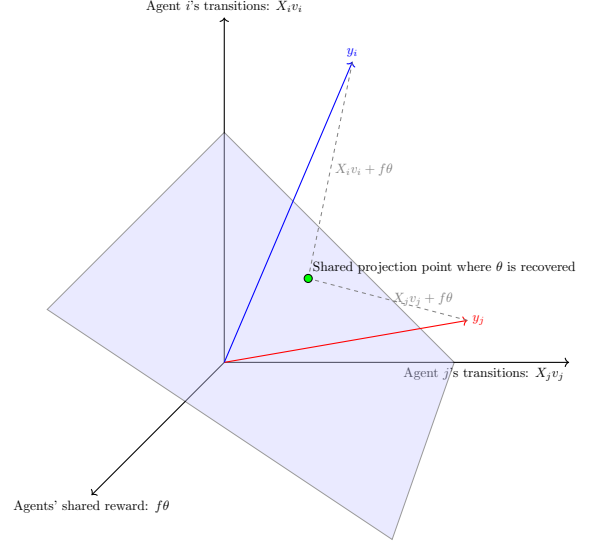
Theorem 3 (Weak Recovery). *Under Assumption 2, the agents' instantaneous reward functions $u_i(s, a)$ and their value functions $V_i^{\pi^*}(s)$, $i = 1, \dots, I$, can be recovered up to affine transformations if and only if $\text{rank}(X) = 2|\mathcal{S}| + d$.*

The proof is provided in [Appendix A](#). The appendix also provides analytic expressions that we will use in our applied framework. Figure 1 in Panel (b) provides a visualization of Theorem 3. To build intuition how θ is recovered, note that the recovery process involves projecting the observed outcomes y_i and y_j , i.e., the log choice probabilities, the respective transitions X_i and X_j , and the shared features f onto the parameter space where v_i , v_j , and θ reside. We can visualize the recovery of θ as finding the point where the two projections cross, that of the vectors y_i and y_j , respectively, onto a subspace defined by X_i , X_j , and f . In this geometric representation, we are looking for the point in this 3D parameter space where the projection of the log choice probabilities aligns with the components.

Imagine the outcomes y_i (blue arrow) and y_j (red arrow) are points in the shared 3D space. The value functions v_i and v_j for agents i and j live in their respective vector spaces. These spaces are defined by the matrices X_i and X_j , which transform the agent-specific transition probability into outcomes. The reward features f interact with both agents and impact the outcomes in a similar way through θ . In Figure 1, the axes represent the projection of the outcome space into the parameter space for v_i , v_j , and $f\theta$. The vectors $X_i v_i$, $X_j v_j$, and $f\theta$ are projected from these points. The goal is to align these projections in such a way that we recover θ , which can be seen as the intersection of the outcome vectors with the shared feature space (blue area). The dashed lines show the projection of y_i and y_j onto the shared point which



(a) Impossibility of reward recovery.



(b) Reward recovery.

Figure 1: Illustration of reward recovery. The figure illustrates the reward recovery as the task of finding the shared projection point, which represents the same combination of $X_i v_i + f\theta$ and $X_j v_j + f\theta$. The blue area depicts the shared projection plane where both agents' choices, y_i and y_j , are projected. This plane represents the span of the agent-specific contributions depending on the shared feature $f\theta$. The left plot illustrates Theorem 1, the right plot reward recovery as the shared projection point where θ is recovered.

represents the same combination of $X_i v_i + f\theta$ and $X_j v_j + f\theta$. This point is the target projection where we recover θ , showing how it aligns with the features and observed outcomes.

1.5 Alignment measure

The methodology recovers the value that managers maximize in practice. Finance theory prescribes that managers should maximize shareholder value. Absent non-pecuniary benefits, shareholder value maximization boils down to shareholder wealth maximization, i.e., maximizing the market value of equity. Absent bondholder expropriation, shareholder wealth maximization amounts to firm value maximization, i.e., maximizing the market value of equity and debt. Scaling everything by capital, this implies maximizing Tobin's q or Total q , i.e., market value of equity and debt divided by book value of total assets.

Theorem 3 and Proposition 2 show that we recover agent i 's true value only up to affine transformations ($v_i = a_v + b_v \hat{v}_i$, where a_v and b_v are unknown). Hence, we cannot directly compare $\hat{v}_i$ and q_i , that is, Tobin's q or Total q which measure the shareholders' valuation of the firm. Instead, we construct a measure of alignment based on the correlation across different states between the recovered value that managers

maximize and observed market values:¹⁰

$$Alignment_i = Corr(\{\hat{v}_i(s)\}_{s \in \mathcal{S}}, \{h(q_i(s))\}_{s \in \mathcal{S}}) \in [-1, 1], \quad (12)$$

for $i = 1, \dots, I$ where $h(\cdot)$ is a statistic aggregating q across observations in state s . In our empirical implementation, we use the median of Total q . If the two are perfectly aligned, the correlation and, hence, the alignment measure will be 1. If the two are perfectly misaligned and the manager maximizes own reward at the expense of shareholders, the correlation and, hence, the alignment measure will be -1.

2 A Sandbox for Recovering Corporate Purpose

We briefly review the neoclassical investment model in which the corporate purpose is shareholder value maximization. Using this model, we show how to map such a model into the IRL framework described in Section 1 and apply our (weak) recovery theorem in this controlled environment.

2.1 A benchmark model for recovery: Neoclassical investment with experimentation

We extend the basic investment model by [Strebulaev and Whited \(2012\)](#) in discrete time with an infinite horizon ($t = 0, 1, \dots, \infty$) to stochastic investment policies. Let $k = k_t$ denote the level of capital this time period and $k' = k_{t+1}$ be the level of capital next period, which evolves according to $k' = (1 - \delta)k + I$, where δ denotes the depreciation rate and I denotes investment in capital. For simplicity, the price of capital is normalized to one. Let $z = z_t$ denote a persistent stochastic productivity shock which evolves according to $\ln(z') = \rho \ln(z) + \sigma \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

Each period, a risk-neutral manager, acting on behalf of shareholders, observes the state $s = (k, z)$ and chooses as action next period's level of capital $a = k'$ with probability $\pi(a|s)$ in order to maximize the value of the firm, which is given by the expected present value of a stream of future cash flows to shareholders. In choosing π , the manager trades off exploitation with exploration. Let $u(s, a)$ denote the cash flow function that depends on state (current capital, productivity shock) and action (next period's

¹⁰While we can recover v_i only up to affine transformations, this does not affect the correlation $Corr(q_i, v_i) = Corr(q_i, a_v + b_v \hat{v}_i) = \text{sign}(b_v)Corr(q_i, \hat{v}_i)$ so long as $b_v > 0$.

capital). The manager maximizes firm value

$$V_\lambda^*(s) = \sup_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t (u(s_t, a_t) - \lambda \ln(\pi(a_t|s_t))) \right], \quad (13)$$

subject to the laws of motion for k' and $\ln(z')$ and given the regularization coefficient $\lambda > 0$. The policy function $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ is stochastic, where $\mathcal{P}(\cdot)$ denotes the set of all probability distributions, as opposed to deterministic $\pi : \mathcal{S} \rightarrow \mathcal{A}$ in the benchmark model in [Strebulaev and Whited \(2012\)](#). As $\lambda \rightarrow 0$, the entropy term in (13) vanishes and the model degenerates to the benchmark model. From the assumption of risk-neutrality, $\gamma = \frac{1}{1+r}$ with risk-free interest rate r .

In the deterministic case ($\lambda = 0$), solving this problem consists in finding the (optimal) policy function $\pi(k, z)$ that maps (capital, profitability)-pairs (k, z) into choices of next period's capital k' . Under the optimal policy function π^* , the value function $V_\lambda^*(s)$ is maximized. Except for specific cases, the pure exploitation model has no analytical solution.

Cash flows and the reward function. We dedicate special attention to the cash flow function, as it is the model component that we want to recover. In [Strebulaev and Whited \(2012\)](#), cash flows are given by

$$u(k, z, k') = \phi(k, z) - I(k, k') - \psi(I, k'), \quad (14)$$

where $\phi(k, z) = zk^\zeta$ is the firm's profit with parameter $\zeta \in (0, 1)$ governing its concavity, $I(k, k') = k' - (1 - \delta)k$ is gross investment, and ψ is an adjustment cost function. Many models share this basic structure. We consider two different cases for adjustment costs that are common in the literature. In the first case, we abstract away from adjustment costs by setting them to zero. As a result of this assumption, even the deterministic model has an analytical solution that can serve as a benchmark. In the second case, we consider convex and fixed adjustment costs. This case has no analytical solution but is more realistic. Consequently,

$$\psi(I, k) = \begin{cases} 0 & , \text{ if no adjustment costs,} \\ \frac{\psi_0}{2} \frac{I^2}{k} + \psi_1 k \mathbb{1}_{\{I \neq 0\}} & , \text{ if adjustment costs,} \end{cases} \quad (15)$$

where $\psi_0, \psi_1 \geq 0$ are parameters that govern how costly is investment and $\mathbb{1}_{\{I \neq 0\}}$ is an indicator function equal to 1 (0) if investment is not equal (equal) to zero. Combining (14) and (15) yields an expression for

the cash flow function

$$u(k, z, k') = \begin{cases} zk^\zeta + (1 - \delta)k - k' & , \text{ if no adjustment costs} \\ zk^\zeta + (1 - \delta)k - k' - \frac{\psi_0}{2} \frac{I^2}{k} - \psi_1 k \mathbb{1}_{\{I \neq 0\}} & , \text{ if adjustment costs} \end{cases} \quad (16)$$

In state-action notation, we can write state $s = (s_1, s_2)$ with state variables $s_1 = k, s_2 = z$ and action $a = k' (= s'_1)$.¹¹ Expressing the cash flow function (16) in terms of states and actions yields the reward function

$$u(s, a) = \begin{cases} \underbrace{s_2 s_1^\zeta}_{\text{Profits}} - \underbrace{\delta s_1}_{\text{Depreciation}} - \underbrace{(a - s_1)}_{\text{Net investment}} & , \text{ if no adjustment costs,} \\ \underbrace{s_2 s_1^\zeta}_{\text{Profits}} - \underbrace{\delta s_1}_{\text{Depreciation}} - \underbrace{(a - s_1)}_{\text{Net investment}} - \underbrace{\frac{\psi_0}{2} \frac{(a - (1 - \delta)s_1)^2}{s_1}}_{\text{Convex Adj. Costs}} - \underbrace{\psi_1 s_1 \mathbb{1}_{\{a - (1 - \delta)s_1 \neq 0\}}}_{\text{Fixed Adj. Costs}} & , \text{ otherwise.} \end{cases} \quad (17)$$

One can see from (17) that the reward function, while non-linear in the state and action variables, is linear in three or five features depending on the case.

Model solution. We solve the model(s) using Q-value Iteration, and by Value Function Iteration in the benchmark model without entropy-regularization. The standard value function iteration algorithm tries to find the optimal deterministic policy function $\pi(s)$ that picks for each state s the optimal action a and maximizes the state value function $V(s)$. The Q-value function iteration algorithm tries to find the optimal stochastic policy function $\tilde{\pi}(s)$ that picks for each state s the optimal (non-zero) probability of picking a given action $P(a|s)$ and maximizes the (state, action)-pair value function $Q(s, a)$. Once we have found the optimal $Q^*(s, a)$ iteratively, we can recover the optimal value $V^*(s)$ and policy $\pi^*(a|s)$ as follows. We have that for any $s \in \mathcal{S}$,

$$V_\lambda^*(s) = \lambda \ln \sum_{a \in A} e^{\frac{1}{\lambda} Q_\lambda^*(s, a)}. \quad (18)$$

¹¹In this section, we limit ourselves to the analysis of problems with only one action variable. However, it is possible to extend the analysis to cases where the action consists in choosing combinations of multiple action variables $a = (a_1, a_2, \dots)$ which we do in our empirical analysis.

and the maximum is achieved by the randomized policy

$$\pi_{\lambda}^*(a|s) = \frac{e^{\frac{1}{\lambda}Q_{\lambda}^*(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\frac{1}{\lambda}Q_{\lambda}^*(s,a')}} \text{ for } a \in \mathcal{A}. \quad (19)$$

2.2 Guidance on feature selection

In line with Assumption 2, we express the reward function (17) as a linear combination of features $f(s, a)$ and weights θ , such that $u(s, a) = f(s, a) \cdot \theta$. Four key considerations guide the specification of features:

1. To avoid problems related to the (non-)invertibility of the X matrix, we avoid including features that are constant across all (state, action)-pairs, i.e., $f(s, a) \neq c \forall s, a$ where c is a constant. Moreover, we avoid including features that are constant for almost all (state, action)-pairs.
2. To avoid problems related to multicollinearity, we avoid including features that are highly correlated with other features.
3. To keep the feature set parsimonious, we avoid adding irrelevant features (i.e., we avoid kitchen sink regressions). However, adding them should not materially affect the reward recovery, provided that they are not overly correlated with other features.
4. There is some degree of freedom when it comes to defining features and weights. In particular, features may be decomposed into linear combinations of other features, provided that they do not conflict with the previous points.

Table 1, Panel A summarizes our feature selection. To illustrate the feature selection, we consider several specifications of features $f(s, a)$ and weights θ . Specification 1 corresponds to the benchmark model without adjustment costs. The first feature is profit, $s_2 s_1^{\zeta}$. The second feature is net investment, $a - (1 - \delta)s_1$. Specification 2 is the same as Specification 1, except that we have split up the second feature into two features, capital today (s_1) and next period (a). Specification 3 corresponds to the benchmark model with adjustment costs. Specifications 4 (5) corresponds to the benchmark model without (with) adjustment costs where we scale all features by capital s_1 . Consequently, $s_2 s_1^{\zeta-1}$ denotes profitability and $\frac{a - (1 - \delta)s_1}{s_1}$ denotes the investment rate.

Table 1: Reward recovery in controlled environment.

Panel A contains different specifications of features and coefficients in our controlled environment. Panel B documents the recovered weights for different feature specifications and experiments in our controlled environment. Column 1 refers to the specification of features and true weights used. Column 2 refers to the recovery experiment. π and T refer to whether we used the true or estimated policy and transition functions. We used the following parameter values: $\zeta = 0.55, \delta = 0.2, \rho = 0.7, r = 0.05, \sigma_\epsilon = 0.04, \psi_0 = 0.01, \psi_1 = 0.01, n_k = 15, d = 3, n_z = 5, m = 2, \lambda = 0.2, \sigma_\epsilon^1 = 0.035, \sigma_\epsilon^2 = 0.045, \gamma^1 = 0.94, \gamma^2 = 0.96, n_{steps} = 1,000,000, \text{seed} = 0$.

Panel A: Reward specifications in controlled environment					
Specification	No. of features	Features $f(s = (s_1, s_2), a = s'_1)$			True Weights (θ)
1	2	$\begin{bmatrix} s_2 s_1^\zeta & a - (1 - \delta)s_1 \end{bmatrix}$			$\begin{bmatrix} 1 & -1 \end{bmatrix}$
2	3	$\begin{bmatrix} s_2 s_1^\zeta & s_1 & a \end{bmatrix}$			$\begin{bmatrix} 1 & 1 - \delta & -1 \end{bmatrix}$
3	4	$\begin{bmatrix} s_2 s_1^\zeta & a - (1 - \delta)s_1 & \frac{(a - (1 - \delta)s_1)^2}{s_1} & s_1 \mathbb{1}_{\{a - (1 - \delta)s_1 \neq 0\}} \end{bmatrix}$			$\begin{bmatrix} 1 & -1 & -\frac{\psi_0}{2} & -\psi_1 \end{bmatrix}$
4	2	$\begin{bmatrix} s_2 s_1^{\zeta-1} & \frac{a - (1 - \delta)s_1}{s_1} \end{bmatrix}$			$\begin{bmatrix} 1 & -1 \end{bmatrix}$
5	4	$\begin{bmatrix} s_2 s_1^{\zeta-1} & \frac{a - (1 - \delta)s_1}{s_1} & \left(\frac{a - (1 - \delta)s_1}{s_1}\right)^2 & \mathbb{1}_{\{a - (1 - \delta)s_1 \neq 0\}} \end{bmatrix}$			$\begin{bmatrix} 1 & -1 & -\frac{\psi_0}{2} & -\psi_1 \end{bmatrix}$
Panel B: Reward recovery in controlled environment					
Specification	Experiment	π	T	True Weights (θ)	Recovered Weights ($\hat{\theta}$)
1	1	True	True	$\{1, -1\}$	$\{1.0, -1.0\}$
1	2	Estimated	True	$\{1, -1\}$	$\{1.005, -1.005\}$
1	3	True	Estimated	$\{1, -1\}$	$\{0.777, -0.796\}$
1	4	Estimated	Estimated	$\{1, -1\}$	$\{0.781, -0.801\}$
2	1	True	True	$\{1, 0.8, -1\}$	$\{1.0, 0.8, -1.0\}$
2	2	Estimated	True	$\{1, 0.8, -1\}$	$\{0.995, 0.786, -0.985\}$
2	3	True	Estimated	$\{1, 0.8, -1\}$	$\{0.626, 0.891, -1.0\}$
2	4	Estimated	Estimated	$\{1, 0.8, -1\}$	$\{0.619, 0.878, -0.986\}$
3	1	True	True	$\{1, -1, -0.005, -0.01\}$	$\{1.0, -1.0, -0.005, -0.01\}$
3	2	Estimated	True	$\{1, -1, -0.005, -0.01\}$	$\{1.011, -1.012, -0.005, -0.01\}$
3	3	True	Estimated	$\{1, -1, -0.005, -0.01\}$	$\{0.724, -0.745, -0.007, -0.01\}$
3	4	Estimated	Estimated	$\{1, -1, -0.005, -0.01\}$	$\{0.744, -0.766, -0.007, -0.009\}$
4	1	True	True	$\{1, -1\}$	$\{1.0, -1.0\}$
4	2	Estimated	True	$\{1, -1\}$	$\{1.001, -0.996\}$
4	3	True	Estimated	$\{1, -1\}$	$\{0.986, -1.0\}$
4	4	Estimated	Estimated	$\{1, -1\}$	$\{0.993, -0.995\}$
5	1	True	True	$\{1, -1, -0.005, -0.01\}$	$\{1.0, -1.0, -0.005, -0.01\}$
5	2	Estimated	True	$\{1, -1, -0.005, -0.01\}$	$\{1.006, -1.01, -0.002, -0.013\}$
5	3	True	Estimated	$\{1, -1, -0.005, -0.01\}$	$\{0.976, -1.021, 0.0, -0.01\}$
5	4	Estimated	Estimated	$\{1, -1, -0.005, -0.01\}$	$\{0.981, -1.03, 0.004, -0.013\}$

2.3 Recovering the feature weights

We now illustrate our recovery theorem by implementing it in this controlled environment. [Appendix B](#) explains in detail how we implement the weak recovery theorem in this controlled environment. We perform several experiments for each specification in which we either use the true model-implied policy π and true transition probability matrix T , or we estimate from our simulated data one or the other, or both. We expect the best recovery results in the true π , true T experiment and the worst in the estimated π , estimated T experiment.

