AI in Corporate Governance: Can Machines Recover Corporate Purpose?

Boris Nikolov*, Norman Schürhoff†, Sam Wagner‡

September, 2024

Abstract

A key question in automating governance is whether machines can recover the corporate objective. We develop a corporate recovery theorem that establishes when machines can do this. Training a machine on a large dataset of firms' investment and financial decisions, we find that managers systematically underestimate investment costs, leading to over-investment and under-exploration. This bias persists even when accounting for intangibles, managerial compensation, and ESG scores. While social and governance concerns influence corporate objectives beyond materiality, environmental concerns do not. Last, we observe that managerial alignment with shareholder value is imperfect, but it has improved over time.

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 recently been challenged by alternative theories focusing on welfare, stakeholders, misalignment of incentives, and non-standard preferences. A firm can and in some cases should have objectives beyond maximizing shareholder value for a number of reasons. Corporations may 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). In addition, CEOs and other managers face incentive problems leading to agency conflicts in which their personal interests do not align with the interests of the shareholders and which can result in suboptimal decisions and outcomes for the corporation (Jensen and Meckling, 1976).[1] Ultimately, the objective function of a firm likely depends on a number of features, including its social objectives, governance, environmental impact, organizational structure, ownership structure, industry, management 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 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 states, one can derive conditions for when 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 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 which objective function of the firm can be recovered based on firms' observed investment, financing, and operational policies.

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 observe the actions of a human expert, in our case the CEO managing the firm, and then infer the reward

[1] To 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.

[2] IRL is the reverse of the reinforcement learning (RL) approach, where the agent learns the optimal policy based on a given reward function (Sutton and Barto, 2018).

function that the expert is maximizing. Once the reward function is inferred, it can be used to train an artificial agent to perform the task in a new 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.

We develop a corporate recovery theorem and apply the IRL methodology to financial data on firms' investment and financial policies which allows us to recover the importance of features affecting firms' objectives. We start by generalizing the results in Cao et al. (2021) and Rolland et al. (2022) to the case where the agents' reward functions contain additively separable preference shocks and the weight on exploration (i.e., the entropy regularization weight) are unknown to the observer/machine, which is a scenario likely to be the case in economics and finance. We show that the agents' instantaneous rewards and value functions can still be recovered up to affine transformations for each state-action pair observed in the data. Based on these results, we can similar to regressions specify a set of features (or independent variables) and recover the coefficients on these features.

We document a number of stylized facts about the corporate purpose. Applying the corporate recovery theorem to a large dataset on investment and financial policies of publicly traded U.S. firms, we document the following findings. 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, or that the reward function is of the form $Profitability - InvestmentRate$, reflecting a price of capital equal to one. However, we find that managers act as if they underestimate the cost of investment by up to 35-55% relative to theory. Simply speaking, managerial policies are consistent with

a reward function of the form $Profitability - p \times InvestmentRate$ where 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 control features, managerial compensation and ESG scores, as well as different discount rates.

These results show that managers de-facto 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. This *corporate reward puzzle* is inconsistent with the predictions from the foundational q-theory model and supports alternative theories predicting over-investment and under-exploration.[3] 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 value-destructing over-investment but silent about exploration/experimentation.

To better understand the source of the investment inefficiency, we explore whether financial factors, intangible capital, compensation policy, managerial risk aversion, and ESG concerns enter the corporate objective. We find that financial factors, including cash holdings, book leverage, and net book leverage, contribute to the corporate reward function, although their impact is neither large nor consistent across models. Intangible capital (following Peters and Taylor (2017)) 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.

We then construct a shareholder alignment measure, *Alignment*, defined as the correlation across different states between the recovered value that managers maximize and observed market values. Our

[3]Over-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.

results suggest that managerial alignment with shareholder value is imperfect, but it has significantly improved over time. Agency theory implies a misalignment of interests, predicting that, absent governance mechanisms, managers take decisions with respect to their own reward function rather than the one of shareholders, predicting $Alignment < 1$. We quantify the degree of alignment relative to both Tobin's q and Total q. We find that compensation policy significantly affects the manager's reward but the improvement in model performance is small from incorporating compensation policy. CEO bonus, CEO ownership, CEO options, and other C-level managers' compensation all enter the corporate objective but do not improve model fit by much. The degree of manager-shareholder alignment is also not materially affected by compensation policy, reaching up to $Alignment = 0.85$ (0.52) for Tobin's q (Total q).

Stakeholder theory predicts that social and environmental performance should positively affect the reward function. Managers' survey responses and shareholder communications confirm that, de jure, they act in stakeholders' interests, including employees, customers, and the environment (Graham, 2022; Rajan et al., 2024). When adding social and environmental performance scores, we find that good social performance (which takes into account criteria such as human capital, product liability, stakeholder opposition and social opportunities[4] and benefits employees and customers) increases the manager's period-by-period reward by approximately 19% of profitability on average. However, environmental performance has little to no effect on the manager's reward on top of profits when considering only investment decisions, and decreases the reward by 11% when considering joint investment and leverage decisions. While this suggests strong materiality of environmental concerns or no care for the environment, the latter is in line with survey evidence that only 20% of CFO responded that they manage the company in the interest of the environment. To the extent that we recover the reward function consistent with all managers' actions, another possibility is that these managers spoke the truth, but were counterbalanced by others that do not care about the environment (Hart and Zingales, 2017).

Overall, firms maximize shareholder value to a large extent in the sense that the alignment measure reaches values of $Alignment = 0.89$, with 1 indicating perfect alignment. However, we find this large alignment only after accounting for the fact that social and governance considerations enter the firm's de-facto objective. The correlation between the recovered value and Total q increases substantially once we incorporate ESG considerations, from 0.20 for 2007-2014 to 0.51 for 2014-2021 without to 0.77 for 2007-2014 to 0.80 for 2014-2021 with ESG considerations. This increase in manager-shareholder alignment

[4]See https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf.

suggests that (environmental), social, and governance considerations play a substantial role in firms' decision-making, and it suggests a strong but imperfect alignment of the reward function of the manager with the valuation of the firm in the market.

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. By contrast, we infer by applying IRL the agents' underlying incentives and rewards based on their actions across different states. Our methodology offers significant potential for use in various finance and economics contexts, including corporate finance, asset management, market microstructure, banking, and household finance, where learning about the agents' reward and value function can provide economic insights without relying on a specific parametric model.

Literature review. The literature on the corporate objective and purpose is vast. The debate over the corporate objective is rooted in two primary perspectives: the shareholder primacy view and the stakeholder theory. The shareholder primacy view, advocated by scholars like Friedman (1970), posits that the sole responsibility of a corporation is to maximize shareholder wealth, operating within the boundaries of the law and ethical customs. In contrast, the stakeholder theory, advanced by Freeman (1984), argues that corporations have a broader responsibility to various stakeholders, including employees, customers, suppliers, and the community, alongside shareholders. This perspective emphasizes that sustainable success is achieved by balancing the interests of all stakeholders.