Our results are summarized in Panel B of Table 1. In Experiment 1, we are able to recover the true coefficients perfectly across all specifications. This result is expected given Proposition 2 and reassures us that the methodology works. In Experiment 2, we are able to recover coefficients that are very close to the true ones. This result indicates that the estimation of the policy function does not materially

affect the recovery of coefficients. In Experiment 3, the coefficients on the first two features are biased for specifications 1, 2 and 3, but are close to the true ones for specifications 4 and 5. We hypothesize that the biased results for the first two features in specification 1 to 3 are due to a combination of (1) insufficient precision in the estimation of the transition functions and (2) the high correlation between capital and profits, which are increasing in capital. Scaling the features by capital, as we do in specifications 4 and 5, solves this problem. The results of Experiment 4 mirror those of Experiment 3, indicating that estimating both policy functions and transition functions does not constitute a problem, provided that the feature definitions are adequate.¹²

In Table B.1 in Appendix B we run additional experiments to quantify the biases from irrelevant and missing features. The results suggest that neither omitted variable bias nor spurious correlation are material concerns due to the non-linear nature of the recovery. In summary, the approach laid out in this section provides a framework for how feature weights can be estimated and reward functions can be recovered in complex intertemporal corporate decision problems. We find that when we simulate realistic sample sizes, the correct feature weights can be recovered with reasonable accuracy. However, noise in π and T matter in that they deteriorate the performance. The feature selection matters although omitted and spurious features do not materially affect the estimates.

3 Data and Empirical Methodology

This section discusses the different datasets and the variable constructions we use to recover the corporate objective of publicly listed U.S. firms.

3.1 Data sources and sample

We download annual firm fundamentals data for all U.S. firms in the Compustat universe. To this, we add CPI data from the U.S. Bureau of Economic Analysis (BEA) and interest rate data from the Federal Reserve Economic Data (FRED). Following Peters and Taylor (2017), we add data on intangible capital and total q downloaded from WRDS. Their measures rely, in part, on the marginal tax rates introduced in Graham (1996), which we download from the author’s website. Coverage starts in 1975, as this is the first year that R&D is reported. We supplement the intangible capital data with data provided by Bena

¹²In unreported results, we confirm that the recovered coefficients converge to the true coefficients as $n_{steps} \rightarrow \infty$, as one would expect.

and Simintzi (2023) on process intangibles. In addition, we collect data on executive compensation from Execucomp and Equilar, which are available from 1992 onward. Finally, we collect data on greenhouse gas emissions, GHG, from Trucost available from 1999 to 2021 as well as data on environmental, social and governance, ESG, scores from MSCI, which are available from 2007 onward.¹³

3.2 Variable definitions

We follow standard screening procedures in the literature. In particular, we remove firms whose SIC code is between 4900 and 4999 (utilities), between 6000 and 6999 (financials), or above 9000 (government agencies). We also remove observations with missing or weakly negative book value of assets (`at` ≤ 0), sales (`sale` ≤ 0) and/or number of common shares outstanding (`csho` ≤ 0). We additionally remove observations with physical capital below \$1 million (`ppent` < 1 or `ppegt` < 1). We impute missing values of amortization of intangibles (`am`), cash and cash equivalents (`che`), short-term debt (`d1c`), long-term debt (`d1tt`), R&D expense (`xrd`) and in-process R&D (`rdip`) by a zero. We impute missing values of deferred taxes and investment tax credit (`txditc`) by deferred taxes (`txdb`) if available or by a zero otherwise.

Table C.1 in Appendix C contains the definitions for all of our variables. We winsorize all variables at the 1% and 99% level by year to mitigate the impact of outliers. Our approach requires beginning of year (BoY) and end of year state variables, as well as action variables, to be observed. Consequently, we remove observations for which either of the following variables are missing: beginning-of-period total capital, end-of-period total capital, beginning-of-period alternative cash flow return on assets, end-of-period alternative cash flow return on assets or the total investment rate.

Table C.2 in Appendix C reports summary statistics for all variables. We provide descriptive statistics across six distinct panels and datasets: Compustat (1975-2021), Peters & Taylor (1975-2021), Bena & Simintzi (2000-2020), ExecuComp and Equilar (1992-2021), Trucost (1999-2021), and MSCI (2007-2021). The table reveals that all variables are in line with the literature in terms of magnitude and variation. Below we highlight some of the key variables of interest in our empirical analysis.

Intangible capital. We account for the the different facets of intangible capital by measuring intangible intensity, knowledge intensity, organizational intensity, and process intensity. Intangible intensity quantifies the proportion of a firm’s resources allocated to intangible assets, such as intellectual property, brand recognition, and proprietary technologies. A higher intangible intensity, given the opaque and spe-

¹³We note that coverage of ESG scores increases dramatically in 2012; see Figure 2 in Pástor et al. (2022)

cialized nature of these assets, can increase managers’ rewards and exacerbate agency conflicts between managers and investors. Knowledge intensity refers to the extent to which a firm’s operations rely on specialized knowledge and expertise. Firms with high knowledge intensity invest significantly in R&D. Such firms may face unique challenges in managing and protecting their intellectual capital, as the specialized nature of their assets can lead to increased agency conflicts. Organizational intensity measures the degree to which a firm’s value is derived from its internal organizational structure, processes, and culture. Managing such firms requires careful alignment of internal structures with innovation strategies to mitigate agency conflicts. Process intensity refers to the extent to which a firm’s operations rely on process-related intangibles, such as efficient workflows, proprietary processes, and operational innovations. Higher process intensity may detract from immediate rewards, possibly reflecting the short-term costs or uncertainties associated with improving complex internal processes.

Deferred compensation. We measure two distinct and complementary dimensions of deferred compensation. First, we measure deferred compensation for all employees of the firms.¹⁴ Second, we measure deferred compensation for the top five highest paid executives. These two measures allow us to assess in a comprehensive way how deferred compensation affects investment decisions and aligns interests.¹⁵

Behavioral factors. We account for behavioral factors by accounting for CEO characteristics. More specifically, we focus on CEO overconfidence. An overconfident CEO is one who holds onto stock options until the very last moment before they expire. Typically, a rational CEO would exercise options earlier to mitigate exposure to firm’s idiosyncratic risk and secure gains. However, if a CEO consistently delays exercising options until their final year, it reflects overconfidence: the belief that the stock price will continue to rise. Additionally, we use an alternative and more stringent measure of CEO overconfidence. A CEO is classified as overconfident if the CEO fails to exercise options with five years remaining before expiration, even when these options are already 67% in the money. This behavior suggests a strong belief that the stock will rise even further, despite having a safe and significant gain readily available.¹⁶

¹⁴Full coverage on employee deferred compensation starts in 2004.

¹⁵Ward (2023) relies on employee deferred compensation to examine the role of intangibles in agency conflicts and investment. Chen et al. (2024) focus on employee deferred compensation to examine the role of process intangibles on agency conflicts and investment.

¹⁶Our two overconfidence measures, derived from Malmendier and Tate (2005), are ‘longholder’ and ‘holder67’.

3.3 Empirical methodology

Our empirical methodology can be summarized in the following steps: (1) specification of expert definitions, states, actions and features, (2) estimation of the policy and transition functions in the data and (3) recovery of the value function, reward function, and feature weights.

Specification of experts, states, actions, and features. Each specification is a combination of 3 elements, (1) definitions of our two experts, (2) a set of state variable(s) and action variable(s) with a corresponding number of bins for each state and action variable, and (3) a set of features.

The first step consists in defining the “experts,” that is, the time period corresponding to different economic conditions. Conceptually, expert definitions are queries that can be applied to a data set in order to obtain a sub-sample of all observations. Our expert definitions are based on time periods and bundle together all firm-year observations that belong to a certain year range. The underlying assumption is that the reward function should be the same for both experts, i.e., not change over time. We remove all observations that do not correspond to the definition of either expert.¹⁷ We compute discount factors as $\gamma_i = \frac{1}{1+\bar{r}_{f,i}}$ for $i = 1, 2$, where $\bar{r}_{f,i}$ denotes the average risk-free rate during the respective period. We assume that both experts share the same preference for entropy parameter $\lambda = 1$.

We represent states as combinations of several state variables and actions as combinations of one or several action variables. We define state and action variables based on quantiles. In line with the theoretical literature on corporate investment, we generally define states as (size quantile, profitability quantile)-pairs and actions as investment rate quantiles or (investment rate quantile, net debt to EBITDA quantile)-pairs. As a baseline, we opt for five bins, i.e., quintiles, for each state and action variable. Table C.2 shows the cutoff values we use to determine quintiles (20%, 40%, 60%, 80%).

Features map (state, action)-pairs into (real) numbers. Generally speaking, the feature array will be of dimension $|\mathcal{S}| \times |\mathcal{A}| \times n_f$. We measure each feature as the median value of a variable across all (non-missing) observations corresponding to a given (state, action)-pair. If a feature is never observed for a given (state, action)-pair, we impute its value by the full sample median. Finding a parsimonious set of features is one of the main empirical explorations in this paper.

Estimation of policy and transition functions. Given a specification of expert definitions for $i = 1, 2$, states s and actions a , we determine for each firm-year observation which state it is in, which

¹⁷We aim to define experts in a way that partitions the full data set into two subsets and avoid losing observations.

action is chosen and which state s' the firm will be in next year. We require that the next state s' be observed and consequently drop all firm-year observations for which s' is missing.

We estimate choice probabilities based on a multinomial logistic regression model. Let the outcome $y = 1, \dots, J$ denote the index of the action a , where $J = |\mathcal{A}|$ denotes the number of possible actions. Then, for each expert $i = 1, 2$, the probabilities of picking a certain action are given by

$$P_i(y = j | \mathbf{x}_i) = \begin{cases} \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_{i,j})}{1 + \sum_h^J \exp(\mathbf{x}_i \boldsymbol{\beta}_{i,h})} & , \text{ if } j = 2, \dots, J \\ \frac{1}{1 + \sum_h^J \exp(\mathbf{x}_i \boldsymbol{\beta}_{i,h})} & , \text{ if } j = 1 \end{cases} \quad (20)$$

where $\mathbf{x}_i$ is a $1 \times K$ vector of first-element unity and $K - 1$ observed conditioning variables. $\boldsymbol{\beta}_{i,j}$ is a $1 \times K$ vector of unknown parameters. Our conditioning variables include all state variables, interactions between state variables and squared state variables. For example, if size and profitability are the state variables, then $\mathbf{x}_i = [1, \text{Size}_i, \text{Profitability}_i, \text{Size}_i \times \text{Profitability}_i, \text{Size}_i^2, \text{Profitability}_i^2]$.

Since we have fully specified the density of y_i given $\mathbf{x}_i$, we estimate this model via maximum likelihood. For each observation $n = 1, \dots, N$ corresponding to expert i , the conditional log likelihood is given by

$$\ell_{i,n}(\boldsymbol{\beta}_i) = \sum_{j=0}^J \mathbb{1}_{\{y_{i,n}=j\}} \log[p_{i,j}(\mathbf{x}_i, \boldsymbol{\beta}_i)] \quad (21)$$

where $\mathbb{1}$ is an indicator function. We estimate $\boldsymbol{\beta}_i$ by maximizing $\sum_{n=1}^N \ell_{i,n}(\boldsymbol{\beta}_i)$ using the gradient-based L-BFGS algorithm.

To estimate the transition probabilities, we proceed in a similar fashion, except that j now represents the index of the next state, such that $J = |\mathcal{S}|$, and that vector $\mathbf{x}_i$ now includes all state variables, action variables, interactions between state and action variables and squared state and action variables. For example, if size and profitability are the state variables and the investment rate is the action variable, then $\mathbf{x}_i = [1, \text{Size}_i, \text{Profitability}_i, \text{InvestmentRate}_i, \text{Size}_i \times \text{Profitability}_i, \text{Size}_i \times \text{InvestmentRate}_i, \text{Profitability}_i \times \text{InvestmentRate}_i, \text{Size}_i^2, \text{Profitability}_i^2, \text{InvestmentRate}_i^2]$.

Recovery of feature weights, reward function and value functions. Given the transition functions $T_i(s'|s, a)$, policy functions $\pi_i(a|s)$, discount factors γ_i and the feature array f , we proceed as follows. First, we create, for each agent $i = 1, 2$, arrays y_i , based on (5), and X_i , based on (6). Next, we define the X matrix based on (11) and test if Theorem 2 holds, i.e., whether $\text{rank}(X) = 2|\mathcal{S}| + d$. If this

condition holds, we compute $(f, M_{ij}, y_i, y_j, P_{ij}, M_j, M_f)$ in the data. Finally, we recover the coefficients θ based on (A.26), the reward from $\hat{u} = f\hat{\theta}$, and the value $\hat{v}_i$ based on (A.27) (for $i, j = 1, 2$ with $i \neq j$).

Performance evaluation. We can assess the fit of each model specification by using the recovered reward to solve for each expert’s optimal policy function that is consistent with this recovered reward. We then compare this recovered policy function to the actual one. Within our framework, investment policies are probability distributions over actions in a given state. To assess how similar the recovered policy functions are to the ones we observe for the real CEOs, we compute, for each state s , the Kullback-Leibler divergence (KLD) between the recovered policy function $\hat{\pi}_i$ and the policy function π_i of expert i as

$$KLD_{i,s}(\hat{\pi}_i, \pi_i \mid s) = \sum_{a \in \mathcal{A}} \hat{\pi}_i(a \mid s) \log \left(\frac{\hat{\pi}_i(a \mid s)}{\pi_i(a \mid s)} \right), \quad (22)$$

and take the average across states to arrive at the mean KLD

$$\overline{KLD}_i = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} KLD_{i,s}. \quad (23)$$

The KLD, also known as relative entropy, measures the difference between the two probability distributions capturing investment across states. A large KLD can be interpreted as a measure for how the probability distribution representing the model diverges from the true empirical distribution. $\overline{KLD}$ captures the expected log difference in investment probabilities; they equal zero if (and only if) $\hat{\pi} = \pi$, and lower values indicate that the two policies are more similar.

4 Recovering Corporate Purpose

We now recover the corporate objective function given different specifications of experts, state variables, action variables and features. In all settings, we verify that the rank conditions of Theorem 3 are satisfied. We organize the empirical analysis around the neoclassical investment theory and alternative theories.

4.1 Neoclassical investment theory vs. alternative models

At its core, the neoclassical theory of investment states that the reward of the corporate investment problem should consist of profits (Π), minus the gross investment ($I = K' - (1 - \delta)K$) needed to generate

them. Scaling the reward by capital (K) to make it independent of firm size, this implies a reward of the form $u = f\theta$, where we have two features, $f = [\frac{\Pi}{K}, \frac{I}{K}]$, in which Π/K denotes profitability and I/K denotes the gross investment rate, and the neoclassical weights are $\theta = [1, p]^T$ and $p = 1$. It would be a puzzle in the corporate reward if managers use $p < 1$ when deciding investment.

Definition 1 (Corporate Reward Puzzle). *The corporate reward puzzle is that the corporate reward function of the form $u = f\theta$, where the investment rate $I/K \in \{f\}$ is among the features and the weight $p = \theta_{I/K} < 1$. In this case, managers de-facto over-invest and under-explore.*

Table 2 shows the results of applying our recovery theorem to Compustat data on managers' decisions assuming the neoclassical theory of investment provides a useful foundation. We report both the raw and normalized (in parantheses) recovered coefficients of the reward function. Since the feature weights are identified up to scaling and λ is set to one in the recovery of the raw coefficients, we normalize θ by the absolute value of the coefficient on profitability, $\theta_{\Pi/K}$, such that profitability always has a normalized coefficient of 1 (or -1) and the magnitude of all other coefficients is expressed relative to profitability.¹⁸

$$\text{Manager reward } u_{it} = \text{Profitability}_{it} - p \times \text{Investment rate}_{it} + \epsilon_{it}. \quad (24)$$

Columns (1) to (5) differ in our choice of sample and expert definition. In the implementation, we opt for a grid of 5 size bins, 5 profitability bins, and 5 investment rate bins. This results in 25 possible economic states (combinations of size quintile and profitability quintile) and 5 possible actions (investment rate quintile). In Appendix E, we expand the state and action sets to allow for a finer grid. We choose year-based expert definitions such that the full sample is split into are a roughly equal number of years and observations for each expert. Expert 1 corresponds to the time period from 1975 until 1998, whereas expert 2 corresponds to the time period from 1999 until 2021. Defining experts based on years implicitly assumes that the reward function stays constant across time. In what follows, we will investigate whether this is true by altering expert definitions. Each time period is associated with a discount factor γ that is the average risk-free rate during that time period. A known limitation of the foundational q -theory model is its inability to account for the growing importance of intangible capital. Peters and Taylor (2017) provide

¹⁸The reasoning is as follows. Our framework assumes that agents have a preference for entropy, governed by the λ parameter whose value is unknown. The raw coefficients are recovered under the assumption that λ is fixed, and λ scales the coefficients (up or down). If we assume that $\lambda = 10$ instead of $\lambda = 1$, we recover coefficients that are ten times larger. However, the relative magnitude of the coefficients is not affected which allows us to normalize all coefficients by the same amount. We choose to normalize by the absolute value of the coefficient on profitability because there is a strong prior in the literature that profitability should have a coefficient of one.

alternative variable definitions that account for total capital, comprised of both tangible and intangible capital, and the corresponding total investment rate, as well as an alternative measure of profitability that takes into account the tax-reducing effects of intangible investment. We make this our base case.

Column (1) reports our results for the simplest possible specification with two features. We find a normalized coefficient of -0.494 on the investment rate. The positive coefficient on profitability indicates that profitability is a desirable feature, as prescribed by the theory. The negative sign of the coefficient on the investment rate is in-line with the theory, as investment is costly. However, the magnitude of the coefficient is only about half of what the foundational model would prescribe. For a given amount of capital, managers act as if \$1 of investment would only cost them \$0.494. This finding suggests several possibilities: the model may be correct but managers' preferences are heavily non-standard, or the benchmark model is overly simplistic and fails to accurately describe the data.

The experimentation parameter equals $\lambda = 0.029$ which suggests the manager faces moderate exploration incentives and, hence, a trade-off between exploitation (maximizing immediate rewards) and exploration (gathering information about the economic environment) with a relative weight of 2.9% ($=0.029/(1+0.029)$) in Equation (1) or (13). We can assess the model fit to the data using the same approach as in the sandbox environment. The mean KL divergence, defined in (23), in our base model for expert 1 equals 0.407 and is lower than for expert 2, which equals 0.554. In economic terms, the $\overline{KLD} = 0.407$ (0.554) means the entropy, or model inefficiency increases by 40.7% (55.4%).

Last, the managerial alignment measure, computed using expression (12), is dependent on the time period. For the period from 1975 until 1998, the manager-shareholder alignment is relatively high at 63% ($Alignment_1=0.63$). However, for the later time period from 1999 until 2021, the manager-shareholder alignment is lower at 58% ($Alignment_2=0.581$) and we will explore the reasons in Section 5.

Column (2) modifies the Peters and Taylor (2017) definition of profitability and investment to check how the results depend on the variable definitions. With the more narrow definition of profitability and physical investment rate defined in Table C.1 of Appendix C, many of the results remain unchanged except for the experimentation parameter and model fit. The price of capital is relatively unchanged at $p = 0.452$. When the model matches physical investment, however, $\lambda = 0.11$ is estimated to be much higher while the model fit is much better but managerial alignment is less uniform across time periods. The reason is likely that neoclassical features better fit physical investment dynamics than they capture intangibles.

Columns (3) and (4) split the sample into large and small firms, respectively, defined as firms having

Table 2: Corporate reward recovery from investment data.