Shareholder wealth maximization (Friedman, 1970) effectively means the maximization of a stream of discounted cash flows to shareholders which is the prevailing norm for the corporate objective function. Assuming that markets are competitive and absent monopolies and externalities, social welfare is also maximized under shareholder wealth maximization. As Jensen and Meckling (1976) show, this is due to the fact that shareholders are the residual claim-holders to the firm's cash flows. All other parties involved in the production of the firm's output (suppliers, employees, debtholders, government) receive their rewards ahead of shareholders. Many studies, for example in the literature on the relation between investment and Tobin's Q (Tobin, 1969; Hayashi, 1982; Abel and Eberly, 1994), focus on firm value maximization as an alternative to shareholder wealth maximization.[5] Absent bondholder expropriation,

[5]Firm value is the sum of the values of all financial claims on the firm, including equity, debt, preferred stock and warrants. Compared to shareholder wealth maximization, it is less restrictive in the sense that it does not require that bondholders must be protected from expropriation. When markets are not competitive or in the presence

firm value maximization implies shareholder wealth maximization.

Proponents of stakeholder value maximization (Freeman, 1984) argue that the firm should maximize the value of all its stakeholders (customers, suppliers, employees, governments, etc.). While there is no consensus on what exactly constitutes the set of stakeholders (Miles, 2012), the shareholders of the firm are free to choose. Stakeholder value maximization implies that the reward function includes at least some altruistic/charitable component in addition to the profit motive (Elhauge, 2005; Graff Zivin and Small, 2005; Baron, 2007; Bénabou and Tirole, 2010; Magill et al., 2015; Hart and Zingales, 2017, 2022; Morgan and Tumlinson, 2019; Ericson, 2024). Without entering into the details, environmental, social and governance (ESG) performance considerations, sustainable development goals (SDGs) and corporate social responsibility (CSR) are all related to stakeholder value maximization in that they also advocate for additional terms beyond profits in the reward function.[6]

Separation of ownership and decision-making gives rise to the principal-agent problem (Jensen and Meckling, 1976), meaning a possible misalignment between the interests of the manager (decision-maker) and those of the shareholders (owners). Absent any governance mechanisms, the manager maximizes their own welfare, consisting of a stream of discounted rewards to the manager, which may be different from the one of shareholders. Agency frictions can arise due to differences in preferences (e.g., risk aversion), time horizon, and/or reward function, which we model as well.

There are two major points of contention in the literature. First, do shareholders care about non-pecuniary benefits? Second, do managers' decisions align with shareholder welfare maximization or not (e.g. due to agency frictions or behavioral biases)? The first question is difficult to answer in practice. For example, a firm that invests into their employees today may do so because they care about their employees' well-being or because they expect these employees to bring in more profits in the future. To clearly establish that the firm cares about non-pecuniary benefits, one would have to observe that the

of negative externalities, social welfare is not maximized under shareholder wealth maximization. One common view states that it is the government's role to impose constraints on the firm to avoid negative externalities on other parties or make the firms internalize the cost of these negative externalities. However, some believe that the government has failed to do this and call for firms to self-regulate in that aspect.

[6] Opponents of stakeholder value maximization (or some implementation thereof) often criticize that stakeholder value maximization is infeasible to implement in practice (Jensen, 2010; Bebchuk and Tallarita, 2020). Non-pecuniary benefits (e.g. *"making the world a better place"*) may be hard or even impossible to measure. Even if perfectly measurable, as the number of shareholders of a firm increases, it becomes increasingly likely that their individual valuations of these non-pecuniary benefits vary. As a result of this, two problems are likely to arise: it may be practically infeasible to aggregate individual preferences into one single objective function and the manager may not be aware of the shareholders' (individual or aggregate) preferences. This, in turn, can exasperate agency problems (holding the firm manager accountable is hard when there is no clear goal) and can render the company less competitive (a firm that under-weights the profit motive may see its survival threatened by competitors).

firm continuously invests in their employees despite this not resulting in increased marginal profits over a long horizon. Hong and Shore (2023) survey the literature and find that, overall, shareholders care about corporate social responsibility and that their interest is predominantly driven by non-pecuniary motives. Graham (2022) surveys managers about their *de jure* objective function (i.e., what managers say that they maximize). When asked in whose interest they think the company should be run in 2020, CFOs responded that the company should be managed to 41% in the interest of stakeholders (compared to 31% in 2010). When asked about which stakeholders are most important, approximately 60% responded employees, 50% responded customers and 20% responded the environment. The study also finds that managers use heuristics and suffer from costly biases.

Our methodology, Inverse Reinforcement Learning (IRL), is not widespread in economics and finance. We contribute to the literature on the corporate objective function by applying IRL to recover the reward function that firm managers use in their decision-making. In doing so, this paper is the first to use IRL in a corporate finance setting to recover corporate policies and managers' reward function in a model-free setting. IRL consists in inferring an agent's unknown underlying preferences or reward function by observing its behavior (Russell, 1998; Ng et al., 2000).

IRL is directly related to revealed preference theory. The economics literature distinguishes between axiomatic and revealed preference theory. In axiomatic preference theory, the reward function is postulated or derived from basic axioms. However, empirical and experimental studies often reject these reward specifications and show that agents display behavioral biases and/or non-standard preferences. Revealed preference theory (Samuelson, 1938) assumes that an agent's reward/preferences can be revealed by their decisions. Our corporate recovery theorem provides conditions for when revealed preferences can be recovered, extending the results in Cao et al. (2021) and Rolland et al. (2022). The use of IRL in economics has clear benefits, as it allows to better understand the rewards of agents in various decision-making problems. It can provide a measurable and interpretable function consistent with agents' behaviour when there is no obvious benchmark.[7] Moreover, this reward function can be used to conduct counterfactual analysis and study how an agent would behave in a new environment.[8] We contribute to this literature by providing conditions for the recovery of an agent's reward and value function and an empirical methodology.

[7]For example, in autonomous driving, the reward function is learned by observing expert drivers. Yet, none of these expert drivers would be able to provide a quantitative estimate of the reward function they implicitly use when while driving.

[8]If one were to simply observe an agent's optimal policy function in one environment, one could not use that policy function to analyse how the same agent would behave in a different environment.

1 The Recovery Framework

One of the main 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 the corporation's 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 economics assumptions and infer the model parameters from panel data on, for instance, firms' investment and financing decisions, including profitability, size, investment, depreciation, leverage, and other features.[9] This section explores conditions for the recovery of an agent's reward and value functions, assuming optimality but a limited set of other economic assumptions.