The table presents the results of recovering the reward function from corporate investment data based on our main sample from 1975 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability and investment rate. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Tobin's q (Specification 1) as well as Total q (Specifications 2 to 5). N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively. Specifications (1) and (2) rely on alternative definitions of profitability and investment. Specifications (3) and (4) present results for large and small firms, respectively. Large/small firms are defined as firms having total assets above/below the 67th/33rd-percentile. In specification (5), Agent 1 and 2 are defined as firms being in recession or non-recession years, respectively.

	Recovered feature importance θ (θ normalized)				
	(1)	(2)	(3)	(4)	(5)
Sample:	All firms	All firms, physical capital	Large firms	Small firms	Alternative expert def.
Alt. Profitability	34.235 (1.000)		32.139 (1.000)	22.512 (1.000)	26.513 (1.000)
Total Investment Rate	-16.920 (-0.494)		-16.045 (-0.499)	-7.238 (-0.322)	-11.800 (-0.445)
Profitability		9.124 (1.000)			
Investment Rate		-4.125 (-0.452)			
λ	0.029	0.110	0.031	0.044	0.038
$\overline{KLD}_1$	0.407	0.158	0.497	0.302	0.356
$\overline{KLD}_2$	0.554	0.134	0.582	0.243	0.408
$Alignment_1$	0.630	0.694	0.719	0.794	0.747
$Alignment_2$	0.581	0.106	0.468	0.257	0.577
N_1	59,743	59,743	19,983	14,349	27,116
N_2	50,890	50,890	33,455	3,333	83,517

total assets above/below the 67th/33rd-percentile. While $p = 0.499$ is slightly larger for large firms, $p = 0.322$ is significantly lower small firms. Comparing the experimentation parameter λ , small firms have more experimentation incentives than large firms ($\lambda=0.044$ vs. 0.031). The mean KL divergences suggest that the model better fits small firms than large firms. Column (5) introduces yet another variation. Here we define the expert regimes based on firms being in recession (1) or non-recession (2) years. The number of observations now varies more across regimes. Results are again similar to the base case in column (2) with $p = 0.445$ and $\lambda = 0.038$. Model fit is also about similar and alignment is better in recession years. Overall, none of the specifications generate a p close to one.

The *corporate reward puzzle* documented in [Table 2](#) is that the weight on current profitability is larger in absolute terms than the weight on the investment rate. For the normalized feature importances reported in parenthesis, $\theta_{\frac{I}{K}} \in (-1, 0)$. In economic terms, these results mean that managers' behavior is as if they use a price of investment $p < 1$ when deciding investment. [Table E.1](#) presents the results across nine industries using the Fama-French classification. We find significant variation in recovered feature importance across industries. Profitability is consistently the most important factor and investment rate is negatively weighted, as one would expect, with the smallest distortions in manufacturing and health

care. The exploration-exploitation trade-off (λ) is lowest in capital-intensive industries like energy and manufacturing, suggesting a stronger focus on exploitation. The $\overline{KLD}$ values indicate moderate divergence between estimated and true policies, with variation across industries. Alignment with shareholder value differs significantly, with business equipment and health care showing the highest alignment, while energy exhibits the weakest. Overall, while the table highlights systematic differences in managerial decision-making and alignment across industries, $p < 1$ across all industries.

4.2 Corporate reward puzzle: Alternative theories and empirical tests

There are several alternative models and explanations that would yield a price of investment less than one. We briefly review these economic channels below and incorporate their economic mechanisms or insights into our analysis to help rationalize the corporate reward puzzle.

Agency and free cash-flow hypothesis. The free cash-flow hypothesis of Jensen (1986) (see also Richardson (2006)) suggests that firms with excess internal funds and limited growth opportunities are prone to over-investment. When managers control substantial free cash-flow but lack strong external monitoring, they may invest in negative net present value projects to increase firm size, managerial perks, or personal job security rather than maximizing shareholder value. This theory predicts that over-investment is more pronounced in firms with low debt and weak governance.

Trade-off theory and financial frictions. The trade-off theory of capital structure argues that firms balance the tax and agency benefits of debt against the potential financial distress and agency costs, thereby influencing investment behavior. In this framework, an under-levered firm leaves the managers with excess free cash-flow, thereby increasing the propensity for the manager to engage in over-investment in projects that do not enhance shareholder value. Also, an over-levered firm is prone to shareholder-bondholder agency conflicts that can lead to over-investment in the form of risk-shifting.

Intangible capital. More recently, Ward (2023) and Chen et al. (2024) examine how the intangible nature of capital affects managerial compensation and investment behavior. They show that the opaque nature of intangibles exacerbates agency conflicts within the firm. In this setup, intangibility is associated with higher investment and deferred managerial compensation mitigates the agency conflicts.

Moral hazard models. Moral hazard models highlight how asymmetric information and agency conflicts between managers and shareholders lead to inefficient investment decisions. When managers control corporate funds but do not fully internalize the costs of their investment choices, they may allocate

resources to projects that enhance their private benefits rather than maximize firm value (Myers and Majluf, 1984; Shleifer and Vishny, 1989; Stulz, 1990). Performance-based compensation is commonly proposed to align managerial incentives with shareholder value and hence mitigate moral hazard problems.

Dynamic contracting models. In the dynamic contracting literature, DeMarzo and Fishman (2007), Bolton et al. (2011), and DeMarzo et al. (2012) examine how incentive constraints influence firms' investment decisions over time. These models reveal a trade-off between incentivizing managers and ensuring efficient capital allocation, often leading to investment distortions. In this setup, incentives are primarily provided through deferred compensation, which aligns managers' interests with long-term firm value. In dynamic contracting models, there is typically under-investment.

Behavioral models. Behavioral models posit that managerial biases can lead to systematic deviations from optimal investment behavior, including over-investment. These models challenge the neoclassical assumption of fully rational managers maximizing shareholder value and instead suggest that overconfidence (Malmendier and Tate, 2005) and miscalibration (Ben-David et al., 2013) drive excessive capital expenditures and acquisitions. Overconfident managers tend to overestimate project returns and underestimate risks, leading them to invest beyond the firm's optimal level, even in the presence of financing constraints.

These theories collectively highlight how financial frictions, agency conflicts, and behavioral biases can lead to excessive investment beyond the optimal shareholder level. The empirical literature documents both over- and under-investment (Blanchard et al., 1994; Cho, 1998; Richardson, 2006; Ferreira and Matos, 2008; Cronqvist and Fahlenbrach, 2009; Ben-David and Chincio, 2024; Jha et al., 2024). Departing from the existing theoretical and empirical literature, our approach can recover managers' incentives including their trade-off between exploration and exploitation. In the model, manager balance learning about uncertain payoffs (exploration) with maximizing immediate rewards (exploitation). Incorporating this trade-off provides insights into how firms dynamically select their investments.

4.3 Financial factors entering manager's reward

Several of the theories, including the free cash-flow and trade-off theory, predict that financial factors enter the manager's reward. One limitation of the foundational model is that it only has features related to the real side of investment, whereas managers may take into account financial conditions. The additional features that are commonly used in empirical corporate finance studies to account for differences in firms' finances and, as such, are plausibly affecting the reward function are related to the firm's financial position

Table 3: Corporate reward recovery - Financial characteristics.

The table presents the results of recovering the reward function from corporate investment data based on our main sample from 1975 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and firm financial characteristics. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)				
	(1)	(2)	(3)	(4)	(5)
Financial factor:	Cash	Leverage	Tangibility	Distress	All
Alt. Profitability	34.698 (1.000)	29.716 (1.000)	36.283 (1.000)	15.981 (1.0000)	19.064 (1.000)
Total Investment Rate	-17.062 (-0.492)	-15.549 (-0.523)	-17.667 (-0.487)	-13.804 (-0.864)	-15.803 (-0.829)
Cash Holdings	-1.468 (-0.042)				
Net Book Leverage		-3.875 (-0.130)			5.771 (0.303)
Tangibility			0.720 (0.020)		1.052 (0.055)
Altman Z-score				1.333 (0.083)	1.817 (0.095)
λ	0.029	0.034	0.028	0.063	0.052
$\overline{KLD}_1$	0.403	0.412	0.403	0.413	0.392
$\overline{KLD}_2$	0.542	0.589	0.580	0.611	0.601
$Alignment_1$	0.634	0.546	0.635	0.410	0.442
$Alignment_2$	0.647	0.534	0.569	0.186	0.209
N_1	59,743	59,743	59,743	59,743	59,743
N_2	50,890	50,890	50,890	50,890	50,890

and likelihood of distress. We therefore include them as additional features:

$$\text{Manager reward } u_{it} = \text{Profitability}_{it} - p \times \text{Investment rate}_{it} + \theta^\top \text{Financial factors}_{it} + \epsilon_{it}. \quad (25)$$

Table 3 shows the results for different financial feature sets that include cash holdings, net book leverage, tangibility, and the Altman Z-score as a summary proxy for financial health. All empirical models reported in columns (1)-(5) assign positive weights to profitability, confirming its significant positive impact on the manager’s reward function. The coefficients for the investment rate are negative, confirming the expected inverse relationship with the manager’s reward function. However, the magnitudes vary across models. As we add financial features, the coefficient on the investment rate remains negative and above -1, ranging from -0.487 in column (3) to -0.864 in column (4), and hence it is now closer to -1 as posited by neoclassical theory.

Financial factors, including cash holdings, net book leverage, tangibility, and the Altman Z-score, contribute to the corporate reward function, though their impact is mixed across models. The normalized coefficient on cash holdings is negative at -0.042, suggesting cash holdings are costly in terms of diminishing

the manager’s reward. The coefficient can be interpreted as a 4.2% annual cost rate. Net book leverage exhibits variation, being negative in column (2) but positive in column (5), likely due to correlation with the other features. Tangibility’s positive influence on the reward function receives relatively low but stable weights. The normalized coefficient on tangibility is positive at 0.020 and 0.055, respectively, suggesting tangibility lowers the costs associated with investment.

Similarly, an increase in financial health improves the manager’s reward. The Altman Z-score has a positive weight of 0.083 (column 4) and 0.095 (column 5), suggesting an important role of the threat of distress in shaping managers’ incentives. An Altman Z-score close to 0 suggests a company is headed for bankruptcy, while a score of 3 is considered a solid financial positioning. Notably, including a proxy for financial health increases p to 0.894 (column 4) and 0.829 (column 5). As we add features, the mean KL divergence of expert 1 remains relatively stable around 0.392-0.413 across models, whereas the KL divergence of expert 2 increases from 0.542 in column (1) to 0.611 in column (4).

This suggests financial factors help explain policy choices in some periods but not uniformly across time, and more financial features may lead to overfitting. Therefore, other features may be needed to improve the models’ predictive accuracy and robustness over time. Also, as we add financial health as a feature the alignment measure drops significantly, indicating that a large distance to default may be to the benefit of managers but not firm-value maximizing.

4.4 Taking account of intangible capital

While the [Peters and Taylor \(2017\)](#) definition of profitability and investment reduce the recovered experimentation incentives and improve managerial alignment measures, they do not resolve the corporate reward puzzle of [Table 2](#). [Ward \(2023\)](#) and [Chen et al. \(2024\)](#) show that the intangible nature of capital affects investment behavior through determining agency conflicts. To describe the different facets of a firm’s intangible assets, we include a granular measurement of intangible capital in the empirical analysis.

[Table 4](#) presents the results of our analysis after including the additional features from [Ward \(2023\)](#). While intangible capital alone does not resolve the corporate reward puzzle from [Table 2](#), intangibility aspects matter and improve model fit. Column (1) indeed finds that intangible intensity increases the manager’s reward, consistent with agency conflicts. These conflicts arise because intangible assets are often difficult to value and monitor, leading to underinvestment despite high valuations and cash flows. The coefficient is 3.333 (0.172 normalized), suggesting a 1-SD increase in intangible intensity keeping

total investment fixed gives 4.5% ($=0.172*0.26$) more reward to the manager. Column (2) shows that knowledge intensity also increases the manager’s reward, with an even larger (normalized) coefficient of 8.168 (0.327). Column (3) shows that organizational intensity decreases the manager’s reward, with a (normalized) coefficient of -8.625 (-0.354), as one would expect.

Incorporating aspects of intangibility improves the model fit with the largest improvement from intangible intensity. The overall model fit, as measured by the mean KL divergence, shows some variation across specifications. The KL divergence ranges from 0.339 to 0.415 for 1975-1998 and from 0.377 to 0.637 for 1999-2021. These values indicate in some cases an improved fit relative to Table 2, suggesting that incorporating intangible capital captures additional aspects of decision-making but does not fundamentally alter the reward structure. Alignment with Total q remains relatively strong, with values ranging from 0.738 to 0.927 for 1975-1998 and from 0.601 to 0.918 for 1999-2021.

The experimentation parameter is stable at between 0.04 and 0.052 across columns 1-3, indicating moderate exploration incentives. The cost of investment parameter p on the total investment rate improves slightly to 0.655 compared to the base case but remains below 1. These findings confirm that managers still act as if investment were less costly than theory predicts, though the magnitude of the deviation is reduced. Understanding intangibility intensities is hence important for understanding how intangible assets influence corporate behavior, investment decisions, and the dynamics between managers and investors.

The inclusion of intangible capital interactions in columns 4-6 provides additional insight into the structure of investment decisions. The interaction between intangible and knowledge intensity is pronounced, with (normalized) coefficients of 13.564 (0.535) and 15.213 (1.248) in columns (4) and (6), respectively, suggesting complementarities between these two dimensions of intangible capital. In contrast, while organizational intensity has a negative stand-alone impact in column (3), its interaction with intangible intensity in columns (5) and (6) yields small positive effects of 3.770 (0.159) and 4.295 (0.352).

In Table E.2 we incorporate process intangibles using data from 2000 to 2021 based on [Chen et al. \(2024\)](#). The table provides a nuanced picture of how different facets of intangible capital influence corporate investment behavior. Overall intangible intensity shows positive weights, suggesting that investing in intangible assets is rewarded. However, when focusing on process-related intangibles, process intensity and process intangibles carry negative weights, implying that aspects of process intangibility may detract from immediate reward, possibly due to their inherent costs or risks associated with their implementation.

Overall, while incorporating intangible capital helps refine our model of corporate investment decisions,

Table 4: Corporate reward recovery - Intangible capital.

The table presents the results of recovering the reward function from corporate investment data based on our main sample from 1975 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing the intangible nature of firm’s asset. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	19.369 (1.000)	24.977 (1.000)	24.341 (1.000)	25.341 (1.000)	23.641 (1.000)	12.190 (1.000)
Total Investment Rate	-12.687 (-0.655)	-14.187 (-0.568)	-13.979 (-0.574)	-14.199 (-0.56)	-14.035 (-0.594)	-10.582 (-0.868)
Intangible intensity	3.333 (0.172)					
Knowledge intensity		8.168 (0.327)				
Organizational intensity			-8.625 (-0.354)			
Intangible int. x Know. int.				13.564 (0.535)		15.213 (1.248)
Intangible int. x Org. int.					3.770 (0.159)	4.295 (0.352)
λ	0.052	0.040	0.041	0.039	0.042	0.082
$\overline{KLD}_1$	0.374	0.415	0.413	0.406	0.376	0.339
$\overline{KLD}_2$	0.377	0.637	0.638	0.618	0.388	0.417
$Alignment_1$	0.738	0.851	0.849	0.845	0.663	0.927
$Alignment_2$	0.918	0.705	0.672	0.771	0.757	0.601
N_1	59,743	59,743	59,743	59,743	59,743	59,743
N_2	50,890	50,890	50,890	50,890	50,890	50,890

it does not resolve the core discrepancy between theory and observed behavior. The finding that $p < 1$ persists, indicating that additional factors beyond intangible capital need to be considered to fully capture the reward function guiding managerial decisions.

4.5 Role of compensation policies

To better understand the nature of agency conflicts and the role of compensation policies in mitigating their impact, as reflected in the corporate reward puzzle documented in [Table 2](#), we now investigate the importance of (non-)managerial compensation. We split this analysis into three elements: Management’s total vs. deferred compensation, composition of managerial compensation, and managerial vs. non-managerial compensation.¹⁹ Moral hazard models highlight that performance-based compensation is a key mechanism to align managerial incentives with shareholder value and hence mitigate incentive problems. We therefore add features relative to performance pay. Dynamic contracting models highlight, in turn, that incentives are primarily provided through deferred compensation, which aligns managers’ interests with long-term

¹⁹The inclusion of these features implies that we need to change expert definitions, as executive compensation data is available only from 1992 onward. Expert 1 is now defined as the time period from 1992 until 2006, whereas expert 2 is defined as the time period from 2007 until 2021.

Table 5: Corporate reward recovery - Compensation policy.

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 1992 to 2021 in columns 1-3, 2004 to 2012 in column 4, and 1992 to 2021 in column 5 depending on the data sources. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing CEO and managerial bonus and ownership. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)				
	(1)	(2)	(3)	(4)	(5)
Compensation:		CEO	Management	Management deferred	Employees
Alt. Profitability	25.175 (1.000)	25.683 (1.000)	26.074 (1.000)	19.468 (1.000)	11.508 (1.000)
Total Investment Rate	-14.118 (-0.561)	-14.471 (-0.563)	-14.658 (-0.562)	-13.805 (-0.709)	-12.378 (-1.076)
CEO Ownership		7.901 (0.308)			
CEO Bonus		-4.777 (-0.186)			
Managerial Ownership			9.389 (0.360)		
Managerial Bonus			-1.868 (-0.072)		
Total Compensation				4.005 (0.206)	
Manag. Deferred Comp.				60.758 (3.121)	
Fraction Deferred Comp.				0.349 (0.018)	
Employee Deferred Comp.					45.781 (3.978)
λ	0.040	0.039	0.038	0.051	0.087
$\overline{KLD}_1$	0.383	0.367	0.368	0.264	0.238
$\overline{KLD}_2$	0.560	0.535	0.542	0.442	0.599
$Alignment_1$	0.433	0.436	0.383	0.205	0.435
$Alignment_2$	0.601	0.647	0.569	0.172	0.815
N_1	39,699	39,699	39,699	19,051	22,193
N_2	29,753	29,753	29,753	13,149	15,441

firm value. We therefore add features relative to present vs. deferred pay. The normalized coefficients θ capture the importance of various facets of compensation:

$$\text{Manager reward } u_{it} = \text{Profitability}_{it} - p \times \text{Investment rate}_{it} + \theta^\top \text{Compensation policy}_{it} + \epsilon_{it}. \quad (26)$$

Table 5 shows the consolidated results when we add various compensation policies. In the Appendix, we report further specifications that provide more detailed analysis where we add compensation features once at a time. Results in Tables E.3, E.4, E.5 are consistent with the consolidated Table 5. For comparison, column (1) reports the two-feature results for the sample of firms from 1992 to 2021 due to the limited availability of compensation data, largely confirming prior findings.²⁰ We interpret the coefficients

²⁰Compared to column (1) of Table 2, the coefficient on the investment rate has increased (in absolute value) to -0.561 (compared to -0.494 before). The mean KL divergence has changed to 0.383/0.560 (compared to 0.407/0.554

considering that under the agency frictions hypothesis ([Jensen and Meckling \(1976\)](#)), managers maximize their own wealth which implies that managerial compensation in the form of a bonus or ownership should have a positive weight in the reward function. Under the no-agency frictions hypothesis, managerial compensation should have a negative coefficient as it is a cash outflow which reduces shareholder profits.