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 $i = 1, ..., I$ 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 agent. The agent must trade off exploitation (maximizing immediate rewards) and exploration (gathering information about the environment), which leads to stochastic policies $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ 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 $\pi_i(a|s_t)$:

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

where $V_i^{\pi^*}(s)$ is the value function under optimal policy π^* and initial state $s_0 = s$, $\mathbb{E}_i$ denotes the

[9] A similar logic applies to other fields of finance, including consumer finance where one tries to explain consumers' financial decisions.

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 $i = 1, ..., I$ 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 $a \in \mathcal{A}$:

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

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

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

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 $\{s_0, a_0, s_1, a_1, s_2, a_2, ..., s_\infty, a_\infty\}$ 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). For any i with fixed policy $\pi_i(a|s) \in (0, 1)$ and an arbitrary choice of function $v_i : \mathcal{S} \to \mathbb{R}$, Cao et al. (2021) show there is a unique corresponding reward function

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

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 identifying assumption we will use to recover the agents' reward is that λ is the same across i.

1.2 Strong reward recovery

In this environment, 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

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

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

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

Define the residual makers for projections on the space spanned by X_i as

$$\mathbf{M}_i = I - X_i(X_i^\top X_i)^{-1} X_i^\top. \tag{7}$$

The problem of reward recovery 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.

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 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.

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

The corporate reward recovery Theorem 1 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:

$$y_i = X_i v_i + u_i. \tag{8}$$

For any two agents i and j,

$$y_i - y_j = \begin{pmatrix} X_i & -X_j \end{pmatrix} \cdot \begin{pmatrix} v_i \\ v_j \end{pmatrix} + \varepsilon_{ij}, \tag{9}$$

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 $rank(X) = 2|\mathcal{S}| - 1$ 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

$$v_i = a_v + b_v (X_i^\top \mathbf{M}_j X_i)^{-1} X_i^\top \mathbf{M}_j y. \tag{10}$$

These expressions correspond to weighted or generalized least squares.

(ii) The agents' individual reward functions can be recovered for all (s,a) from (13) and any i by $u_i = y_i - X_i \widehat{v}_i$. Finally, the shared common reward can be recovered for all (s,a) by

$$u = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} u_i. \tag{11}$$

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

$$u_i(s,a) = a_u + b_u(\ln \pi_i^*(a|s) - \gamma_i \sum_{s'\in\mathcal{S}} T_i(s'|s,a) v_i(s') + v_i(s)), \tag{12}$$

and $u(s,a)$ can be approximated by $\frac{1}{I}\sum_i u_i(s,a)$. Expression (12) 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 (13). With the stacked vector $v = (v_i^\top \; v_j^\top)^\top \in \mathbb{R}^{2|\mathcal{S}|\times 1}$ and constants $(a_v, b_v) \in \mathbb{R}^2$, (10) yields

$$v = a_v \mathbf{1} + b_v (X^\top X)^{-1} X^\top y. \tag{13}$$

The proposition provides conditions and presents a method for recovering the value functions $V_i^{\pi^*}(s)$ and individual reward functions $u_i(s,a)$ for each agent i. The value functions $V_i^{\pi^*}(s)$ are recovered using closed-form expressions that involve weighted least squares. These expressions allow for the estimation of the value functions up to affine transformations. The individual reward functions $u_i(s,a)$ 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 enables the determination of the reward functions for each agent and each state-action pair. Additionally, the proposition describes how a shared common reward $u(s,a)$ can be recovered by averaging the individual reward functions across all agents. The recovered value and reward functions provide insights into the underlying decision-making processes of the agents. By estimating these functions from observed data, the proposition offers a method for understanding and analyzing the behavior of the agents.

1.3 Weak reward recovery

We now make the additional assumption that the reward function u is linear in features $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, 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 $(i,j) \in \{1,...,I\}^2$ the matrix

$$X = \begin{pmatrix} X_i & -X_j & 0 \\ X_i & 0 & f \end{pmatrix}. \tag{14}$$

Let Ω be the matrix that is composed of four sub-matrices:

$$\Omega \equiv X^\top X = \begin{pmatrix} A & C^\top \\ C & F \end{pmatrix}, \tag{15}$$

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

$$A = \begin{pmatrix} 2X_i^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}, \quad C = \begin{pmatrix} f^\top X_i & 0 \end{pmatrix}, \quad F = (f^\top f). \tag{16}$$

Denote the Schur complements of Ω with respect to A and F by Ω/A and Ω/F, respectively. With these definitions at hand, we obtain the following recovery result:

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

Proof. This is an extension of Theorem 7 in Rolland et al. (2022). Consider the linear reward recovery problem. For any agent $i = 1, ..., I$,

$$y_i = X_i v_i + f\theta + \varepsilon_i, \tag{17}$$

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:

$$\underbrace{\begin{pmatrix} y_i - y_j \\ y_i \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} X_i & -X_j & 0 \\ X_i & 0 & f \end{pmatrix}}_{X} \cdot \begin{pmatrix} v_i \\ v_j \\ \theta \end{pmatrix} + \varepsilon. \tag{18}$$

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

$$\begin{pmatrix} v_i \\ v_j \\ \theta \end{pmatrix} = \underbrace{\begin{pmatrix} (\Omega/F)^{-1} & -(\Omega/F)^{-1} C^\top F^{-1} \\ -(\Omega/A)^{-1} CA^{-1} & (\Omega/A)^{-1} \end{pmatrix}}_{\Omega^{-1}} X^\top (y - \varepsilon). \tag{19}$$

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.

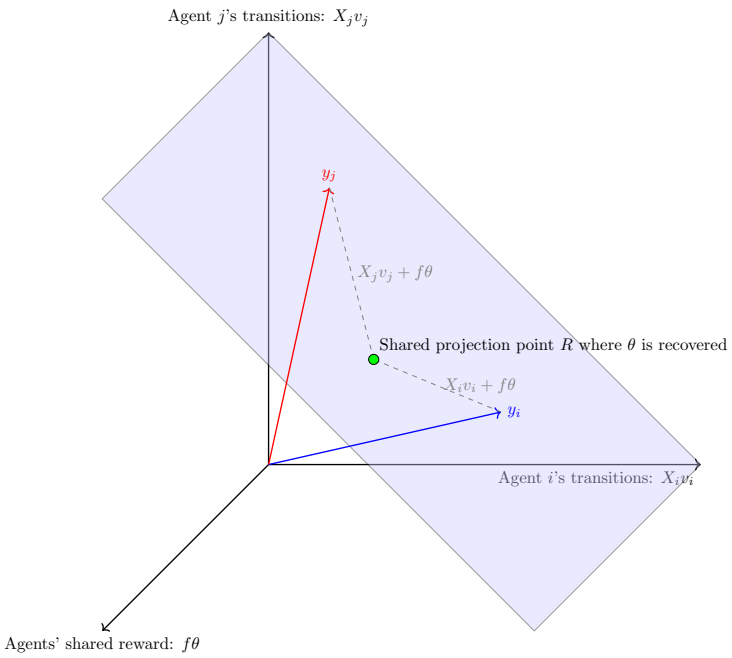


Figure 1: Illustration of reward recovery

The figure illustrates the reward recovery as the task of finding the shared projection point R, 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 y_i and y_j are projected. This plane represents the span of the agent-specific contributions, allowing us to see how the choices y_i and y_j interact with the shared feature $f\theta$.

□

Illustration. Figure 1 provides a visualization of Theorem 2. 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 R, which represents the same combination of $X_i v_i + f\theta$ and $X_j v_j + f\theta$. The point R is the target projection where we recover θ, showing how it aligns with the features and observed outcomes.

Analytic expressions. To obtain analytic expressions based on these results, note that the feature weights θ and the value functions $v_i, i = 1, ..., I$, 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 (7) and

$$\mathbf{M}_f = I - f(f^\top f)^{-1} f^\top, \tag{20}$$

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

$$A/D = 2X_i^\top X_i - X_i^\top X_j (X_j^\top X_j)^{-1} X_j^\top X_i. \tag{21}$$

Expression (21) 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 $\mathbf{P}_{ij} x = \operatorname{argmin}_{y \in \text{range}(X_i)} \|x - y\|_D^2$. Then under this norm:

$$\begin{array}{lll} \text{Projection relative to } \|\cdot\|_D\text{:} & \mathbf{P}_{ij} & = X_i \left(A/D\right)^{-1} X_i^\top, \\ \text{Residuals relative to } \|\cdot\|_D\text{:} & \mathbf{M}_{ij} & = I - X_i \left(A/D\right)^{-1} X_i^\top. \end{array} \tag{22}$$

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

$$\begin{array}{rcl} \Omega/A & = & f^\top \mathbf{M}_{ij} f, \\ \Omega/F & = & \begin{pmatrix} X_i^\top (I + \mathbf{M}_f) X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}. \end{array} \tag{23}$$

The double Schur complement of Ω with respect to F and D will also be important and equals $(\Omega/F)/D = X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i$. 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*

$$\theta = (f^\top \mathbf{M}_{ij} f)^{-1} f^\top \left(\mathbf{M}_{ij} y_i + \mathbf{P}_{ij} \mathbf{M}_j (y_j - y_i)\right). \tag{24}$$

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

$$v_i = a_v + b_v \left(X_i^\top \left(\mathbf{M}_j + \mathbf{M}_f\right) X_i\right)^{-1} X_i^\top \left((\mathbf{M}_j + \mathbf{M}_f) y_i - \mathbf{M}_j y_j\right). \tag{25}$$

Proof. See Appendix. □

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 involving agents interacting with their environment. The mathematical tools needed in Proposition 2 are typical in econometrics and statistical learning for handling confounding influences and separating out the effects of interest. The end goal is to estimate the parameters θ that best describe how agents' decisions are influenced by their context and actions.

1.4 A step-by-step example of recovering corporate purpose

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

A benchmark model for recovery: Neoclassical investment with exploration. We extend the basic investment model by Strebulaev and Whited (2012) in discrete time with an infinite horizon ($t = 0, 1, \ldots, \infty$) to stochastic 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 capital). The manager maximizes firm value

$$V_\lambda^*(s) = \max_{\{a_t\}_{t\geq 0}} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (u(s_t, a_t) - \lambda \ln(\pi(a_t|s_t)))], \tag{26}$$

subject to the laws of motion for k' and $\ln(z')$ and given the regularization coefficient $\lambda > 0$. The policy function $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is stochastic, where $\mathcal{P}(\cdot)$ denotes the set of all probability distributions, as opposed to deterministic $\pi : \mathcal{S} \to \mathcal{A}$ in the benchmark model in Strebulaev and Whited (2012). As $\lambda \to 0$, the entropy term in (26) 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

$$u(k, z, k') = \phi(k, z) - I(k, k') - \psi(I, k'), \tag{27}$$

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 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,

$$\psi(I, k) = \begin{cases} 0 & \text{, if no adjustment costs,} \\ \frac{\psi_0}{2}\frac{I^2}{k} + \psi_1 k \mathbb{1}_{\{I \neq 0\}} & \text{, if adjustment costs,} \end{cases} \tag{28}$$

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 (27) and (28) yields an expression for the cash flow function

$$u(k, z, k') = \begin{cases} zk^{\zeta} + (1-\delta)k - k' & \text{, if no adjustment costs} \\ zk^{\zeta} + (1-\delta)k - k' - \frac{\psi_0}{2}\frac{I^2}{k} - \psi_1 k \mathbb{1}_{\{I \neq 0\}} & \text{, if adjustment costs} \end{cases} \tag{29}$$

In state-action notation, we can write state $s = (s_1, s_2)$ with state variables $s_1 = k, s_2 = z$ and action $a = k' (= s_1')$.[10] Expressing the cash flow function (29) in terms of states and actions yields the reward function

$$u(s, a) = \begin{cases} \underbrace{s_2 s_1^{\zeta}}_{\text{Profits}} - \underbrace{\delta s_1}_{\text{Depreciation}} - \underbrace{(a - s_1)}_{\text{Net investment}} & \text{, if no adjustment costs,} \\ \underbrace{s_2 s_1^{\zeta}}_{\text{Profits}} - \underbrace{\delta s_1}_{\text{Depreciation}} - \underbrace{(a - s_1)}_{\text{Net investment}} - \underbrace{\frac{\psi_0}{2}\frac{(a - (1-\delta)s_1)^2}{s_1}}_{\text{Feature } f_4} - \underbrace{\psi_1 s_1 \mathbb{1}_{\{a-(1-\delta)s_1 \neq 0\}}}_{\text{Feature } f_5}, & \text{otherwise.} \end{cases} \tag{30}$$

[10] In 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.

One can see from (30) that the reward function, while non-linear in the state and action variables, is linear in three or five features depending on the case.

Feature selection. In line with Assumption 2, we express the reward function (30) 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 illustrative 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.