In column (2), we add CEO compensation features including CEO ownership and bonus (we have also analyzed CEO options separately). We find that CEO ownership has a positive coefficient of 7.901 (0.308 normalized), whereas CEO bonus has a negative coefficient of -4.777 (-0.186 normalized). The positive coefficient on CEO ownership suggests that this feature is desirable, because shareholders align the CEO’s incentives with their own and/or the CEO likes control. The negative coefficient on CEO bonus suggests that the manager is not acting purely in their self-interest. Adding CEO compensation features improves the model’s performance in terms of mean KLD and alignment, but not in aligning p with theory.

While the CEO is ultimately responsible for any decision taken within the firm, other executives contribute to the decision-making process. We therefore repeat our analysis taking into account the compensation of the top five executives by total current compensation. Column (3) shows that the results are qualitatively unchanged if we use top managerial instead of CEO compensation. The results in column (3) for top managerial compensation are qualitatively similar to those in column (2) for CEO compensation. The coefficient on managerial ownership (9.389, 0.360 normalized) is higher than that for CEO ownership (7.901, 0.308 normalized). This suggests that broader managerial ownership plays a stronger role in aligning incentives. At the same time, the managerial bonus coefficient (-1.868, -0.072 normalized) is less negative than the CEO bonus coefficient (-4.777, -0.186 normalized), implying that CEO-specific incentives through bonuses may be more distortive than those at the broader executive level.

Dynamic contracting theories highlight the importance of deferred compensation in aligning incentives and limiting investment inefficiencies. Our next set of specifications, reported in column (4) and [Table E.4](#), focus on the temporal composition of managerial compensation. We include total compensation from Execucomp, managerial deferred compensation, and the fraction of deferred compensation. Introducing deferred compensation variables improves model performance. The absolute value of the investment rate coefficient increases to -0.709, indicating stronger discipline on investment decisions. Model fit also improves. The mean KL divergence decreases (compared to prior specifications), implying that the estimated policy is closer to the true policy. However, alignment is very low. The alignment measures

before). This indicates a similar fit.

drop significantly (to 0.205 for *Alignment*₁ and 0.172 for *Alignment*₂), suggesting that while deferred compensation plays a role in shaping managerial incentives, it does not strongly align them with Total q. This indicates the potential presence of other frictions affecting firm value beyond deferred pay incentives.

In our final analysis, we focus on non-executive employees. We obtain data from Compustat on deferred compensation for all employees. Compared to the other specifications, column (5) highlights the stronger role of employee deferred compensation. Employee deferred compensation has a large positive coefficient (45.781, 3.978 normalized), suggesting that deferred pay for employees plays a significant role in shaping firm incentives. This effect is even stronger than managerial deferred compensation (60.758, 3.121 normalized in column 4), implying that long-term compensation mechanisms for the broader workforce contribute to firm performance. The investment rate coefficient becomes closer to -1 (-1.076 normalized), suggesting a better investment discipline when including employee compensation. This could indicate that firms with substantial deferred employee pay face optimal investment dynamics, potentially due to lower short-term pressure on managers. Model fit also improves in terms of mean KL divergence (0.238 for $\overline{KLD}_1$ and 0.599 for $\overline{KLD}_2$), though not as strongly as in column (4). Unlike in column (4), alignment measures increase sharply (to 0.435 for *Alignment*₁ and 0.815 for *Alignment*₂). This suggests that employee deferred compensation is a means to align with firm value (Total q), more than managerial deferred pay. A potential explanation is that broader deferred compensation aligns employees' efforts with firm success, reducing agency conflicts at multiple levels.

Overall, ownership effects are stronger than bonus incentives for both CEOs and managers. Deferred compensation improves model fit, but its impact on alignment varies. Managerial deferred compensation improves fit but reduces alignment measures. Employee deferred compensation improves both fit and alignment, suggesting a firm-wide incentive effect. Investment discipline improves when including employee deferred pay, implying different mechanisms at play compared to managerial incentives.

4.6 Behavioral theories and other hypotheses

Behavioral theories are harder to test in our context since data on CEO's and other managers' behavioral traits are not widely available. We collect data on CEO age, tenure, overconfidence from Execucomp and Equilar. Table E.6 in the Appendix examines the influence of CEO behavioral characteristics on the recovered reward function in a reduced data set and, therefore, results should be interpreted with caution. Across all specifications, profitability remains the most important factor, with a consistently high positive

coefficient (around 25.2) and a normalized value of 1.000. In contrast, investment rates consistently exhibit negative coefficients ranging from -0.65 to -0.56. The inclusion of CEO age (column 2) results in a small negative coefficient (-0.034) but is nearly zero in normalized terms (-0.001), suggesting little economic significance. CEO tenure (column 3) has a slightly positive coefficient (0.018), but again, its normalized effect is negligible (0.001). These findings suggest at face value that age and tenure do not meaningfully shape firm incentives to invest.

The standard measure of CEO overconfidence (column 5) is strongly positive (0.905, normalized 0.041), suggesting that overconfident CEOs place greater weight on profitability in decision-making. However, the alternative measure of overconfidence (column 6) is negative (-0.014, normalized -0.001), hinting at inconsistencies across different ways of capturing overconfidence. This indicates that CEO overconfidence influences investment decisions, but its exact effect depends on how overconfidence is measured.

The exploration-exploitation parameter λ is relatively stable across specifications, ranging from 0.040 to 0.046, suggesting a similar level of decision flexibility across different models. Model fit improves when behavioral characteristics are included, as seen in the lower KL divergence values. $\overline{KLD}_1$ drops from 0.383 (column 1) to 0.260 (column 5) and $\overline{KLD}_2$ decreases from 0.560 to 0.477. This suggests that incorporating CEO overconfidence improves the model’s ability to match observed firm behavior.

Alignment metrics decrease when CEO behavioral traits are included. $Alignment_1$ drops from 0.433 to 0.302 (column 5), and $Alignment_2$ follows a similar decline. This suggests that while CEO overconfidence helps explain decision-making, it reduces alignment with long-term firm value, potentially reflecting suboptimal investment choices driven by overconfidence.

Overall, the analysis in this section suggest that the corporate recovery theorem from Section 1 can be used to recover the firm’s de-facto reward function. Material factors such a profitability and investment have the right signs, positive and negative, respectively, but the relative magnitude of current profits in managers’ rewards is larger than the cost of investments. Incorporating financial frictions, intangible capital, and (non-)managerial compensation partially resolve this corporate reward puzzle. In turn, CEO behavioral traits play a minor role overall, with overconfidence showing the strongest effect, while including behavioral characteristics improves model fit but reduces alignment with firm value, suggesting potential distortions in decision-making due to CEO biases. In the next section, we explore the extent to which the manager’s incentives align with shareholder value maximization as well as the impact of ESG.

5 Machine-Run Firms and the Alignment of Manager Rewards with Shareholder-Value Maximization

Even if none of the models from the previous section fully resolve the corporate reward puzzle, machine-run firms trained on behavioral data may still function properly if managerial behavior is aligned with the interests of shareholders. This section explores the biases arising from the corporate reward puzzle, the extent of manager-shareholder alignment, and the effect of non-pecuniary benefits, ESG, joint investment-financing behavior, and other extensions.

5.1 Biases of machine-run firms

We explore the impact on investment behavior of the corporate reward puzzle through the lens of the foundational q -theory model from Section 2.1. [Appendix D](#) summarizes the details of the experiment. To explore the impact of $p < 1$ or, equivalently, $\theta_{\frac{I}{K}} \in (-1, 0)$ on the dynamics of investment, we simulate the neoclassical investment model in two ways. In the benchmark case, we set $p = 1$, simulate profitability shocks and compute the optimal investment rates in each state. In the de-facto case, we set $p = 0.5$, simulate the same profitability shocks as in the benchmark case and compute the de-facto investment rates in each state.

The results in [Table D.1](#) show that managers de-facto behave as if they over-invest and under-explore compared to the case when $p = 1$. The over-investment is pervasive across profitability levels, with the relative increase the largest for low levels of profitability. In addition, managers appear to do too little exploration, by experimenting with varying investment less than optimal. This behavior is inconsistent with the predictions from the foundational q -theory model and supports some of the theories in [Section 4](#).

5.2 Alignment measure

To understand the source of this bias, we explore the extent of manager-shareholder alignment. Absent non-pecuniary benefits in the reward function, shareholder value maximization amounts to maximizing the market value of equity scaled by capital, or Tobin's q . Absent bondholder expropriation, shareholder wealth maximization amounts to firm value maximization, i.e., maximizing the market value of equity and debt. Scaling everything by capital, this implies maximizing Total q . Our methodology allows us to recover the value that managers actually maximize in practice. We can thus compare the manager's

recovered value $\hat{v}$ computed using expression (A.27) from Proposition 2 to Tobin’s q and Total q .

As discussed in prior sections, we can recover an agent’s value function only up to affine transformations, $v_i = a_v + b_v \hat{v}_i$, where a_v and b_v remain unknown. Hence, we cannot directly measure the manager’s value and compare it to shareholder value. Instead, we can construct the scale-free alignment measure (12) based on the correlation across different states between the recovered value that managers maximize and observed market values. For the purpose of this analysis, we implement it as

$$Alignment_i = Corr(\{\hat{v}_i(s)\}_{s \in \mathcal{S}}, \{Median(Total\ q_i(s))\}_{s \in \mathcal{S}}), \quad (27)$$

where $\hat{v}$ is recovered using expression (A.27) and the median is taken across all observations in state s .

Figure 2 shows the *Alignment* of the recovered value function with Total q and its relation with model complexity (measured by the number of features in the reward function) for all specifications in the main text. The blue dots represent *Alignment*₁ (corresponding to the early period, Expert 1), while the red squares represent *Alignment*₂ (the later period, Expert 2). The dotted/dashed lines correspond to fitted values. From the data, we observe that manager-shareholder alignment is generally high with an average (median) of 62% (64%) in the earlier period and 54% (58%) in the later period across the 21 specifications. The maximum values are 92-93%, with a minimum of 19-23%. The fitted lines are rather on top of each other, suggesting that alignment does not differ starkly between the early and late sample periods. When we test whether alignment has increased over time, we can reject this hypothesis, while it has decreased at 10% significance, with a p value of 0.0753 (though this test does not account for the sampling error in each alignment estimate). Across specifications, the alignment measure increases with model complexity up to a point, peaking at complexity 4 with 0.927, and then declines sharply. The alignment measures are more volatile in the later than earlier period. High model complexity is associated with a drop in alignment for both experts, suggesting overfitting in the most complex models or complex managerial reward structures not being well aligned with shareholder value.

Overall, these results suggest that machine-run firms would perform poorly since manager-shareholder alignment is imperfect and better-fitting more complex models of reward structures would imply even larger misalignment. Hence, there is a trade-off between model structure and alignment, which is novel and different from known ML/AI biases.

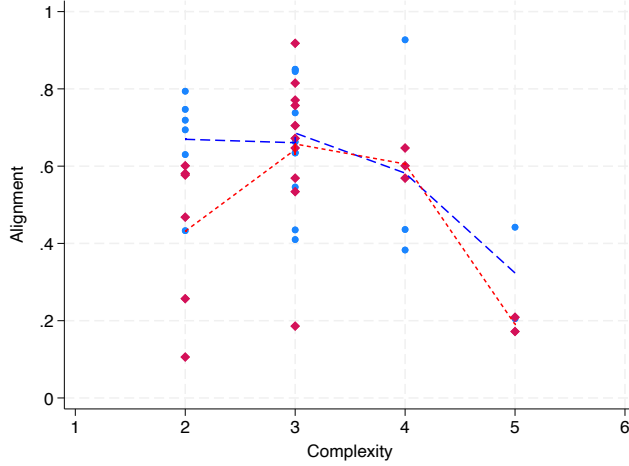


Figure 2: Alignment between recovered value and q . The figure plots the alignment between the recovered value function and Total q against model complexity. The blue dots and dashed line correspond to the early period (expert 1), the red squares and dotted line correspond to the later period (expert 2).

5.3 (Mis)alignment of manager’s reward with shareholder-value

The misalignment between the manager’s value and shareholders’ valuation can be better understood by comparing the valuations state-by-state across different firm sizes and profitability levels. If the alignment is high, the two should exhibit a similar shape. To do this we select a parsimonious specification that combines features from several of the considered theories, fitted to the data according to²¹

$$\begin{aligned}
 \text{Manager reward } u_{it} = & \text{Profitability}_{it} - 0.864 \times \text{Investment rate}_{it} - 0.108 \times \text{Net book leverage}_{it} \\
 & + 0.183 \times \text{Intangible intensity}_{it} + 0.972 \times \text{Deferred compensation}_{it} \\
 & + 0.014 \times \text{Fraction Deferred compensation}_{it}. \quad (28)
 \end{aligned}$$

This managerial reward specification yields an experimentation coefficient of $\lambda = 0.07$ with KLD of 0.298 and 0.318, respectively. The alignment measures are $Alignment_1 = 0.309$ in the earlier and $Alignment_2 = 0.732$ in the later time period.

Figure 3 illustrates the recovered value for the manager for each firm size-profitability combination in the left plot using specification (28). For comparison and to illustrate the (mis)alignment between manager value and shareholder value, the right plot depicts the median value of Total q for each corresponding firm

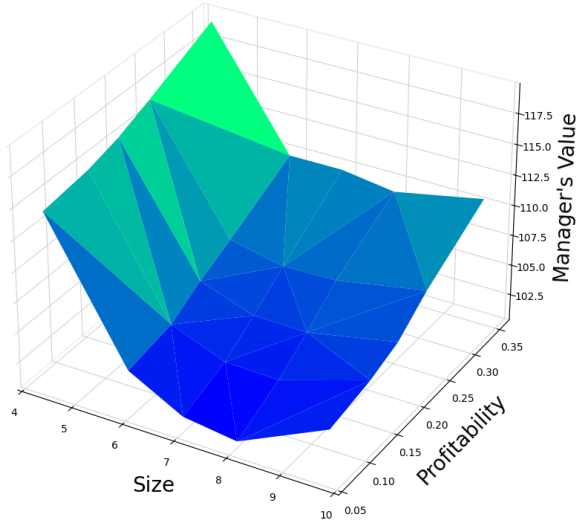
²¹We have also used a formal model selection procedure based on out-of-sample KLD and J score (log loss, or cross entropy). However, the small sample size limits precision.

size-profitability combination. The *Alignment* measure is the correlation between the two, with a value of 1 if the two plots have the same shape, though not necessarily the same levels.

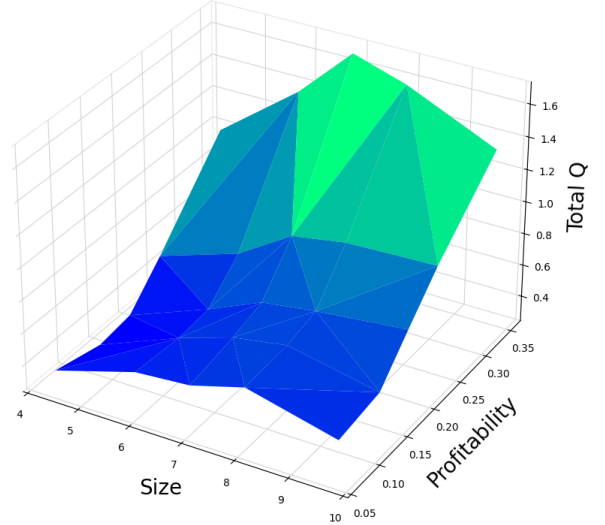
The figure illustrates there is a clear disconnect between the recovered value, especially for the lowest size quintile, and firm value. Both the manager's and firm value are monotonically increasing in profitability, as one would expect. However, mid-sized firms have the largest q while the smallest and largest firms have the lowest q in the Compustat data. Since Total q is the ratio of a firm's market value to the replacement cost of its assets, this inverse U-shaped relationship with firm size suggests that mid-sized firms achieve a more favorable balance. Small firms often face higher financial frictions, liquidity constraints, distress and uncertainty, and lack of diversification, which can depress their market valuations relative to asset replacement costs. Conversely, large firms tend to experience lower growth prospects, managerial inefficiencies, or legacy assets that are not as efficiently utilized, which also leads to lower q . Mid-sized firms, on the other hand, strike an optimal balance: they are large enough to benefit from access to financial markets and economies of scale, yet small enough to remain innovative and agile in product markets, resulting in higher q values.

In contrast, the figure illustrates that the manager's value, while increasing in profitability, is U shaped in firm size. The manager's value is largest for small, highly profitable firms and lowest for unprofitable mid-sized to large firms. The large increase in manager value for the smallest firm size category is particularly astonishing. Most notably, the manager's value and firm value are inversely related to each other in the firm size dimension. The inverse relationship between the manager's recovered value and firm value along the firm size dimension likely reflects differences in managerial discretion, agency frictions, and governance. In small, highly profitable firms, managers often have greater discretion and weaker external monitoring, allowing them to capture substantial private benefits—through compensation, perks, or personal control—that drive up their recovered value. In contrast, in mid-sized to large firms, stronger governance and institutional oversight constrain managerial behavior, reducing the ability to extract personal rents even when firm value is high. Thus, while firm value (Total q) peaks for mid-sized firms, the manager's value is highest in small firms presumably due to the enhanced scope for private benefit extraction and agency frictions that are less prevalent in larger organizations.

Overall, by comparing state-by-state valuations, it is evident that while both the manager's and firm's values increase with profitability, their relationship with firm size diverges. The inverse relationship highlights a disconnect between managerial priorities and shareholder value, particularly when firms are small.



(a) Manager's recovered value as a function of size and profitability.



(b) Shareholders' valuation of the company as a function of size and profitability.

Figure 3: Alignment and Total q . The figure plots the manager's recovered value (left) and company valuation (right) for different firm sizes and profitability levels.

5.4 Extensions of the analysis and robustness

We consider several extensions of our analysis in [Appendix E](#), including the effect of non-pecuniary benefits or detriments on corporate decision-making and the alignment between manager and shareholder value, and joint investment and financing behavior.

Under the shareholder or stakeholder welfare hypothesis, non-pecuniary benefits should have a positive effect on the reward. We measure these non-pecuniary benefits by environmental, social, and governance scores. Under shareholder wealth maximization, the environmental and social scores should have little to no effect on the manager's reward. Under the agency frictions hypothesis, the sign of the coefficient on governance is ambiguous. From the manager's point of view, strong governance mechanisms reflected in a high G score are undesirable because they limit their ability to misbehave. From the shareholder's point of view, strong governance is desirable, but costly.

Table [E.7](#) shows the results when we include ESG scores. Breaking up the ESG score into its individual components reveals that the environmental score has a negative effect on the reward (-0.004), whereas the social (0.005) and governance (0.005) scores have small positive effects. A 1-SD change in the E/S/G score changes the reward by -0.8%/0.8%/1.0%. Including ESG scores improves the model's performance in terms of mean KLD. From this perspective, ESG does not appear to be of large economic relevance for

firms decisions beyond their materiality.²²

Table E.8 presents results on the recovered reward function with a focus on the role of greenhouse gas (GHG) emissions, and we discuss these results in detail in Appendix E. The results suggest that firms prioritize profitability but also incorporate GHG emissions into their decision-making. Scope 1 and Scope 2 emissions are not penalized while only Scope 3 emissions diminish corporate value.