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

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: $\zeta = 0.55, \delta = 0.2, \rho = 0.7, r = 0.05, \sigma_\epsilon = 0.04, \psi_0 = 0.01, \psi_1 = 0.01, n_k = 15, d = 3, n_z = 5, m = 2, \lambda = 0.2, \sigma_\epsilon^1 = 0.035, \sigma_\epsilon^2 = 0.045, \gamma^1 = 0.94, \gamma^2 = 0.96, n_{steps} = 1,000,000,$ 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 ($\boldsymbol{\theta}$)
1	2	$[\ s_2 s_1^\zeta \quad a-(1-\delta)s_1\ ]$	$[\ 1 \quad -1\ ]$
2	3	$[\ s_2 s_1^\zeta \quad s_1 \quad a\ ]$	$[\ 1 \quad 1-\delta \quad -1\ ]$
3	4	$[\ s_2 s_1^\zeta \quad a-(1-\delta)s_1 \quad \frac{(a-(1-\delta)s_1)^2}{s_1} \quad s_1 \mathbb{1}_{\{a-(1-\delta)s_1 \neq 0\}}\ ]$	$[\ 1 \quad -1 \quad -\frac{\psi_0}{2} \quad -\psi_1\ ]$
4	2	$[\ s_2 s_1^{\zeta-1} \quad \frac{a-(1-\delta)s_1}{s_1}\ ]$	$[\ 1 \quad -1\ ]$
5	4	$[\ s_2 s_1^{\zeta-1} \quad \frac{a-(1-\delta)s_1}{s_1} \quad \left(\frac{a-(1-\delta)s_1}{s_1}\right)^2 \quad \mathbb{1}_{\{a-(1-\delta)s_1 \neq 0\}}\ ]$	$[\ 1 \quad -1 \quad -\frac{\psi_0}{2} \quad -\psi_1\ ]$

Panel B: Reward recovery in controlled environment

Specification	Experiment	π	T	True Weights ($\boldsymbol{\theta}$)	Recovered Weights ($\hat{\boldsymbol{\theta}}$)
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	$\{1, -1, -0.005, -0.01\}$	$\{0.981, -1.03, 0.004, -0.013\}$

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 $s \in \mathcal{S}$,

$$V_\lambda^*(s) = \lambda \ln \sum_{a \in A} e^{\frac{1}{\lambda} Q_\lambda^*(s,a)}. \tag{31}$$

and the maximum is achieved by the randomized policy

$$\pi_\lambda^*(a|s) = \frac{e^{\frac{1}{\lambda} Q_\lambda^*(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\frac{1}{\lambda} Q_\lambda^*(s,a')}} \text{ for } a \in \mathcal{A}. \tag{32}$$

Recovering the feature weights. We now illustrate our recovery theorem by implementing it in a 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. In unreported results, we confirm that the recovered coefficients converge to the true coefficients as $n_{steps} \to \infty$, as one would expect.

2 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.

2.1 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 this data with executive compensation metrics from Execucomp, which are available from 1992 onward, as well as environmental, social and governance scores from MSCI, which are available from 2007 onward.[11]

2.2 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). We impute missing values of amortization of intangibles (`am`), cash and cash equivalents (`che`), short-term debt (`dlc`), long-term debt (`dltt`), 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 2 shows the summary statistics for all variables. We provide descriptive statistics across four distinct panels and datasets: Compustat (1975-2021), Peters & Taylor (1975-2021), ExecuComp (1992-2021), and MSCI (2007-2021). The summary highlights that the key variables such as Size, Profitability, Investment Rate, and ESG Scores among others are in line with the prior literature. Panel A shows that average firm size is 4.66 with a standard deviation of 2.24 (median size is 4.52). On average, profitability is 0.51 with large variability evident from its standard deviation of 0.91. The main corporate actions, investment rate and net debt to EBITDA, display averages of 0.24 and 1.78 respectively, indicating varying levels of corporate investment and debt management practices. Average tangibility is reported at 0.34, and Tobin's q averages at 1.56. In Panel B, ln(Total Capital) is 6.06 and alternative profitability is 0.21, indicating

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

Table 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
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
Panel C: ExecuComp (1992-2021)								
CEO Bonus (%)	30,481	0.03	0.06	0.00	0.00	0.00	0.00	0.04
CEO Ownership	29,647	0.02	0.05	0.00	0.00	0.00	0.00	0.02
CEO Own. & Options	29,642	0.03	0.05	0.00	0.01	0.01	0.01	0.03
CEO Own. & Opt. 2	29,642	0.03	0.06	0.00	0.01	0.01	0.02	0.04
Mana. Bonus (%)	30,481	0.07	0.13	0.00	0.00	0.01	0.03	0.10
Mana. Ownership	30,420	0.03	0.07	0.00	0.00	0.01	0.01	0.03
Mana. Own. & Options	30,420	0.05	0.08	0.01	0.01	0.02	0.03	0.05
Mana. Own. & Opt. 2	30,420	0.05	0.08	0.01	0.02	0.03	0.04	0.07
Panel D: 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

substantial intangible capital and modest profitability. Total q has a lower average compared to traditional Tobin's q, with a mean of 0.88. In Panel C, the ExecuComp CEO and management compensation metrics highlight both fixed (bonuses) and variable (ownership and options) compensation elements. Ownership structures reflect a relatively modest level of stock holding by CEOs and management, indicating the potential for incentive misalignment. In Panel D, the ESG scores from MSCI (2007-2021) have averages around 4.58 to 5.38, reflecting moderate to good ESG performance across the typical firm. The standard deviation of the ESG scores is quite large, ranging between 1.58 for S to 2.11 for E.

2.3 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. We refer to one specification of our model as a combination of 3 elements, namely (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 our two experts. Conceptually, we think of expert definitions as 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.[12] 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 choose to represent states as combinations of state variables, e.g., pairs in the case of two state variables, and actions as combinations of action variables. We choose to 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 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 one parsimonious set of features is one of the main empirical challenges we face in this paper. Consequently, our feature set will vary across specifications.

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

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 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 $y = 1, \ldots, J$ 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

$$P_i(y = j|\boldsymbol{x}_i) = \begin{cases} \frac{exp(\boldsymbol{x}_i\boldsymbol{\beta}_{i,j})}{1+\sum_h^J exp(\boldsymbol{x}_i\boldsymbol{\beta}_{i,h})} & \text{, if } j = 2, \ldots, J \\ \frac{1}{1+\sum_h^J exp(\boldsymbol{x}_i\boldsymbol{\beta}_{i,h})} & \text{, if } j = 1 \end{cases} \tag{33}$$

where $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i = [1, Size_i, Profitability_i, Size_i \times Profitability_i, Size_i^2, Profitability_i^2]$.

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

$$\ell_{i,n}(\boldsymbol{\beta}_i) = \sum_{j=0}^{J} \mathbb{1}_{\{y_{i,n}=j\}} \log[p_{i,j}(\boldsymbol{x}_i, \boldsymbol{\beta}_i)] \tag{34}$$

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 $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i = [1, Size_i, Profitability_i, InvestmentRate_i, Size_i \times Profitability_i, Size_i \times InvestmentRate_i, Profitability_i \times InvestmentRate_i, Size_i^2, Profitability_i^2, InvestmentRate_i^2]$.

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 (14) and test if Theorem 2 holds, i.e., whether $rank(X) = 2|\mathcal{S}| + d$. 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 (24), the reward from $\hat{u} = f\hat{\theta}$, and the value $\hat{v}_i$ based on (25) (for $i, j = 1, 2$ with $i \neq j$).