Tables E.9 and E.10 investigate how the corporate recovery theorem performs when we specify the actions to be two-dimensional choices as combinations of investment rate and net debt to EBITDA. Investment needs to be financed through, for example, debt or equity. As such, one would expect that managers take investment and leverage decisions jointly, rather than implement separate investment and financial policies. Overall, our main results remain unchanged. Managers still act as if investment were less costly than prescribed by theory, with shadow cost coefficients ranging from 0.42 to 0.60. The experimentation parameter is now larger, ranging between 0.19 and 0.35. The alignment between the recovered value and Total q goes down when we consider joint investment and leverage decisions, compared to investment decisions alone. This suggests that managers are less aligned with shareholders when it comes to making leverage decisions than investment decisions.

Last, Tables E.9 and E.10 also explore the robustness to variations in the estimation approach for choice and transition probabilities and the state and action variables. Overall, the results remain unchanged.

6 Conclusion

This paper employs inverse reinforcement learning (IRL) to recover the reward functions that managers implicitly use when making investment and financial policy decisions. We first establish a corporate recovery theorem that provides conditions for the unique identification of the reward function, and then we apply our method to a comprehensive dataset of U.S. public corporations spanning 1975–2021, which includes detailed financial accounting, measures of intangible capital, managerial compensation, behavioral traits, and ESG scores.

Our empirical findings document several novel facts. First, we show that managers’ reward functions can be non-parametrically recovered from observed behavior. The recovered estimates reveal that managers

²²Bonnefon et al. (2022) find that investors buy shares of companies whose (social) values align with their own. In the extreme case, this would imply that the manager and investors are always aligned (at least when it comes to social values) irrespective of whether the manager places positive, zero, or negative weight on the social score. Iliewa et al. (2024) suggest an (im-)morality-payoff associated with certain corporate actions.

act as if they underestimate the shadow cost of investment, with the estimates ranging from 45 to 65 cents on the dollar. Although incorporating financial frictions and intangible intensity improve model fit, they do not fully resolve this corporate recovery puzzle. Moreover, managerial compensation policies—particularly deferred compensation and ownership stakes—along with social and governance considerations significantly influence the recovered reward function. In contrast, better environmental scores have negative impact on managers’ rewards.

Importantly, we find that the de-facto alignment between managers’ reward functions and shareholder-value maximization is imperfect and declines with model complexity. These results suggest that machine-run firms would perform poorly since manager-shareholder alignment is imperfect and better-fitting models would imply even larger misalignment. Hence, there is a trade-off between model structure and alignment, which is novel compared to known ML/AI biases.

We show the misalignment is driven primarily by a discrepancy between firm valuation and managerial value in distressed, low firm size states, rather than by differences in incentive alignment across varying profitability levels. In the later period of our sample, the correlation between the recovered value function and Total q can exceed 80%, suggesting that interpretable ML/AI methods, such as IRL, could be used to debias managerial decision-making.

Overall, our study contributes to the literature by demonstrating that interpretable and generalizable machine-learning techniques can be applied in a finance to uncover de-facto managerial objectives and other reward structures. These insights have implications for automating corporate governance and understanding the economic frictions underlying financial decision-making.

References

- Abel, Andrew B. and Janice C. Eberly (1994) “A Unified Model of Investment Under Uncertainty,” *The American Economic Review*, 84 (5), 1367–1384.
- Baron, David P. (2007) “Corporate Social Responsibility and Social Entrepreneurship,” *Journal of Economics & Management Strategy*, 16 (3), 683–717.
- Ben-David, Itzhak and Alex Chinco (2024) “Modeling Managers As EPS Maximizers,” Working paper.
- Ben-David, Itzhak, John R. Graham, and Campbell R. Harvey (2013) “Managerial Miscalibration,” *Quarterly Journal of Economics*, 128 (4), 1547–1584.
- Bena, Jan and Elena Simintzi (2023) “Machines could not compete with Chinese labor: Evidence from U.S. firms’ innovation,” Working Paper, University of British Columbia.
- Blanchard, Olivier J., Florencio Lopez-de Silanes, and Andrei Shleifer (1994) “What do firms do with cash windfalls?” *Journal of Financial Economics*, 36 (3), 337–360.
- Bolton, Patrick, Hui Chen, and Neng Wang (2011) “A Unified Theory of Tobin’s Q, Corporate Investment, Financing, and Risk Management,” *Journal of Finance*, 66 (5), 1545–1578.
- Bonnefon, Jean-François, Augustin Landier, Parinita R. Sastry, and David Thesmar (2022) “The Moral Preferences of Investors: Experimental Evidence,” NBER Working Papers 29647.
- Bénabou, Roland and Jean Tirole (2010) “Individual and Corporate Social Responsibility,” *Economica*, 77 (305), 1–19.
- Campello, Murillo, Lin William Cong, and Luofeng Zhou (2024) “AlphaManager: A Data-Driven-Robust-Control Approach to Corporate Finance,” April, Working paper.
- Cao, Haoyang, Samuel Cohen, and Lukasz Szpruch (2021) “Identifiability in inverse reinforcement learning,” in *Advances in Neural Information Processing Systems*, 34.
- Chen, Hui, Yuhan Cheng, Yanchu Liu, and Ke Tang (2023) “Teaching Economics to the Machines,” Working paper.

- Chen, Hui, Ali Kakhbod, Maziar Kazemi, and Hao Xing (2024) “Process Intangibles and Agency Conflicts,” *Working Paper*.
- Cho, Myeong-Hyeon (1998) “Ownership structure, investment, and the corporate value: an empirical analysis,” *Journal of Financial Economics*, 47 (1), 103–121.
- Cronqvist, Henrik and Rüdiger Fahlenbrach (2009) “Large shareholders and corporate policies,” *Review of Financial Studies*, 22 (10), 3941–3976.
- DeMarzo, Peter M. and Roni Fishman (2007) “Agency and Optimal Investment Dynamics,” *Review of Financial Studies*, 20 (1), 151–188.
- DeMarzo, Peter M., Roni Fishman, Zhiguo He, and Neng Wang (2012) “Dynamic Agency and the Q Theory of Investment,” *Journal of Finance*, 67 (6), 2295–2340.
- Elhauge, Einer (2005) “Sacrificing Corporate Profits in the Public Interest,” *New York University Law Review*, 80, 733–869.
- Ferreira, Miguel A. and Pedro Matos (2008) “The colors of investors’ money: The role of institutional investors around the world,” *Journal of Financial Economics*, 88 (3), 499–533.
- Freeman, R.E. (1984) *Strategic Management: A Stakeholder Approach*, Business and Public Policy Series.
- Friedman, Milton (1970) “The social responsibility of business is to increase its profits,” *The New York Times Magazine*, 122–126.
- Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther (2022) “Predictably Unequal? The Effects of Machine Learning on Credit Markets,” *The Journal of Finance*, 77 (1), 5–47.
- Graff Zivin, Joshua and Arthur Small (2005) “A Modigliani-Miller Theory of Altruistic Corporate Social Responsibility,” *The B.E. Journal of Economic Analysis & Policy*, 5 (1), 1–19.
- Graham, John R. (1996) “Debt and the Marginal Tax Rate,” *Journal of Financial Economics*, 41 (1), 41–73.
- (2022) “Presidential Address: Corporate Finance and Reality,” *The Journal of Finance*, 77 (4), 1975–2049.

- Hart, Oliver and Luigi Zingales (2017) “Companies Should Maximize Shareholder Welfare Not Market Value,” *Journal of Law, Finance, and Accounting*, 2 (2), 247–275.
- (2022) “The New Corporate Governance,” *The University of Chicago Business Law Review*, 1 (1), 195–216.
- Hayashi, Fumio (1982) “Tobin’s Marginal q and Average q: A Neoclassical Interpretation,” *Econometrica*, 50 (1), 213–224.
- Iliewa, Zwetelina, Elisabeth Kempf, and Oliver Spalt (2024) “Corporate Actions as Moral Issues,” Working paper.
- Jensen, Michael C. (1986) “Agency costs of free cash flow, corporate finance, and takeovers,” *American Economic Review*, 76 (2), 323–329.
- Jensen, Michael C. and William H. Meckling (1976) “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure,” *Journal of Financial Economics*, 3 (4), 305–360.
- Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang (2024) “ChatGPT and Corporate Policies,” Chicago Booth Research Paper No. 23-15, Fama-Miller Working Paper 2023-103, University of Chicago, Becker Friedman Institute for Economics.
- Malmendier, Ulrike and Geoffrey Tate (2005) “CEO Overconfidence and Corporate Investment,” *Journal of Finance*, 60 (6), 2661–2700.
- Myers, Stewart C. and Nicholas S. Majluf (1984) “Corporate financing and investment decisions when firms have information that investors do not have,” *Journal of Financial Economics*, 13 (2), 187–221.
- Ng, Andrew Y., Daishi Harada, and Stuart Russell (1999) “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping,” in *Proceedings of the Sixteenth International Conference on Machine Learning*, 278–287: Morgan Kaufmann.
- Peters, Ryan H. and Lucian A. Taylor (2017) “Intangible Capital and the Investment-q Relation,” *Journal of Financial Economics*, 123 (2), 251–272.
- Pástor, Luboš, Robert F. Stambaugh, and Lucian A. Taylor (2022) “Dissecting Green Returns,” *Journal of Financial Economics*, 146 (2), 403–424.

- Rajan, Raghuram, Pietro Ramella, and Luigi Zingales (2024) “What Purpose Do Corporations Purport? Evidence from Letters to Shareholders,” Working paper.
- Richardson, Scott (2006) “Over-investment of free cash flow,” *Review of Accounting Studies*, 11, 159–189.
- Rolland, Paul, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher (2022) “Identifiability and Generalizability from Multiple Experts in Inverse Reinforcement Learning,” in *Advances in Neural Information Processing Systems*, 35.
- Russell, Stuart (1998) “Learning agents for uncertain environments (extended abstract),” in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT’ 98*.
- Samuelson, P. A. (1938) “A Note on the Pure Theory of Consumer’s Behaviour,” *Economica*, 5 (17), 61–71.
- Shleifer, Andrei and Robert W. Vishny (1989) “Management entrenchment: The case of manager-specific investments,” *Journal of Financial Economics*, 25 (1), 123–139.
- Strebulaev, Ilya A. and Toni M. Whited (2012) “Dynamic models and structural estimation in corporate finance,” *Foundations and Trends in Finance*, 6 (1–2), 1–163.
- Stulz, René M. (1990) “Managerial discretion and optimal financing policies,” *Journal of Financial Economics*, 26 (1), 3–27.
- Sutton, Richard S. and Andrew G. Barto (2018) *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 2nd edition.
- Tauchen, George (1986) “Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions,” *Economics Letters*, 20 (2), 177–181.
- Tobin, James (1969) “A General Equilibrium Approach to Monetary Theory,” *Journal of Money, Credit and Banking*, 1 (1), 15–29.
- Ward, Colin (2023) “Agency in Intangibles,” Working Paper, University of Alberta.

APPENDIX

Appendix A Proofs and Analytic Expressions

Preliminary derivations. With X defined in (11), let Ω be the matrix that is composed of four sub-matrices:

$$\Omega \equiv X^\top X = \begin{pmatrix} A & C^\top \\ C & F \end{pmatrix}, \quad (\text{A.1})$$

with submatrices corresponding to the quadratic form $X^\top X$:

$$A = \begin{pmatrix} 2X_i^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}, \quad C = (f^\top X_i \quad 0), \quad F = (f^\top f). \quad (\text{A.2})$$

A and F are square invertible and C is of conformable dimension. Let $B \equiv -X_j^\top X_i$ and $D \equiv X_j^\top X_j$. Denote the Schur complements of Ω with respect to A and F by Ω/A and Ω/F , respectively. If A is nonsingular, the Schur complement of Ω with respect to A is defined as $\Omega/A = F - CA^{-1}C^\top$. If F is nonsingular, the Schur complement of Ω with respect to F is defined as $\Omega/F = A - C^\top F^{-1}C$. Then

$$\Omega^{-1} = \begin{pmatrix} (\Omega/F)^{-1} & -(\Omega/F)^{-1}C^\top F^{-1} \\ -(\Omega/A)^{-1}CA^{-1} & (\Omega/A)^{-1} \end{pmatrix}. \quad (\text{A.3})$$

To compute A^{-1} , the Schur complement of the block D in A equals

$$A/D = 2X_i^\top X_i - X_i^\top X_j (X_j^\top X_j)^{-1} X_j^\top X_i. \quad (\text{A.4})$$

One obtains

$$A^{-1} = \begin{pmatrix} (A/D)^{-1} & -(A/D)^{-1}B^\top D^{-1} \\ -D^{-1}B(A/D)^{-1} & D^{-1} + D^{-1}B(A/D)^{-1}B^\top D^{-1} \end{pmatrix}. \quad (\text{A.5})$$

Define the projection matrix under the norm defined by D as

$$\begin{aligned} \mathbf{P}_{ij} &= X_i (A/D)^{-1} X_i^\top, \\ \mathbf{M}_{ij} &= I - X_i (A/D)^{-1} X_i^\top. \end{aligned} \quad (\text{A.6})$$

The Schur complement Ω/A then equals

$$\begin{aligned} \Omega/A &= (f^\top f) - (f^\top X_i \quad 0) A^{-1} \begin{pmatrix} X_i^\top f \\ 0 \end{pmatrix} \\ &= (f^\top f) - (f^\top X_i (A/D)^{-1} X_i^\top f) \\ &= f^\top (I - X_i (A/D)^{-1} X_i^\top) f \\ &= f^\top \mathbf{M}_{ij} f, \end{aligned} \quad (\text{A.7})$$

with inverse

$$(\Omega/A)^{-1} = (f^\top \mathbf{M}_{ij} f)^{-1}. \quad (\text{A.8})$$

The Schur complement Ω/F equals

$$\begin{aligned} \Omega/F &= \begin{pmatrix} 2X_i^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix} - \begin{pmatrix} X_i^\top f \\ 0 \end{pmatrix} (f^\top f)^{-1} (f^\top X_i \quad 0) \\ &= \begin{pmatrix} 2X_i^\top X_i - X_i^\top f (f^\top f)^{-1} f^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}, \end{aligned} \quad (\text{A.9})$$

and the Schur complement $(\Omega/F)/D$ equals

$$\begin{aligned}
(\Omega/F)/D &= 2X_i^\top X_i - X_i^\top f(f^\top f)^{-1} f^\top X_i - X_i^\top X_j (X_j^\top X_j)^{-1} X_j^\top X_i \\
&= X_i^\top (I - f(f^\top f)^{-1} f^\top) X_i + X_i^\top (I - X_j (X_j^\top X_j)^{-1} X_j^\top) X_i \\
&= X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i.
\end{aligned} \tag{A.10}$$

The inverse $(\Omega/F)^{-1}$ can now be written

$$= \begin{pmatrix} ((\Omega/F)/D)^{-1} & -((\Omega/F)/D)^{-1} B^\top D^{-1} \\ -D^{-1} B ((\Omega/F)/D)^{-1} & D^{-1} + D^{-1} B ((\Omega/F)/D)^{-1} B^\top D^{-1} \end{pmatrix} = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \tag{A.11}$$

with

$$\begin{aligned}
E_{11} &= (X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i)^{-1}, \\
E_{12} &= (X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i)^{-1} X_i^\top X_j (X_j^\top X_j)^{-1}, \\
E_{21} &= (X_j^\top X_j)^{-1} X_j^\top X_i (X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i)^{-1}, \\
E_{22} &= (X_j^\top X_j)^{-1} + (X_j^\top X_j)^{-1} X_j^\top X_i (X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i)^{-1} X_i^\top X_j (X_j^\top X_j)^{-1}.
\end{aligned} \tag{A.12}$$

The inverse Ω^{-1} can now be written explicitly as

$$\begin{aligned}
&\begin{pmatrix} (\Omega/F)^{-1} & -(\Omega/F)^{-1} C^\top F^{-1} \\ -(\Omega/A)^{-1} C A^{-1} & (\Omega/A)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} ((\Omega/F)/D)^{-1} & -((\Omega/F)/D)^{-1} B^\top D^{-1} & -((\Omega/F)/D)^{-1} X_i^\top f (f^\top f)^{-1} \\ -D^{-1} B ((\Omega/F)/D)^{-1} & D^{-1} + D^{-1} B ((\Omega/F)/D)^{-1} B^\top D^{-1} & D^{-1} B ((\Omega/F)/D)^{-1} X_i^\top f (f^\top f)^{-1} \\ -(\Omega/A)^{-1} f^\top X_i (A/D)^{-1} & (\Omega/A)^{-1} f^\top X_i (A/D)^{-1} B^\top D^{-1} & (\Omega/A)^{-1} \end{pmatrix}.
\end{aligned} \tag{A.13}$$

We get the analytical expressions

$$\begin{pmatrix} v_1 \\ v_2 \\ \theta \end{pmatrix} = \Omega^{-1} X^\top \begin{pmatrix} y_i - y_j \\ y_i \end{pmatrix} = \Omega^{-1} \begin{pmatrix} X_i^\top (2y_i - y_j) \\ -X_j^\top (y_i - y_j) \\ f^\top y_i \end{pmatrix} = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix}, \tag{A.14}$$

with

$$\begin{aligned}
E_1 &= ((\Omega/F)/D)^{-1} X_i^\top (2y_i - y_j) + ((\Omega/F)/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) \\
&\quad - ((\Omega/F)/D)^{-1} X_i^\top f (f^\top f)^{-1} f^\top y_i, \\
E_2 &= -D^{-1} B ((\Omega/F)/D)^{-1} X_i^\top (2y_i - y_j) - D^{-1} X_j^\top (y_i - y_j) \\
&\quad - D^{-1} B ((\Omega/F)/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) + D^{-1} B ((\Omega/F)/D)^{-1} X_i^\top f (f^\top f)^{-1} f^\top y_i, \\
E_3 &= -(\Omega/A)^{-1} f^\top X_i (A/D)^{-1} X_i^\top (2y_i - y_j) \\
&\quad - (\Omega/A)^{-1} f^\top X_i (A/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) + (\Omega/A)^{-1} f^\top y_i.
\end{aligned} \tag{A.15}$$

We now obtain our main results:

$$\begin{aligned}
v_1 &= ((\Omega/F)/D)^{-1} X_i^\top (I - X_j (X_j^\top X_j)^{-1} X_j^\top) y_i \\
&\quad + ((\Omega/F)/D)^{-1} X_i^\top (I - f (f^\top f)^{-1} f^\top) y_i \\
&\quad - ((\Omega/F)/D)^{-1} X_i^\top (I + X_j (X_j^\top X_j)^{-1} X_j^\top) y_j \\
&= (X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i)^{-1} X_i^\top [(\mathbf{M}_f + \mathbf{M}_j) y_i - \mathbf{M}_j y_j],
\end{aligned} \tag{A.16}$$

and

$$\begin{aligned}
v_2 &= -D^{-1} B ((\Omega/F)/D)^{-1} X_i^\top (I - X_j (X_j^\top X_j)^{-1} X_j^\top) (y_i - y_j) - D^{-1} X_j^\top (y_i - y_j) \\
&\quad - D^{-1} B ((\Omega/F)/D)^{-1} X_i^\top (I - f (f^\top f)^{-1} f^\top) y_i \\
&= (X_j^\top X_j)^{-1} X_j^\top [\mathbf{Q}_{ij} \mathbf{M}_f y_i - (I - \mathbf{Q}_{ij} \mathbf{M}_j) (y_i - y_j)],
\end{aligned} \tag{A.17}$$

with projection matrix

$$\mathbf{Q}_{ij} = X_i((\Omega/F)/D)^{-1}X_i^\top, \quad (\text{A.18})$$

and

$$\begin{aligned} \theta &= (\Omega/A)^{-1}f^\top \left(I - X_i(A/D)^{-1}X_i^\top \right) y_i \\ &\quad - (\Omega/A)^{-1}f^\top X_i(A/D)^{-1}X_i^\top (I + X_j(X_j^\top X_j)^{-1}X_j^\top) (y_i - y_j) \\ &= (\Omega/A)^{-1}f^\top \mathbf{M}_{ij}y_i + (\Omega/A)^{-1}f^\top \mathbf{P}_{ij}\mathbf{M}_j(y_j - y_i). \end{aligned} \quad (\text{A.19})$$

Proof of Theorem 3. This is an extension of Theorem 7 in [Rolland et al. \(2022\)](#) applied to our setting. Consider the linear reward recovery problem. For any agent $i = 1, \dots, I$,

$$y_i = X_i v_i + f\theta + \varepsilon_i, \quad (\text{A.20})$$

where $y_i = \ln \pi_i^*(a|s)$ is defined in (5), $X_i = -(I - \gamma_i T_i)$ is defined in (6), and ε_i is iid noise that is independent of X_i and the features f . In matrix notation, the conditions for agents (i, j) can be written:

$$\underbrace{\begin{pmatrix} y_i - y_j \\ y_i \end{pmatrix}}_y = \underbrace{\begin{pmatrix} X_i & -X_j & 0 \\ X_i & 0 & f \end{pmatrix}}_X \cdot \begin{pmatrix} v_i \\ v_j \\ \theta \end{pmatrix} + \varepsilon. \quad (\text{A.21})$$

This highlights the joint influence of agent-specific value functions v_i , v_j , and the shared parameters θ . The next step involves recovering v_i , v_j , and θ by solving the system of equations. After pre-multiplying by $\Omega^{-1}X^\top$, so long as Ω is invertible, the solution is given by

$$\begin{pmatrix} v_i \\ v_j \\ \theta \end{pmatrix} = \underbrace{\begin{pmatrix} (\Omega/F)^{-1} & -(\Omega/F)^{-1}C^\top F^{-1} \\ -(\Omega/A)^{-1}CA^{-1} & (\Omega/A)^{-1} \end{pmatrix}}_{\Omega^{-1}} X^\top (y - \varepsilon). \quad (\text{A.22})$$

This structure reflects a decomposition of the reward function into agent-specific and shared components. The matrix inversion and multiplication by Ω^{-1} means that we are solving a system of linear equations, using regularization (implicit in Ω) to ensure numerical stability.