3 Recovering Corporate Purpose

We now recover the corporate objective function given different specifications of experts, state variables, action variables and features.

3.1 Corporate reward puzzle: Reward recovery in foundational model

At its core, the neoclassical theory of investment states that the reward of the corporate investment problem should consist of profits (Π), net of the investment (I) 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 $f = [\frac{\Pi}{K}, \frac{I}{K}]$, $\theta = [1, -1]^T$, Π/K denotes profitability and I/K denotes the investment rate.

Table 3 shows the results of applying our recovery theorem to empirical data on managers' decisions. Expert 1 corresponds to the time period from 1975 until 1998, whereas expert 2 corresponds to the time period from 1999 until 2021. The year-based expert definitions were chosen such that the full sample is split into are a roughly equal number of years and observations for each expert. 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, which is based on the average risk-free rate during that time period. We opt for a relatively coarse grid of 5 size bins, 5 profitability bins and 5 investment rate bins. This results in 25 possible 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 report both the raw and normalized (in parantheses) recovered coefficients of the reward function. Columns (1) to (6) differ in our choice of the features f.[13]

[13] We normalize the coefficients by the absolute value of the coefficient on profitability, such that profitability always has a normalized coefficient of 1 (or -1) and the magnitude of all other coefficients is expressed relative to profitability. The 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 $\lambda = 1$ and are proportional to λ. In other words, λ scales the coefficients up or down. For example, if

Table 3: Reward recovery from corporate investment data.

The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include profitability, investment rate, tangibility, cash holdings, book leverage, and net book leverage. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 size (BoY) bins, 5 profitability (BoY) bins, and for the action variable 5 investment rate bins. The discount factors are $\gamma_1 = 0.923$ and $\gamma_2 = 0.967$.

	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Profitability	4.649	7.580	4.483	3.675	4.646	7.453
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Investment Rate	-1.193	-3.314	-1.181	-0.721	-1.192	-3.773
	(-0.257)	(-0.437)	(-0.263)	(-0.196)	(-0.256)	(-0.506)
Tangibility		8.882				14.704
		(1.172)				(1.973)
Cash Holdings			-12.06			
			(-2.690)			
Book Leverage				-3.650		
				(-0.993)		
Net Book Leverage					-0.013	-8.515
					(-0.003)	(-1.142)
$\overline{KLD}_1$	0.248	0.204	0.249	0.242	0.248	0.184
$\overline{KLD}_2$	0.096	0.115	0.092	0.109	0.096	0.142
N_1	59,743	59,743	59,743	59,743	59,743	59,743
N_2	49,836	49,836	49,836	49,836	49,836	49,836

Column (1) reports our results for the simplest possible specification with two features. We find a normalized coefficient of -0.257 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 one fourth of what theory would prescribe. For a given amount of capital, managers act as if \$1 of investment would only cost them \$0.257. This finding suggests one of two possibilities: either the model is correct and managers' preferences are heavily non-standard or the benchmark model is too simple and fails to accurately describe the data. Assuming the latter explanation, we investigate possible limitations of the base model.

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

we assume that $\lambda = 10$, 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. In other words, we are more confident that profitability has a coefficient of one than we are that λ equals one. For this reason, we focus on normalized coefficients in the rest of our analysis.

functions are to the ones we observe for the real CEOs, we compute, for each state s, the Kullback-Leibler divergence between the recovered policy function $\hat{\pi}_i$ and the policy function π_i of expert i as

$$KLD_i(\hat{\pi}_i, \pi_i \mid s) = \sum_{a \in \mathcal{A}} \hat{\pi}_i(a \mid s) \log \left(\frac{\hat{\pi}_i(a \mid s)}{\pi_i(a \mid s)} \right), \tag{35}$$

and take the average across states to arrive at the mean Kullback-Leibler divergence

$$\overline{KLD}_i = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} KLD_i. \tag{36}$$

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

The mean KL divergence in our base model for expert 1 equals 0.248 and is more than double than that for expert 2, which equals 0.096. The KL divergence should not be interpreted as a "distance measure" between the distributions, but rather as a measure of entropy increase due to the use of an approximation to the true distribution rather than the true distribution itself. In other words, the KL divergence measures the increase in entropy in $\hat{\pi}$ over and above the unavoidable entropy of distribution π. In economic terms, the $\overline{KLD} = 0.248$ (0.096) means the entropy increases by 24.8% (9.6%).

Financial factors entering the manager's reward. Columns (2) to (6) of Table 3 show the results for different feature sets that include tangibility, cash holdings, book leverage, and net book leverage. One limitation of the base model may be that it only has two features, whereas managers may take into account additional features. The additional features are commonly used in empirical studies to account for differences in firms and, as such, it is plausible that they affect the reward function. We therefore include them as additional features. All models consistently assign positive weights to profitability, indicating its significant positive impact on the manager's reward function. The coefficients for investment rate are negative, suggesting an inverse relationship with the manager's reward function. However, the magnitude varies across models, indicating some variability in its influence. As we add additional features, the coefficient on the investment rate increases (in absolute value) from -0.257 in column (1) to -0.506 in

column (6), but still falls short of the -1 posited by the theory.

Financial factors, including cash holdings, book leverage, and net book leverage, also contribute to the reward function, although their impact appears less consistent across models. The coefficients vary in magnitude and sign, suggesting a nuanced relationship with the reward. The normalized coefficient on tangibility is positive at 1.172 and 1.973, respectively, whereas the ones on cash holdings, book leverage and net book leverage are negative. Tangibility's positive influence on the reward function receives substantial weights; however, the normalized coefficients show considerable variation, indicating its importance varies across specifications. Similar, book leverage has a normalized weight of -1 (column 4), while net book leverage has a normalized weight of -1.142 (column 6).

As we add features, the mean KL divergence of expert 1 decreases from 0.248 in column (1) to 0.184 in column (6), whereas the one of expert 2 increases from 0.096 to 0.142. This suggests that the additional features help explain the policy choices of expert 1, but not expert 2.

Corporate reward puzzle. The *corporate reward puzzle* documented in Table 3 is that managers use $p \approx 0.5$ when deciding investment. We explore the impact of this corporate reward puzzle through the lens of the foundational q-theory model from Section 1.4. We use this model 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 3 show that $p = 0.45$-0.55.

To explore the impact of $p < 1$ 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 4 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.

Table 4: 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.

		Investment rate ($p = 0.5$)			Benchmark investment rate ($p = 1$)		
Profitability		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

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 de-facto 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. This corporate reward puzzle is, hence, inconsistent with the predictions from the foundational q-theory model and supports alternative theories predicting over-investment and under-exploration.

Over-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. 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), but under-exploration has not. In studies with related findings, Ben-David and Chinco (2024) demonstrate that a 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, which is also consistent with over-investment but silent about exploration/experimentation.

3.2 Taking account of intangible capital

A limitation of the base model is its inability to account for the growing importance of intangible capital. Peters and Taylor (2017) provide 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.

Table 5 presents the results of our analysis after using these alternative measures. Intangible capital does not resolve the corporate reward puzzle from Table 3. We still find that managers act as if investment were half as costly as theory posits, with coefficients ranging from -0.478 in column (1) to -0.496 in column (6). The coefficient for tangibility is relatively small, ranging from 0.025 to 0.031. This suggests that while tangibility may have some influence on investment decisions, its impact is relatively minor compared to profitability and investment rate. The coefficient for cash holdings is -0.042, indicating a relatively small impact on the reward function. The coefficients for book leverage and net book leverage are negative and relatively small in magnitude, ranging from -0.411 to -0.142. This suggests that higher levels of leverage are associated with lower rewards. Overall, these coefficients provide insights into the factors driving corporate investment decisions, highlighting the importance of profitability and investment costs while also considering the influence of other financial indicators such as tangibility and leverage.

Surprisingly, adding additional features does not improve the overall fit of the model to the data in terms of mean KL divergence when we use these alternative measures. The KL divergence ranges from 0.400 to 0.421 for Expert 1 and from 0.546 to 0.623 for Expert 2 across the different model specifications. These values suggest that the models exhibit a reasonable but not very close level of alignment with the true policies, indicating their limited effectiveness in capturing the decision-making dynamics of corporate investment. Hence, a more detailed analysis incorporating other features is needed to improve the models'

Table 5: Reward recovery from investment in tangible and intangible capital.
The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, cash holdings, book leverage, and net book leverage. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the action variable 5 total investment rate bins. The discount factors are $\gamma_1 = 0.923$ and $\gamma_2 = 0.967$.

	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	32.236	34.7	32.637	23.279	27.439	30.033
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-15.416	-16.301	-15.534	-12.279	-13.947	-14.882
	(-0.478)	(-0.470)	(-0.476)	(-0.527)	(-0.508)	(-0.496)
Tangibility		0.866				0.933
		(0.025)				(0.031)
Cash Holdings			-1.377			
			(-0.042)			
Book Leverage				-9.559		
				(-0.411)		
Net Book Leverage					-4.223	-4.278
					(-0.154)	(-0.142)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.414	0.409	0.411	0.400	0.421	0.419
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.561	0.593	0.546	0.588	0.623	0.662
N_1	59,745	59,745	59,745	59,745	59,745	59,745
N_2	49,836	49,836	49,836	49,836	49,836	49,836

predictive accuracy and robustness.

3.3 Role of managerial compensation

To better understand the source for the corporate reward puzzle documented in Table 3, we now investigate the effect of managerial compensation. 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.

Table 6 shows the results when we add CEO compensation.[14] In columns (1) and (2), we repeat the previous analysis to show the recovered coefficients with updated expert definitions. Looking at column (1) of Table 6 and comparing it to column (1) of Table 5, the coefficient on the investment rate has increased (in absolute value) to -0.562 (compared to -0.478 before). The mean KL divergence of expert 1 has decreased to 0.379 (compared to 0.414 before), whereas the one of expert 2 has stayed almost the same at 0.562 (compared to 0.561 before). This indicates a better fit for expert 1 and overall. These findings

[14]The 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 6: Reward recovery including CEO compensation.

The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, and CEO compensation. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the action variable 5 total investment rate bins. The discount factors are $\gamma_1 = 0.947$ and $\gamma_2 = 0.975$.

	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	26.408	24.033	24.247	24.264	24.359	24.655
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-14.847	-14.303	-14.474	-14.398	-14.355	-14.543
	(-0.562)	(-0.595)	(-0.597)	(-0.593)	(-0.589)	(-0.590)
Tangibility		0.232	0.584	0.104	0.236	0.6
		(0.01)	(0.024)	(0.004)	(0.01)	(0.024)
Net Book Leverage		-2.46	-3.057	-2.416	-2.347	-2.937
		(-0.102)	(-0.126)	(-0.100)	(-0.096)	(-0.119)
CEO Bonus			-5.729			-5.915
			(-0.236)			(-0.240)
CEO Ownership				10.247		
				(0.422)		
CEO Own. & Opt.					3.448	4.241
					(0.142)	(0.172)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.379	0.382	0.368	0.380	0.383	0.368
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.562	0.606	0.601	0.607	0.609	0.604
N_1	39,305	39,305	39,305	39,305	39,305	39,305
N_2	29,093	29,093	29,093	29,093	29,093	29,093

suggest that the corporate objective function has changed over time.

In column (2), we add our control features and in columns (3) to (5), we introduce the CEO bonus, CEO ownership and CEO ownership & options one-by-one. In column (6), we find that CEO bonus has a negative coefficient of -0.240, whereas CEO ownership and options has a positive coefficient of 0.172. The negative coefficient on CEO bonus suggests that the manager is not acting purely in their self-interest. The positive coefficient on CEO ownership and options suggests that this feature is desirable, either because the CEO likes control or because the shareholders like to align the CEO's incentives with their own. Adding CEO compensation features improves the model's performance in terms of mean KLD for expert 1, but worsens it for expert 2.

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. Appendix E.3 shows that the results are qualitatively unchanged if we use top managerial instead of CEO compensation.

Overall, the results in this section suggest that the corporate recovery theorem from Section 1 can be used to recover the manager's reward function. We find that 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 intangible capital and features of managerial compensation packages do not resolve this corporate reward puzzle. 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 considerations.

4 Alignment of Manager's Reward with Shareholder-Value Maximization and the Role of ESG

Corporate finance theory prescribes that managers should maximize shareholder value. Absent non-pecuniary benefits in the reward function, shareholder value maximization boils down to shareholder wealth maximization, i.e., the maximization of the market value of equity. Absent bondholder expropriation, shareholder wealth maximization amounts to firm value maximization, i.e., the maximization of the market value of equity and debt. Scaling everything by capital, this implies maximizing Tobin's q, i.e., the market value of equity and debt divided by the book value of total assets. Our methodology allows us to recover the value that managers actually maximize in practice. We can thus compare the manager's recovered value $\hat{v}$ to Tobin's q and Total q which measure the shareholders' valuation of the firm.

4.1 Alignment measure

We recover expert 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. 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:[15]

$$Alignment_i = Corr\left(\{\hat{v}_i(s)\}_{s\in\mathcal{S}}, \{Median(q_i(s))\}_{s\in\mathcal{S}}\right), \tag{37}$$

for $i = 1, ..., I$. If the two were perfectly aligned, the correlation and, hence, the alignment measure will be 1.