Analytic expressions. To obtain analytic expressions based on the results in Theorem 3, note that the feature weights θ and the value functions $v_i, i = 1, \dots, I$, can be expressed explicitly in terms of the features f , X_i , y_i , and $y_j, j \neq i$. The projections and residual makers corresponding to X_i and, respectively, f from matrix X are given by $\mathbf{M}_i$ and $\mathbf{M}_f = I - f(f^\top f)^{-1}f^\top$, which is the residual maker for the projection on f . To compute A^{-1} , the Schur complement of the block $D \equiv X_j^\top X_j$ in A equals

$$A/D = 2X_i^\top X_i - X_i^\top X_j(X_j^\top X_j)^{-1}X_j^\top X_i. \quad (\text{A.23})$$

Expression (A.23) defines a norm that can be used to define an alternate projection on X_i . The positive definite matrix D defines an inner product $\langle x, y \rangle_D = y^\top D x$, and the projection $\mathbf{P}_{ij}$ relative to D is given by $\mathbf{P}_{ij}x = \operatorname{argmin}_{y \in \operatorname{range}(X_i)} \|x - y\|_D^2$. Then under this norm:

$$\begin{aligned} \text{Projection relative to } \|\cdot\|_D: \quad \mathbf{P}_{ij} &= X_i(A/D)^{-1}X_i^\top, \\ \text{Residuals relative to } \|\cdot\|_D: \quad \mathbf{M}_{ij} &= I - X_i(A/D)^{-1}X_i^\top. \end{aligned} \quad (\text{A.24})$$

So long as A and F are square invertible and C is of conformable dimension, the Schur complements of Ω with respect to A and, respectively, F are equal to

$$\begin{aligned} \Omega/A &= f^\top \mathbf{M}_{ij}f, \\ \Omega/F &= \begin{pmatrix} X_i^\top(I + \mathbf{M}_f)X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}. \end{aligned} \quad (\text{A.25})$$

The double Schur complement of Ω with respect to F and D will also be important and equals $(\Omega/F)/D = X_i^\top (\mathbf{M}_f + \mathbf{M}_j) X_i$. It captures a double projection and reports the residuals corresponding to both subspaces generated by f and X_j , respectively. We are now ready to state expressions for θ and v_i .

Proposition 2. *The reward function $u(s, a) = f(s, a) \cdot \theta$ can be recovered from*

$$\theta = (f^\top \mathbf{M}_{ij} f)^{-1} f^\top (\mathbf{M}_{ij} y_i + \mathbf{P}_{ij} \mathbf{M}_j (y_j - y_i)). \quad (\text{A.26})$$

Up to affine transformations represented by constants $(a_v, b_v) \in \mathbb{R}$, the agents' value functions for $i = 1, \dots, I$ can be recovered by

$$v_i = a_v + b_v (X_i^\top (\mathbf{M}_j + \mathbf{M}_f) X_i)^{-1} X_i^\top ((\mathbf{M}_j + \mathbf{M}_f) y_i - \mathbf{M}_j y_j). \quad (\text{A.27})$$

In the reward function recovery, the matrix $\mathbf{M}_{ij}$ is a residual maker matrix that orthogonalizes vectors with respect to the span of f concerning agents i and j . The term involving $\mathbf{P}_{ij}$ and $\mathbf{M}_j$ are adjustments based on the residuals from agent j 's outcomes y_j relative to i 's outcomes y_i . In the value function recovery, that is defined up to an affine transformation (scaling and shifting by a_v and b_v), the matrices $\mathbf{M}_j$ and $\mathbf{M}_f$ are both residual makers for different contexts: $\mathbf{M}_j$ with respect to agent j 's influence, and $\mathbf{M}_f$ with respect to the features f . The appearance of $\mathbf{M}_f$ is crucial as it adds an additional layer of orthogonalization regarding the features used in the model. This inclusion shows that the value function recovery takes into account not just the difference with other agents but also the specific features of the environment and the agents themselves (captured by f). This additional complexity is necessary since the agent's behavior is mediated/modulated through the features of the agent and the environment, which would otherwise confound the estimation of individual agent value functions.

Proposition 2 differs from Proposition 1 in that the recovery of v_i now involves the residual maker for the features f , $\mathbf{M}_f$, in addition to the residual maker for the second agent, $\mathbf{M}_j$. The proposition provides a method for feature importance and value function recovery in our framework. The ultimate goal is to estimate the parameters θ that describe how agents' decisions are influenced by their context and actions.

Appendix B Implementing Recovery in a Controlled Environment

Discretization. We start by creating the environment, consisting of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, real-valued features $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, coefficients θ and rewards $u(s, a) = f(s, a) \cdot \theta$. The model has two state variables, capital k and profitability shock z . For k , we create a set $\mathcal{K}$ containing n_k grid points centered around the steady state value of k , denoted k^* . We then add $(n_k - 1)/2$ grid points to either side, given by $k^* \prod_{j=1}^{(n_k-1)/2} (1 - \delta)^{\pm \frac{j}{d}}$, respectively, where d is a density parameter. For z , we create a set $\mathcal{Z}$ containing (a preferably odd number of) n_z equi-spaced grid points between $-m\sigma_\epsilon$ and $m\sigma_\epsilon$, where m represents a multiple of the unconditional standard deviation of ϵ . We then follow [Tauchen \(1986\)](#) to approximate the corresponding shock transition probabilities $P(z' | z)$. Based on these grids, we can create our set of states $\mathcal{S}$ as all possible pairs of (k, z) s.t. $k \in \mathcal{K}, z \in \mathcal{Z}$. Our action a consists in picking the next period's capital k' . Consequently, the action set $\mathcal{A}$ equals the set of discrete values for capital $\mathcal{K}$. The reward $u(s, a)$, feature matrix $f(s, a)$ and true coefficients θ^* vary across specifications and are defined as in the table above.

Experts. Next, we create two experts $i = 1, 2$ acting in this environment. Both experts share the same preference for entropy (captured by the λ parameter), but differ in their discount factor γ , as well as their subjective assessment of the unconditional standard deviation of profitability shocks $\sigma_{\epsilon,i}$ (γ_i should be different, but similar in magnitude to $\gamma = \frac{1}{1+r}$. Similarly, $\sigma_{\epsilon,i}$ should be different, but similar in magnitude to σ_ϵ). The latter part implies different shock transition probabilities $P_i(z' | z)$. These, in turn, lead to different state transition probabilities $P_i(s' | s, a)$, which we collect in a three-dimensional array $T_i(s' | s, a)$. We solve for each expert's optimal stochastic policy function $\pi_i(s)$ using an algorithm called Q-value function iteration.

Simulations and estimation. For each expert i , we simulate a trajectory of n_{steps} steps, starting from a the steady state capital $k_0 = k^*$ and a neutral profit shock $z_0 = 1$ and following $T_i(s' | s, a)$ and $\pi_i(a | s)$. We start by drawing the initial action a_0 from $\pi(a | s = (k_0, z_0))$. For each time step $t = 1, \dots, n_{steps}$, we first draw the next state s_t from $T_i(s' | s = s_{t-1}, a = a_{t-1})$ and then draw the next action a_t from $\pi_i(a | s = s_t)$. For each expert, we then estimate $\hat{T}_i(s' | s, a)$ and $\hat{\pi}_i(a | s)$ using the consistent count-based estimators

$$\hat{\pi}_{i,N}(a | s) = \frac{\#(a_{i,t} = a \text{ and } s_{i,t} = s ; t \leq N)}{\#(s_{i,t} = s ; t \leq N)} \rightarrow \pi_i(a | s) \text{ a.s. as } N \rightarrow \infty \quad (\text{B.1})$$

and

$$\hat{T}_{i,N}(s' | s, a) = \frac{\#(s_{i,t+1} = s \text{ and } s_{i,t} = s \text{ and } a_{i,t} = a ; t \leq N)}{\#(s_{i,t} = s \text{ and } a_{i,t} = a ; t \leq N)} \rightarrow T_i(s' | s, a) \text{ a.s. as } N \rightarrow \infty \quad (\text{B.2})$$

Experiments. To test how our recovery depends on the estimation of the policy and transition functions, we run 4 experiments for each specification. In Experiment 1, we recover the coefficients of the reward function based on the known policy and transition functions. In Experiment 2, we estimate the policy functions based on our simulated trajectories and recover the coefficients of the reward function based on the estimated policy functions and true transition functions. In Experiment 3, we estimate the transition functions based on our simulated trajectories and recover the coefficients of the reward function based on the true policy functions and estimated transition functions. In Experiment 4, we estimate both the policy functions and transition functions based on our simulated trajectories and recover the coefficients of the reward function based on the estimated policy and transition functions. To test how our recovery depends on the specification of features and coefficients, we repeat all four experiments for each of the specifications listed above.

Biases from irrelevant and missing features. So far, we have assumed that the set of features used in the recovery coincides with the true set of features that enter the experts' reward function, i.e., that the feature array is correctly specified. In this section, we assess the robustness of our reward recovery methodology to feature set misspecifications. We examine the consequences of discrepancies between the feature set employed for recovery and the true feature set underlying the expert's reward function. This analysis encompasses scenarios wherein the true feature set is either more extensive or more restrictive than the recovery feature set, analogous to omitted variable bias and spurious correlation, respectively, in a regression framework. Building upon the established results from Table 1, we focus on Specification 4, Experiment 4, which incorporates profitability and investment rate as features, utilizing estimated choice and transition probabilities.

Table B.1: Robustness of reward recovery in controlled environment.

Panel A contains different specifications of features and coefficients in our controlled environment and in the recovery. Panel B documents the recovered weights for different experiments in our controlled environment. Column 1 refers to the recovery experiment. π and T refer to whether we used the true or estimated policy and transition functions. We used the following parameter values: $\zeta = 0.55, \delta = 0.2, \rho = 0.7, r = 0.05, \sigma_\epsilon = 0.04, \psi_0 = 0.01, \psi_1 = 0.01, n_k = 15, d = 3, n_z = 5, m = 2, \lambda = 0.2, \sigma_\epsilon^1 = 0.035, \sigma_\epsilon^2 = 0.045, \gamma^1 = 0.94, \gamma^2 = 0.96, n_{steps} = 1,000,000, \text{seed} = 0$.

Panel A: Reward specifications in controlled environment				
Experiment	Feature Set	Features $f(s = (s_1, s_2), a = s'_1)$		True Weights (θ)
1	True	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1} \epsilon]$		$[1 \quad -1 \quad 1]$
1	Recovery	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1}]$		
2	True	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1} 0.8 \frac{a-(1-\delta)s_1}{s_1} + \epsilon]$		$[1 \quad -1 \quad 1]$
2	Recovery	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1}]$		
3	True	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1} -0.8 \frac{a-(1-\delta)s_1}{s_1} + \epsilon]$		$[1 \quad -1 \quad 1]$
3	Recovery	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1}]$		
4	True	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1}]$		$[1 \quad -1]$
4	True	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1} \epsilon]$		$[1 \quad -1 \quad 0]$
4	Recovery	$[s_2 s_1^{\zeta-1} \frac{a-(1-\delta)s_1}{s_1} \epsilon]$		
Panel B: Reward recovery in controlled environment				
Experiment	π	T	True Weights (θ)	Recovered Weights ($\hat{\theta}$)
1	Estimated	Estimated	$\{1, -1, 1\}$	$\{0.987, -1.003\}$
2	Estimated	Estimated	$\{1, -1, 1\}$	$\{0.995, -1.007\}$
3	Estimated	Estimated	$\{1, -1, 1\}$	$\{1.002, -1.015\}$
4	Estimated	Estimated	$\{1, -1, 0\}$	$\{0.991, -0.996, 0.000\}$

Table B.1 details in Panel A the feature sets used for simulation and recovery. Across all experiments, profitability ($s_2 s_1^{\zeta}$) and net investment ($a - (1 - \delta)s_1$) constitute the core feature set. Variations in feature sets arise from the inclusion or exclusion of a third feature and the manipulation of its correlation structure. In Experiment 1, a third feature comprising uncorrelated Gaussian noise ($\epsilon \sim N(0, 1)$), with a true weight of 1, is included in the true feature set but omitted from the recovery process. Experiments 2 and 3 introduce a third feature, with a true weight of 1, that is highly positively (0.8) or negatively (-0.8) correlated with the investment rate, respectively. These features are subsequently omitted during recovery. In Experiment 4, an irrelevant feature consisting of uncorrelated Gaussian noise is included in the recovery feature set, despite its absence from the true feature set (or equivalently, a true weight of zero). We expect the recovered weights to be unbiased in experiments 1 and 4, and the recovered weight on the investment rate to be upward (downward) biased in experiment 2 (3), akin to the bias resulting from omitting a highly positively (negatively) correlated variable in a regression framework. Our results are summarized in Panel B of Table B.1. The procedure recovers weights that are very close to the true ones across all four experiments. These results suggest that neither omitted variable bias nor spurious correlation are material concerns due to the non-linear nature of the recovery.

Appendix C Data Definitions and Summary Statistics

Table C.1: Variable Definitions.

This table contains the definitions and sources of all variables used in our empirical analysis.

Variable Name	Definition	Formula
U.S. Bureau of Economic Analysis (BEA)		
CPI_{Adj}	CPI adjustment factor (base year=2000)	$\frac{CPI_{2000}}{CPI_t}$
Federal Reserve Economic Data (FRED)		
R_f	Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity	$DSG10$
Compustat		
Size	Natural Logarithm of (real) net property, plant and equipment (ppent)	$\ln(ppent_t \times CPI_{Adj,t})$
Profitability	Earnings before extraordinary items (ib) plus depreciation and amortization (dp); divided by lagged net property, plant and equipment (ppent)	$\frac{ib_t + dp_t}{ppent_{t-1}}$
Investment Rate	Change in net property, plant and equipment (ppent) plus depreciation and amortization (dp) minus amortization (am); divided by lagged net property, plant and equipment (ppent). Values above 1 are treated as missing.	$\frac{ppent_t - ppent_{t-1} + dp_t - am_t}{ppent_{t-1}}$
Net Debt / EBITDA	Debt in current liabilities (d1c) plus long-term debt (d1tt) minus cash and cash equivalents (che); divided by EBITDA (ebitda) if EBITDA is positive. Treated as missing if EBITDA is negative.	$\frac{d1c_t + d1tt_t - che_t}{ebitda_t}$
Book Equity	Stockholders' equity (seq) plus deferred taxes and investment tax credit (txditc) minus the book value of preferred stock (preference order if available: redemption value (pstkrv) > liquidation value (pstk1) > carrying value (pstk) > 0). Negative values are treated as missing.	$seq_t + txditc_t - coalesce(pstkrv, pstk1, pstk, 0)$
Tobin's q	Total Assets (at) minus BookEquity plus common shares outstanding (csho) times share price (prcc_f); divided by total assets (at)	$\frac{at_t - BookEquity_t + csho_t \times prcc_f_t}{at_t}$
Tangibility	Net property, plant and equipment (ppent); divided by total assets (at)	$\frac{ppent_t}{at_t}$
Cash Holdings	Cash and cash equivalents (che); divided by total assets (at)	$\frac{che_t}{at_t}$
Book Leverage	Debt in current liabilities (d1c) plus long-term debt (d1tt); divided by total assets (at). Negative values are treated as missing.	$\frac{d1c_t + d1tt_t}{at_t}$
Net Book Leverage	Debt in current liabilities (d1c) plus long-term debt (d1tt) minus cash and cash equivalents (che); divided by total assets (at)	$\frac{d1c_t + d1tt_t - che_t}{at_t}$
Altman Z-score	1.2 times the ratio of working capital (ac-lct) to total assets (at) plus 1.4 times the ratio of retained earnings (re) to total assets (at) plus 3.3 times the ratio of earnings before interest and taxes (ni+xint+txt) to total assets (at) plus 0.6 times the ratio of market value of equity (prcc_f * csho) to total liabilities (lt) plus 1.0 times the ratio of sales (sale) to total assets (at)	$1.2 \frac{act_t - lct_t}{at_t} + 1.4 \frac{re_t}{at_t} + 3.3 \frac{ni_t + xint_t + txt_t}{at_t} + 0.6 \frac{prcc_f_t * csho_t}{lt_t} + 1.0 \frac{sale_t}{at_t}$

(Table C.1 continued)

Variable Name	Definition	Formula
Employee deferred compensation	Employee deferred compensation (stkco) divided by lagged gross property, plant and equipment (ppeg)	$\frac{stkco_t}{ppeg_{t-1}}$
Peters & Taylor (2017) Data (Downloaded from WRDS)		
Total Capital (K_{tot})	Gross property, plant and equipment (ppeg) plus intangible capital (K_int)	$ppeg_t + K_{int,t}$
ln(Total Capital)	Natural Logarithm of (real) total capital (K_tot)	$\ln(K_{tot,t} \times CPI_{Adj,t})$
Phys. Inv. Rate (I_{phy})	Change in net property, plant and equipment (ppent) plus depreciation and amortization (dp) minus amortization (am); divided by lagged total capital (K_{tot}).	$\frac{ppent_t - ppent_{t-1} + dp_t - am_t}{K_{tot,t-1}}$
Int. Inv. Rate (I_{int})	Research and development expense (xrd) plus 0.3 times the selling, general and administrative expense (xsga) plus 0.15 times lagged intangible capital (K_{int}); divided by lagged total capital (K_{tot})	$\frac{xrd + 0.3 \times xsga + 0.15 \times K_{int,t-1}}{K_{tot,t-1}}$
Total Inv. Rate (I_{tot})	Physical investment rate (I_{phy}) plus intangible investment rate (I_{int})	$I_{phy,t} + I_{int,t}$
Alt. Profitability	Alternative measure of profitability. Earnings before extraordinary items (ib) plus depreciation and amortization (dp); divided by lagged total capital (K_{tot}) plus tax-adjusted intangible investment rate. The marginal tax rate (κ) is taken from Graham (1996).	$\frac{ib_t + dp_t}{K_{tot,t-1}} + (1 - \kappa)I_{int,t}$
Intangible intensity	Intangible capital (K_int) divided by total capital (K_tot)	$K_{int,t} / K_{tot,t}$
Knowledge intensity	Knowledge capital (K_know) divided by intangible capital (K_int)	$K_{know,t} / K_{int,t}$
Organizational intensity	Organizational capital (K_org) divided by intangible capital (K_int)	$K_{org,t} / K_{int,t}$
Bena and Simintzi (2023)		
Process intensity	Sum over patents of number of process claims per patent (cl_process_N) divided by sum of number of claims per patent (cl_N)	$\frac{\sum_i cl_process_N_{i,t}}{\sum_i cl_N_{i,t}}$
Process intangibility	Sum over patents of number of process claims per patent (cl_process_N) divided by sum of number of claims per patent (cl_N); multiplied by the ratio of intangible capital (K_int) to total capital (K_tot)	$\frac{\sum_i cl_process_N_{i,t}}{\sum_i cl_N_{i,t}} \frac{K_{int,t}}{K_{tot,t}}$
Non-process intangibility	Ratio of intangible capital (K_int) to total capital (K_tot) minus process intangibility	$\frac{K_{int,t}}{K_{tot,t}} (1 - \frac{\sum_i cl_process_N_{i,t}}{\sum_i cl_N_{i,t}})$
Execucomp and Equilar		
CEO Bonus	Bonus (bonus) received by the CEO (ceoann =1); divided by total assets (at) (in %)	$\frac{bonus_t}{at_t}$
CEO Ownership	Shares excluding options (shown_excl_opts) owned by the CEO; divided by common shares outstanding (csho)	$\frac{shown_excl_opts_t}{csho_t}$
Managerial Bonus	Sum of the bonuses (bonus) received by the top 5 managers ranked by total current compensation (total_curr) in a given year; divided by total assets (at) (in %)	$\frac{\sum_{i=1}^5 bonus_{i,t}}{at_t}$
Mana. Ownership	Sum of the shares excluding options (shown_excl_opts) owned by the top 5 managers; divided by number of shares outstanding (csho)	$\frac{\sum_{i=1}^5 shown_excl_opts_{i,t}}{csho_t}$

(Table C.1 continued)

Variable Name	Definition	Formula
Total compensation	Sum of total compensation (tdc1) received by the top 5 managers ranked by total current compensation (total_curr) in a given year; divided by lagged gross property, plant and equipment (ppeg)	$\frac{\sum_{i=1}^5 tdc1_{i,t}}{ppeg_{t-1}}$
Manag. deferred comp.	Sum of stock awards (stock_awards) and option awards (option_awards) received by the top 5 managers ranked by total current compensation (total_curr) in a given year; divided by lagged gross property, plant and equipment (ppeg)	$\frac{\sum_{i=1}^5 stock_awards_{i,t} + option_awards_{i,t}}{ppeg_{t-1}}$
Fraction deferred comp.	Sum of stock awards (stock_awards) and option awards (option_awards) received by the top 5 managers ranked by total current compensation (total_curr) in a given year; sum of the total compensation (tdc1) received by the top 5 managers ranked by total current compensation (total_curr) in a given year	$\frac{\sum_{i=1}^5 stock_awards_{i,t} + option_awards_{i,t}}{\sum_{i=1}^5 tdc1_{i,t}}$
CEO age	CEO age (age) in years	age_t
CEO tenure	Difference between current year (fyear) and the year of CEO appointment (becameceo)	$fyear_t - becameceo$
CEO overconfidence	Dummy variable equal to 1 if for firm i and CEO j the following two conditions are met: number of exercisable unexercised options (options_exercisable) is strictly positive and the residual maturity of unexercised exercisable options (options_expiration_date) is less or equal to 1 year	1 if $options_exercisable_{ijt} > 0$ and $ options_expiration_date_{ijt} - t \leq 1$
CEO overconfidence alt.	Dummy variable equal to 1 if there are two occurrences of a CEO failing to exercise an option with five years of maturity remaining that is at least 67% in the money	There is more than one date $t_m, m \in 1, 2, \dots$ for which $options_exercisable_{ijt_m} > 0$ and 4.5 years $\leq options_expiration_date_{ijt_m} - t_m \leq 5.5$ years and $\frac{price_{it_m} - options_exercise_price_{ijt_m}}{options_exercise_price_{ijt_m}} \geq 67\%$
Trucost		
Scope 1 intensity	Scope 1 emissions (scope1) divided by lagged sales (sale) divided by 1,000	$scope1_t / sale_{t-1} / 1000$
Scope 2 intensity	Scope 2 emissions (scope2) divided by lagged sales (sale) divided by 1,000	$scope2_t / sale_{t-1} / 1000$
Scope 3 intensity	Scope 3 emissions (scope3) divided by lagged sales (sale) divided by 1,000	$scope3_t / sale_{t-1} / 1000$
MSCI ESG Ratings		
The environmental score (E_score), social score (S_score), governance score (G_score), weighted-average ESG score (ESG_score) and industry-adjusted ESG score (ESG_score_adj) are taken directly from MSCI without adjustment.		

Table C.2: Summary Statistics.

This table contains summary statistics for our variables across different samples.

	N	Mean	S.D.	20%	40%	50%	60%	80%
Panel A: Compustat (1975-2021)								
Size	121,773	4.66	2.24	2.62	3.93	4.52	5.14	6.65
Profitability	121,773	0.51	0.91	0.13	0.26	0.32	0.41	0.72
Investment Rate	121,773	0.24	0.23	0.08	0.16	0.20	0.25	0.39
Net Debt to EBITDA	121,773	1.78	5.60	-0.51	0.64	1.15	1.70	3.42
Cash Holdings	121,772	0.12	0.13	0.02	0.05	0.07	0.10	0.19
Book Leverage	120,488	0.26	0.20	0.06	0.19	0.24	0.29	0.41
Net Book Leverage	121,772	0.15	0.30	-0.08	0.10	0.16	0.23	0.37
Tangibility	121,772	0.34	0.22	0.14	0.24	0.29	0.35	0.54
Tobin's q	112,151	1.56	1.05	0.92	1.12	1.24	1.40	1.96
Altman Z-score	116,857	3.88	3.47	1.82	2.79	3.24	3.73	5.24
Employee deferred comp.	44,552	0.04	0.11	0.00	0.01	0.01	0.02	0.04
Panel B: Peters & Taylor (1975-2021)								
ln(Total Capital)	121,773	6.06	2.07	4.18	5.37	5.92	6.48	7.88
Alternative Profitability	121,773	0.21	0.13	0.12	0.17	0.20	0.22	0.29
Physical Investment Rate	121,773	0.07	0.08	0.02	0.04	0.05	0.06	0.11
Intangible Inv. Rate	121,773	0.16	0.10	0.05	0.12	0.15	0.18	0.24
Total Investment Rate	121,773	0.23	0.13	0.13	0.19	0.22	0.25	0.32
Total q	116,617	0.88	1.29	0.10	0.38	0.52	0.68	1.24
Intangible intensity	121,773	0.44	0.26	0.17	0.38	0.46	0.54	0.69
Knowledge intensity	121,773	0.14	0.21	0.00	0.00	0.00	0.07	0.31
Organizational intensity	121,773	0.85	0.23	0.67	0.91	0.99	1.00	1.00
Panel C: Bena and Simintzi (2000-2020)								
Process intensity	13,605	0.33	0.25	0.00	0.09	0.31	0.50	0.83
Process intangibility	13,605	0.20	0.25	0.00	0.00	0.09	0.41	0.74
Non-process intangibility	13,605	0.18	0.07	0.03	0.11	0.21	0.24	0.25
Panel D: ExecuComp and Equilar (1992-2021)								
CEO Bonus (%)	30,481	0.03	0.06	0.00	0.00	0.00	0.00	0.04
CEO Ownership	30,481	0.02	0.05	0.00	0.00	0.00	0.00	0.02
Mana. Bonus (%)	30,481	0.07	0.13	0.00	0.00	0.01	0.03	0.10
Mana. Ownership	30,481	0.03	0.07	0.00	0.00	0.01	0.01	0.03
Total compensation	30,481	0.01	0.02	0.00	0.00	0.00	0.01	0.05
Management deferred comp.	18,321	0.01	0.02	0.00	0.00	0.00	0.01	0.03
Fraction deferred comp.	18,321	0.54	0.36	0.00	0.36	0.56	0.70	0.90
CEO age	30,481	56.28	7.07	50.00	54.00	56.00	58.00	62.00
CEO tenure	30,481	7.37	7.41	2.00	4.00	5.00	7.00	12.00
CEO overconfidence	14,486	0.66	0.47	0.00	1.00	1.00	1.00	1.00
CEO overconfidence alt.	16,422	0.47	0.50	0.00	0.00	0.00	1.00	1.00
Panel E: Trucost (1999-2021)								
Scope 1 intensity GHG emissions	15,486	0.12	0.32	0.01	0.02	0.02	0.03	0.07
Scope 2 intensity GHG emissions	15,486	0.04	0.05	0.01	0.02	0.02	0.03	0.05
Scope 3 intensity GHG emissions	15,486	0.57	1.50	0.07	0.13	0.18	0.26	0.49
Panel F: MSCI (2007-2021)								
ESG Score	11,299	4.58	0.99	3.80	4.40	4.60	4.90	5.40
Industry-adj. ESG Score	14,455	4.58	2.15	2.80	3.90	4.40	5.10	6.50
E Score	14,454	4.86	2.11	3.00	4.15	4.70	5.30	6.60
S Score	14,455	4.53	1.58	3.20	4.17	4.60	4.90	5.80
G Score	14,450	5.38	1.97	3.90	4.82	5.30	5.70	6.80

Appendix D Corporate Reward Puzzle

We use the model from Section 2.1 to document the effects of underestimating the cost of investing, p , as managers seem to do. In the benchmark model, a dollar of investment costs as much as a dollar of profits today, hence, $p = 1$ in the benchmark model. However, the estimation results in Table 2 show that $p = 0.45$ - 0.55 or even less.

To explore the impact of $p < 1$ or, equivalently, $\theta_{\frac{I}{K}} \in (-1, 0)$ on the dynamics of investment, we simulate the neoclassical investment model in two ways. In the benchmark case, we set $p = 1$, simulate profitability shocks and compute the optimal investment rates in each state. In the de-facto case, we set $p = 0.5$, simulate the same profitability shocks as in the benchmark case and compute the de-facto investment rates in each state. We then summarize and compare them.

Table D.1 documents statistics for the investment rates in the de-facto case (columns 1-3) and the benchmark case (columns 4-6). Across rows, we condition on different profitability levels (z) ranging from low to high, with specific values of $z = 0.894, 0.946, 1, 1.058, 1.119$. We report the mean investment rate, the variance of the investment rate, and the Kullback-Leibler divergence (KLD) between the policy function and a uniform distribution. The KLD represents a measure of exploration, with the uniform distribution representing maximum entropy (i.e., maximum exploration). A lower value of the KLD captures proximity to the uniform distribution and, hence, higher entropy and more exploration.

The table shows that the manager's de-facto reward function with $p < 1$ has a significant effect on the dynamics of investment. Higher profitability corresponds to higher expected investment rates. For $p = 0.5$, as profitability increases, both the mean investment rate and KLD increase, while the variance decreases slightly. In the benchmark case, mean investment rates are significantly lower compared to $p = 0.5$, particularly at low profitability levels. The KLD is also lower, indicating more exploration in the benchmark case, while variance is somewhat stable. The KLD values represent how much the manager explores different investment strategies. The higher KLD in the de-facto case indicates less exploration, while the lower KLD (closer to the uniform distribution) in the benchmark case means the manager is exploring more investment options.

These results show that managers behave as if they over-invest and under-explore. The over-investment is pervasive across profitability levels, with the relative increase the largest for low levels of profitability. In addition, managers do too little exploration, or experimentation by varying investment less than optimal.

Table D.1: Corporate reward puzzle.

This table documents, for different profitability shocks z and different prices of investment (as perceived by the manager) p , the mean investment rate, variance of the investment rate, and the Kullback-Leibler divergence (KLD) between the policy function and a uniform distribution. The KLD represents a measure of exploration, with the uniform distribution representing maximum entropy/exploration. A lower value of the KLD captures proximity to the uniform distribution, i.e., higher entropy and more exploration.

Profitability		Investment rate ($p = 0.5$)			Benchmark investment rate ($p = 1$)		
		Mean	Variance	KLD	Mean	Variance	KLD
Low	($z=0.894$)	0.380	0.113	0.294	0.005	0.102	0.051
2	($z=0.946$)	0.401	0.109	0.342	0.027	0.106	0.035
3	($z=1$)	0.423	0.103	0.395	0.051	0.111	0.022
4	($z=1.058$)	0.444	0.097	0.452	0.076	0.115	0.014
High	($z=1.119$)	0.464	0.092	0.508	0.103	0.119	0.011

Appendix E Extensions of the Analysis and Robustness

In this section, we explore several extensions of the analysis, including process intangibles, joint investment and leverage problems, and the impact of managerial preferences and beliefs.

Process intangibles. Table E.2 summarizes our reward recovery for process intangibles using data from 2000 to 2021 based on Chen et al. (2024). The table presents the recovered feature importance weights (θ) for several key variables in the manager’s reward function. Alt. Profitability is used as the baseline (normalized to 1), and its coefficients remain relatively stable across specifications. The Total Investment Rate consistently appears with negative coefficients (normalized values ranging from -0.550 to -1.188), indicating that higher investment rates reduce the reward, which is in line with standard investment discipline.

The table further decomposes intangible capital into several components. In columns (2) and (4), the measure of overall Intangible Intensity shows positive weights (normalized 0.345 and 0.258), suggesting that, in general, investing in intangible assets is rewarded. However, when focusing specifically on process-related intangibles, the picture becomes more nuanced. Process Intensity (columns (3) and (4)) and Process Intangibles (columns (5) and (7)) carry negative coefficients (normalized between -0.115 and -0.136), implying that aspects of process intangibility may detract from immediate reward, possibly due to their inherent costs or risks associated with their implementation. Conversely, Non-Process Intangibles (columns (6) and (7)) have positive weights (normalized around 0.135 – 0.161), suggesting that other forms of intangible assets contribute favorably to the reward function.

The λ parameter varies notably across specifications (ranging from 0.042 to 0.115), while the Kullback-Leibler divergence and the alignment measures reveal differences in model fit and the extent to which the recovered reward function correlates with firm value (Total q). Overall, these results indicate that while overall intangible investments may enhance the firm’s reward, the process-specific components have a more complex, sometimes negative, influence on managerial decision-making.

Non-pecuniary benefits and ESG. Table E.7 shows the results when we include ESG scores. ESG data is available only from 2007 onward. Expert 1 is now defined as the time period from 2007 until 2014, whereas expert 2 is defined as the time period from 2015 until 2021. In column (1), we repeat the previous analysis with the updated definitions. The coefficient on the investment rate is now -0.629 , i.e., closer to the theoretical benchmark. The KLD decreases to 0.294 and 0.488 , respectively, indicating a better fit to the data from 2007 onward. The correlation between the recovered value and Total q decreases compared to Table 2.

In column (2) to (7), we add the weighted-average ESG score, industry-adjusted ESG score, and individual E, S, and G scores. We find a small positive normalized coefficient of 0.008 for the weighted-average ESG score and 0.001 for the industry-adjusted ESG score. Economically, a 1-SD change in the weighted-average/industry-adjusted ESG score ($SD=0.99/2.15$) changes the reward by $0.8\%/0.2\%$, which is a small effect.

Breaking up the ESG score into its individual components reveals that the environmental score has a negative effect on the reward (-0.004), whereas the social (0.005) and governance (0.005) scores have small positive effects. While these coefficients appear small, the average E, S, and G scores are an order of magnitude larger than profitability. A 1-SD change in the E/S/G score ($SD=2.11/1.58/1.97$) changes the reward by $-0.8\%/0.8\%/1.0\%$. By comparison, a 1-SD change in alternative profitability ($SD=0.13$) changes the reward by 13% , and a 1-SD change in total investment ($SD=0.13$) changes the reward by 7.9% .

To further investigate the relevance of E, S, and G scores relative to profitability, we compute how much each of these features contributes to the reward when measured at their mean. On average, environmental performance results in a period-by-period loss in reward that is roughly equal to 2.5% of profitability (in absolute value). Social performance and governance performance result in a period-by-period benefit in reward that is roughly equal to 19% and 33% of profitability, respectively.

Including ESG scores improves the model’s performance in terms of mean KLD, which decreases from $0.294/0.488$ in column (1) to $0.287/0.481$ in column (7). This improvement in model performance is moderate. From this perspective, ESG does not appear to be of large economic relevance for firms.

Greenhouse gas emissions. Table E.8 presents results on the recovered reward function with a focus on the role of greenhouse gas (GHG) emissions. The estimated feature importance for profitability is consistently high and the total investment rate has a negative weight (≈ -12.5), implying that firms trade off investment intensity against profitability as in the prior analysis.

GHG emissions intensity has varying effects. Scope 1 emissions (direct emissions) appear in models (2) and (5) with a small but positive effect (0.464 and 0.699). Scope 2 emissions (indirect emissions from purchased electricity)

appear in models (3) and (5), with a stronger positive effect ($\approx 3.2 - 3.6$). This suggests that firms may not get heavily penalized for indirect emissions. Scope 3 emissions (supply chain emissions) appear in models (4) and (5) with negative weights (-0.135 and -0.187), indicating that firms get penalized for extensive supply-chain emissions in their investment decisions. The trade-off parameter λ is stable across models with values around 0.05. The KLD remains consistent across models, suggesting robustness in the inferred reward function. The highest alignment is in model (5), suggesting this specification best captures firms' implicit reward structures.

Joint investment and leverage problem. As before, we choose 5 bins each for the total investment rate and the net debt to EBITDA ratio, resulting in 25 possible actions (compared to 5 before). This has two main implications. First, the dimensionality of our policy functions and transition functions increase and we estimate more parameters using the same number of observations. As a result, we expect the results to be a little more noisy. Second, given that the definition of the action has changed, it is no longer possible to compare the Kullback-Leibler divergence with earlier results.

Table E.9, column (1) shows that our results remain largely unchanged. Managers still act as if investment were less costly than prescribed by theory, with $p = 0.586$. The KL divergence increases to 0.678 and 0.876. The correlation between the recovered value and Total q drops to 0.365 and 0.292. The alignment between the recovered value and Total q goes down when we consider joint investment and leverage decisions, compared to investment decisions alone. This suggests that managers are less aligned with shareholders when it comes to making leverage decisions than investment decisions.

Impact of managerial preferences. So far, we have assumed that the manager acts risk-neutral and discounts at the risk-free rate. This may not be true in practice. We now increase the manager's discount factor such that

$$\gamma_i = \frac{1}{1 + \bar{r}_{f,i} + c} \text{ for } i = 1, 2, \quad (\text{E.1})$$

where $\bar{r}_{f,i}$ denotes the average risk-free rate during the respective period and $c = 0.06$ is a constant representing risk-aversion or impatience. If the manager is indeed risk-averse or impatient, the higher discount rates (and lower discount factors) should improve the model's explanatory performance in terms of KL divergence. Moreover, these higher discount rates are likely more aligned with the weighted average cost of capital used by investors to value the firm. As such, the correlation between the recovered value and Total q should also increase.

Table E.9, column (2) shows that higher discount rates help explain managers' policies and align their recovered value with shareholders. As we increase the discount rate, the KL divergence decreases to 0.283 and 0.450. The correlation between the recovered value and Total q increases to 0.593 and 0.729.

Estimation of choice and transition probabilities. In the main analysis, we have estimated choice and transition probabilities using multinomial logit models, which gives an estimate of choice probabilities conditional on state variables and an estimate of transition probabilities conditional on state and action variables. To better understand how our results depend on the method of estimating the empirical choice and transition probabilities, we repeat the analysis using an alternative non-parametric count estimator. Since we need to exclude zero-probability events and cannot guarantee ex-ante that we will observe every possible (state, action)-pair and (state, action, next state)-triplet in the data, we perform *additive smoothing*. Additive smoothing ensures that there are no zero-probability events by adding α fictional observations to each event. Consequently, we estimate the choice probabilities based on

$$\hat{\pi}_i(a|s) = \frac{\#(a_{i,t} = a \text{ and } s_{i,t} = s) + \alpha}{\#(s_{i,t} = s) + \alpha|\mathcal{A}|}, \quad (\text{E.2})$$

and the transition probabilities based on

$$\hat{T}_i(s'|s, a) = \frac{\#(s_{i,t+1} = s' \text{ and } s_{i,t} = s \text{ and } a_{i,t} = a) + \alpha}{\#(s_{i,t} = s \text{ and } a_{i,t} = a) + \alpha|\mathcal{S}|}, \quad (\text{E.3})$$

where $\alpha > 0$ is a constant that can be arbitrarily small.

Table E.9, column (3) shows the results when we use the count-based estimator of probabilities in the joint investment and leverage problem.

Coarseness of the state and action variable grid. In our analysis, we have opted for a grid of 25 states and 5 (for the investment problem) or 25 actions for the joint investment and leverage problem. We repeat the analysis using finer grids, with up to 100 possible states and actions.

Table E.10 shows the results when increase the number of state and action variable bins in the joint investment and leverage problem. We still find that managers underestimate the cost of investment and the p is relatively stable between 0.416 and 0.602. The λ coefficient is more variable, ranging between 0.19 and 0.345. The explanatory performance in terms of KL divergence decreases due to the curse of dimensionality. As the number of state and action variable bins increases, the number of choice probabilities $p(a|s)$ and transition probabilities $T(s'|s, a)$ increases exponentially while the number of observations used to estimate these parameters stays constant.

Table E.1: Corporate reward recovery - Industry sample splits.

The table presents the results of recovering the reward function from corporate investment data based on our main sample from 1975 to 2021. It displays the recovered feature importance weights, denoted by θ , for nine industries. Industry definitions are based on the Fama-French twelve industries. The reward features considered include profitability and investment rate. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

Industry:	Recovered feature importance θ (θ normalized)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Non durables	Durables	Manufactur.	Energy	Business Eq.	Telecom.	Shops	Health Care	Other
Alt.	3.258	8.573	27.901	31.619	8.334	4.763	6.397	17.118	17.044
Profitability	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total	-0.621	-3.067	-14.316	-11.253	-3.835	-1.461	-0.387	-8.629	-6.272
Investment Rate	(-0.191)	(-0.358)	(-0.513)	(-0.356)	(-0.46)	(-0.307)	(-0.061)	(-0.504)	(-0.368)
λ	0.307	0.117	0.036	0.032	0.120	0.210	0.156	0.058	0.059
$\overline{KLD}_1$	0.526	0.339	0.357	0.342	0.736	0.132	0.713	0.472	0.263
$\overline{KLD}_2$	0.385	0.373	0.471	0.211	0.472	0.190	0.755	0.759	0.257
$Alignment_1$	0.527	0.433	0.515	0.123	0.883	0.453	0.932	0.935	0.424
$Alignment_2$	0.507	0.343	0.682	0.088	0.201	0.190	0.513	0.688	0.270
N_1	6,877	2,916	13,709	3,369	7,192	1,722	9,387	3,019	8,611
N_2	4,242	2,035	8,479	2,847	8,62	2,981	7,107	3,986	8,251

Table E.2: Corporate reward recovery - Process intangibles.

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 2000 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing process intangibility. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Alt. Profitability	20.939	8.694	21.091	10.247	23.818	16.765	19.537
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-13.508	-10.329	-12.093	-10.017	-13.098	-12.954	-12.587
	(-0.645)	(-1.188)	(-0.573)	(-0.978)	(-0.550)	(-0.773)	(-0.644)
Intangible intensity		3.001		2.639			
		(0.345)		(0.258)			
Process intensity			-2.421	-1.191			
			(-0.115)	(-0.116)			
Process intang.					-2.868		-2.658
					(-0.120)		(-0.136)
Non-process intang.						2.699	2.631
						(0.161)	(0.135)
λ	0.048	0.115	0.047	0.098	0.042	0.060	0.051
$\overline{KLD}_1$	0.308	0.246	0.290	0.245	0.295	0.297	0.283
$\overline{KLD}_2$	0.570	0.385	0.515	0.382	0.558	0.492	0.480
$Alignment_1$	0.412	0.532	0.109	0.214	0.220	0.293	0.140
$Alignment_2$	0.682	0.573	0.139	0.688	0.473	0.689	0.241
N_1	28,088	28,088	28,088	28,088	28,088	28,088	28,088
N_2	18,637	18,637	18,637	18,637	18,637	18,637	18,637

Table E.3: Corporate reward recovery - CEO and managerial ownership.

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 1992 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing CEO and managerial bonus and ownership. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Alt. Profitability	25.175 (1.000)	25.596 (1.000)	25.172 (1.000)	25.683 (1.000)	25.833 (1.000)	25.304 (1.000)	26.074 (1.000)
Total Investment Rate	-14.118 (-0.561)	-14.342 (-0.560)	-14.196 (-0.564)	-14.471 (-0.563)	-14.499 (-0.561)	-14.293 (-0.565)	-14.658 (-0.562)
CEO Ownership		6.506 (0.254)		7.901 (0.308)			
CEO Bonus			-4.529 (-0.180)	-4.777 (-0.186)	-4.576 (-0.177)		
Managerial Ownership					8.198 (0.317)		9.389 (0.360)
Managerial Bonus						-1.695 (-0.067)	-1.868 (-0.072)
λ	0.040	0.039	0.040	0.039	0.039	0.040	0.038
$\overline{KLD}_1$	0.383	0.382	0.368	0.367	0.367	0.371	0.368
$\overline{KLD}_2$	0.560	0.561	0.534	0.535	0.537	0.541	0.542
$Alignment_1$	0.433	0.347	0.554	0.436	0.364	0.593	0.383
$Alignment_2$	0.601	0.440	0.849	0.647	0.524	0.899	0.569
N_1	39,699	39,699	39,699	39,699	39,699	39,699	39,699
N_2	29,753	29,753	29,753	29,753	29,753	29,753	29,753

Table E.4: Corporate reward recovery - Management total vs. deferred compensation.

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 2004 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing management deferred compensation. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	20.965 (1.000)	20.108 (1.000)	19.468 (1.000)	20.637 (1.000)	19.433 (1.000)	19.468 (1.000)
Total Investment Rate	-12.989 (-0.620)	-13.449 (-0.669)	-13.557 (-0.696)	-13.742 (-0.666)	-13.758 (-0.708)	-13.805 (-0.709)
Total Compensation		27.285 (1.357)				4.005 (0.206)
Manag. Deferred Comp.			72.654 (3.732)		69.578 (3.580)	60.758 (3.121)
Fraction Deferred Comp.				0.960 (0.047)	0.287 (0.015)	0.349 (0.018)
λ	0.048	0.050	0.051	0.048	0.051	0.051
$\overline{KLD}_1$	0.305	0.277	0.265	0.295	0.264	0.264
$\overline{KLD}_2$	0.510	0.461	0.444	0.494	0.442	0.442
$Alignment_1$	0.351	0.115	0.156	0.642	0.208	0.205
$Alignment_2$	0.504	0.123	0.197	0.751	0.176	0.172
N_1	19,051	19,051	19,051	19,051	19,051	19,051
N_2	13,149	13,149	13,149	13,149	13,149	13,149

Table E.5: Corporate reward recovery - Employee deferred compensation.

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 1992 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing employee deferred compensation. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively. Specifications (1) presents results for the benchmark model. Specifications (2) accounts for employee deferred compensation. Specifications (3) and (4) present results for large and small firms, respectively. Large/small firms are defined as firms having total assets above/below the 67th/33rd-percentile.

	Recovered feature importance θ (θ normalized)			
	(1)	(2)	(3)	(4)
Alt. Profitability	21.511 (1.000)	11.508 (1.000)	12.010 (1.000)	11.948 (1.000)
Total Investment Rate	-13.285 (-0.618)	-12.378 (-1.076)	-13.233 (-1.102)	-8.985 (-0.752)
Employee Deferred Comp.		45.781 (3.978)	79.800 (6.644)	11.660 (0.976)
λ	0.046	0.087	0.083	0.084
$\overline{KLD}_1$	0.289	0.238	0.242	0.225
$\overline{KLD}_2$	0.526	0.599	0.559	0.514
$Alignment_1$	0.345	0.435	0.330	0.897
$Alignment_2$	0.616	0.815	0.563	0.875
N_1	22,193	22,193	11,592	3,803
N_2	15,441	15,441	9,170	2,179

Table E.6: Corporate reward recovery - Behavioral characteristics

The table presents the results of recovering the reward function from corporate investment data. Specifications (1) to (3) are based on a sub-sample from 1992 to 2021. Specifications (4) to (6) are based on a sub-sample from 2006 to 2021. The table displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing CEO behavioral characteristics. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	25.175 (1.000)	24.933 (1.000)	25.202 (1.000)	22.317 (1.000)	21.944 (1.000)	22.321 (1.000)
Total Investment Rate	-14.118 (-0.561)	-14.381 (-0.577)	-14.155 (-0.562)	-14.384 (-0.645)	-14.272 (-0.65)	-14.347 (-0.643)
CEO age		-0.034 (-0.001)				
CEO tenure			0.018 (0.001)			
CEO overconfidence					0.905 (0.041)	
CEO overconfidence alt.						-0.014 (-0.001)
λ	0.040	0.040	0.040	0.045	0.046	0.045
$\overline{KLD}_1$	0.383	0.375	0.382	0.285	0.260	0.285
$\overline{KLD}_2$	0.560	0.542	0.563	0.512	0.477	0.512
$Alignment_1$	0.433	0.427	0.396	0.324	0.302	0.318
$Alignment_2$	0.601	0.608	0.579	0.615	0.537	0.555
N_1	39,699	39,699	39,699	25,332	25,332	25,332
N_2	29,753	29,753	29,753	16,194	16,194	16,194

Table E.7: Do ESG considerations affect corporate purpose?

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 2013 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing firm's environmental, societal, and governance (ESG) performance. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Alt. Profitability	21.921 (1.000)	21.021 (1.000)	21.820 (1.000)	22.477 (1.000)	21.48 (1.000)	21.485 (1.000)	21.626 (1.000)
Total Investment Rate	-13.781 (-0.629)	-13.601 (-0.647)	-13.776 (-0.631)	-13.673 (-0.608)	-13.537 (-0.630)	-13.747 (-0.640)	-13.353 (-0.617)
ESG Score		0.162 (0.008)					
Adj. ESG Score			0.013 (0.001)				
Environmental Score				-0.090 (-0.004)			-0.109 (-0.005)
Social Score					0.114 (0.005)		0.120 (0.006)
Governance Score						0.106 (0.005)	0.122 (0.006)
λ	0.046	0.048	0.046	0.044	0.047	0.047	0.046
$\overline{KLD}_1$	0.294	0.289	0.293	0.295	0.292	0.289	0.287
$\overline{KLD}_2$	0.488	0.478	0.488	0.489	0.487	0.482	0.481
$Alignment_1$	0.260	0.318	0.271	0.199	0.264	0.332	0.257
$Alignment_2$	0.537	0.581	0.405	0.355	0.547	0.567	0.384
N_1	18,822	18,822	18,822	18,822	18,822	18,822	18,822
N_2	10,931	10,931	10,931	10,931	10,931	10,931	10,931

Table E.8: Do GHG emissions matter?

The table presents the results of recovering the reward function from corporate investment data based on a sub-sample from 2002 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability, investment rate, and measures capturing the intensity of firm's GHG emissions. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively.

	Recovered feature importance θ (θ normalized)				
	(1)	(2)	(3)	(4)	(5)
Alt. Profitability	19.723 (1.000)	19.950 (1.000)	20.364 (1.000)	19.750 (1.000)	20.820 (1.000)
Total Investment Rate	-12.333 (-0.625)	-12.419 (-0.622)	-12.525 (-0.615)	-12.525 (-0.634)	-12.944 (-0.622)
Scope 1 intensity		0.464 (0.023)			0.699 (0.034)
Scope 2 intensity			3.203 (0.157)		3.578 (0.172)
Scope 3 intensity				-0.135 (-0.007)	-0.187 (-0.009)
λ	0.051	0.050	0.049	0.051	0.048
$\overline{KLD}_1$	0.315	0.312	0.312	0.313	0.304
$\overline{KLD}_2$	0.544	0.541	0.543	0.54	0.535
$Alignment_1$	0.369	0.370	0.368	0.416	0.435
$Alignment_2$	0.440	0.436	0.440	0.468	0.465
N_1	28,172	28,172	28,172	28,172	28,172
N_2	22,718	22,718	22,718	22,718	22,718

Table E.9: Extensions and robustness checks.

The table presents the results of recovering the reward function from corporate investment data based on our main sample from 1975 to 2021. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability and investment rate. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively. Specification (1) presents results from an extension of the main model that includes the joint investment and leverage decisions. Specifications (2) presents results of the main model that accounts for managerial risk-aversion or impatience. Specification (3) presents results of the main model where the estimation of choice and transition probabilities relies on an alternative method. These model robustness tests are described into details in [Appendix E](#).

Panel A: Recovered feature importance θ (θ normalized)			
	(1)	(2)	(3)
	Joint investment and financing	Managerial risk-aversion	Alternative choice and transition probabilities
Alt. Profitability	5.068 (1.000)	23.353 (1.000)	2.434 (1.000)
Total Investment Rate	-2.970 (-0.586)	-14.047 (-0.602)	-1.324 (-0.544)
λ	0.197	0.043	0.411
$\overline{KLD}_1$	0.678	0.283	0.630
$\overline{KLD}_2$	0.876	0.450	0.608
$Alignment_1$	0.365	0.593	0.554
$Alignment_2$	0.292	0.729	0.357
N_1	18,822	18,822	18,822
N_2	10,931	10,931	10,931

Table E.10: Robustness to implementation choices.

The table presents the results of recovering the reward function from corporate investment and leverage data based on our main sample from 1975 to 2021 when we change our binning approach. It displays the recovered feature importance weights, denoted by θ . The reward features considered include profitability and investment rate. Variable definitions are provided in [Appendix C](#). λ measures the trade-off between exploration and exploitation. The Kullback-Leibler performance measure, $\overline{KLD}$, reports the divergence between estimated and true policies. The alignment measure captures the correlation between the recovered value function and Total q. N reports the number of observations. Subscripts 1 and 2 refer to Agent 1 and Agent 2, respectively. Model specifications include for the two state variables 5-10 ln(Total Capital) (BoY) bins, 5-10 alternative profitability (BoY) bins, and for the two action variables 5-10 total investment rate bins, and 5-10 net debt to EBITDA bins. These model robustness tests are described into details in [Appendix E](#).

Panel A: Number of bins for states and action variables						
	(1)	(2)	(3)	(4)	(5)	(6)
State Variable(s) and Bins						
ln(Total Capital) (BoY)	5	10	5	5	5	10
Alt. Profitability (BoY)	5	5	10	5	5	10
Action Variable(s) and Bins						
Total Investment Rate	5	5	5	10	5	10
Net Debt to EBITDA	5	5	5	5	10	10
Panel B: Recovered feature importance θ (θ normalized)						
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	5.264 (1.000)	3.484 (1.000)	5.168 (1.000)	3.485 (1.000)	3.904 (1.000)	2.894 (1.000)
Total Investment Rate	-2.788 (-0.530)	-1.913 (-0.549)	-2.833 (-0.548)	-1.875 (-0.538)	-2.349 (-0.602)	-1.205 (-0.416)
λ	0.190	0.287	0.193	0.287	0.256	0.345
$\overline{KLD}_1$	0.677	0.716	0.812	0.743	0.752	0.917
$\overline{KLD}_2$	0.876	0.906	1.028	0.929	0.926	1.149
$Alignment_1$	0.490	0.375	0.646	0.404	0.394	0.435
$Alignment_2$	0.347	0.284	0.480	0.324	0.344	0.311
N_1	18,822	18,822	18,822	18,822	18,822	18,822
N_2	10,931	10,931	10,931	10,931	10,931	10,931

Swiss Finance Institute

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public-private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.

about ECGI

The European Corporate Governance Institute has been established to improve *corporate governance through fostering independent scientific research and related activities*.

The ECGI will produce and disseminate high quality research while remaining close to the concerns and interests of corporate, financial and public policy makers. It will draw on the expertise of scholars from numerous countries and bring together a critical mass of expertise and interest to bear on this important subject.

The views expressed in this working paper are those of the authors, not those of the ECGI or its members.

ECGI Working Paper Series in Finance

Editorial Board

Editor	Nadya Malenko, Professor of Finance, Boston College
Consulting Editors	Renée Adams, Professor of Finance, University of Oxford Franklin Allen, Nippon Life Professor of Finance, Professor of Economics, The Wharton School of the University of Pennsylvania Julian Franks, Professor of Finance, London Business School Mireia Giné, Associate Professor, IESE Business School Marco Pagano, Professor of Economics, Facoltà di Economia Università di Napoli Federico II
Editorial Assistant	Asif Malik, Working Paper Series Manager

Electronic Access to the Working Paper Series

The full set of ECGI working papers can be accessed through the Institute's Web-site (www.ecgi.global/content/working-papers) or SSRN:

Finance Paper Series	http://www.ssrn.com/link/ECGI-Fin.html
Law Paper Series	http://www.ssrn.com/link/ECGI-Law.html