Table 7 shows the correlation between the recovered value of expert i, $\hat{v}_i$, and Tobin's q (in Panel A)

[15] While we can recover v_i only up to affine transformations, this should not affect the correlation $Corr(q_i, v_i) = Corr(q_i, a_v + b_v\hat{v_i}) = sign(b_v)Corr(q_i, \hat{v_i})$ as it is unlikely that $b_v < 0$.

or Total q (in Panel B). Panel A computes the alignment between manager's value and Tobin's q based on the recovered value from the specification in Table 3. The different columns in Panel A replicate the specifications in columns (1) to (6) of Table 3. Panel B computes the alignment between manager's value and Total q based on the recovered value from the specification in Table 5. We measure Tobin's q at the end of year t. To mitigate possible concerns related to the timing of the action within the year or the incorporation of information into market prices, we also report correlations for q measured at the beginning of year $(t-1)$ and end of next year $(t+1)$, and show that the timing does not materially affect the results.

In column (1), we find that the degree of alignment between the manager's value and Tobin's q is positive and equal to 0.716 for expert 1, but -0.519 for expert 2. The latter suggests that managers' incentives are perversely distorted, with managers gaining from lower market valuations. This is likely due to the fact that the specification is incomplete. In columns (2)-(6), when we include additional features, we find that the degree of alignment between the manager's value and Tobin's q improves. In column (6) it equals to 0.756 for expert 1, but still only 0.082 for expert 2. This suggests that the objective function of the firm has become less aligned with firm value maximization over time, as expert 1 is defined as the time period from 1975 to 1998 and expert 2 is defined as the time period from 1999 until 2021.

A potential explanation for this finding is that, over time, intangible capital has become more important, but this is not properly accounted for in the standard measure of q in the specification in Table 3. In Panel B, we repeat the analysis with our alternative measures that incorporate intangible capital. Columns (1) to (6) correspond to the specifications in columns (1) to (6) of Table 5. Incorporating intanglible capital we find that the correlation between the recovered value and Total q is 0.385 for expert 1 and 0.473 for expert 2 in column (6). This suggests incorporating intangible capital helps us in finding a more reasonable positive alignment between manager rewards and firm value maximization. The correlations are, however, still far from perfect which is what one would expect if managers maximize shareholder value.

4.2 ESG and non-pecuniary benefits

Next, we investigate the effect of non-pecuniary benefits on corporate decision-making and the alignment between manager and shareholder value. 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

Table 7: Alignment between recovered value and q without ESG considerations.
The table documents the alignment between the recovered value function and Tobin's q (Panel A) as well as Total q (Panel B). Panel A computes the alignment between manager's value and Tobin's q based on Table 3. Panel B computes the alignment between manager's value and Total q based on Table 5. † indicates a violation of the economic prior.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Alignment between manager's value and Tobin's q						
Features included:						
Profitability	X	X	X	X	X	X
Investment Rate	X	X	X	X	X	X
Tangibility		X				X
Cash Holdings			X			
Book Leverage				X		
Net Book Leverage					X	X
Alignment 1975-1998	0.712	0.846	0.835	0.615	0.712	0.756
Alignment 1999-2021	-0.519†	0.144	0.001	-0.588†	0.111	0.082
1975-1998						
Corr($\hat{v}_1, q_{1,t-1}$)	0.716	0.825	0.826	0.625	0.716	0.745
Corr($\hat{v}_1, q_{1,t+1}$)	0.695	0.859	0.832	0.590	0.695	0.758
1999-2021						
Corr($\hat{v}_2, q_{2,t-1}$)	-0.491†	0.186	0.047	-0.564†	0.153	0.125
Corr($\hat{v}_2, q_{2,t+1}$)	-0.539†	0.161	0.033	-0.612†	0.135	0.093
Panel B: Alignment between manager's value and Total q						
Features included:						
Alt. Profitability	X	X	X	X	X	X
Total Investment Rate	X	X	X	X	X	X
Tangibility		X				X
Cash Holdings			X			
Book Leverage				X		
Net Book Leverage					X	X
Alignment 1975-1998	0.471	0.478	0.474	0.276	0.377	0.385
Alignment 1999-2021	0.443	0.430	0.523	0.187	0.485	0.473
1975-1998						
Corr($\hat{v}_1, q^{tot}_{1,t-1}$)	0.539	0.546	0.542	0.354	0.451	0.459
Corr($\hat{v}_1, q^{tot}_{1,t+1}$)	0.364	0.372	0.367	0.157	0.263	0.272
1999-2021						
Corr($\hat{v}_2, q^{tot}_{2,t-1}$)	0.454	0.440	0.535	0.201	0.495	0.483
Corr($\hat{v}_2, q^{tot}_{2,t+1}$)	0.353	0.340	0.437	0.087	0.397	0.386

and social scores should have little to no effect. 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 8 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 -0.611, i.e., closer to the theoretical benchmark. The KL divergence decreases to 0.268 for expert 1 and 0.503 for expert 2, indicating a better fit to the data from 2007 onward. The correlation between the recovered value and Total q decreases for expert 1 (Corr=0.20),

Table 8: Do ESG considerations affect corporate purpose?

The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, and ESG scores. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the action variable 5 total investment rate bins. The discount factors are $\gamma_1 = 0.971$ and $\gamma_2 = 0.980$.

Panel A: Feature importance in reward recovery specification $u = f\theta$					
	Recovered feature importance θ ($\theta_{\text{normalized}}$)				
	(1)	(2)	(3)	(4)	(5)
Alt. Profitability	21.833	22.749	21.589	22.505	21.662
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-13.344	-13.464	-13.335	-13.572	-13.300
	(-0.611)	(-0.592)	(-0.618)	(-0.603)	(-0.614)
Tangibility		-0.024	0.141	0.093	0.091
		(-0.001)	(0.007)	(0.004)	(0.004)
Net Book Leverage		0.927	0.657	0.939	0.976
		(0.041)	(0.030)	(0.042)	(0.045)
ESG Score			0.231		
			(0.011)		
Adj. ESG Score				0.073	
				(0.003)	
Environmental Score					-0.010
					(-0.000)
Social Score					0.189
					(0.009)
Governance Score					0.286
					(0.013)
$\overline{KLD}_1$	0.268	0.271	0.265	0.269	0.258
$\overline{KLD}_2$	0.503	0.492	0.478	0.492	0.482
N_1	18,518	18,518	18,518	18,518	18,518
N_2	10,575	10,575	10,575	10,575	10,575
Panel B: Alignment between recovered value $\hat{v}$ and Total q					
Alignment 2007-2014	0.204	0.428	0.508	0.548	0.773
Alignment 2015-2021	0.509	0.837	0.797	0.889	0.802
2007-2014					
Corr($\hat{v}_1, q^{tot}_{1,t-1}$)	0.252	0.469	0.543	0.58	0.793
Corr($\hat{v}_1, q^{tot}_{1,t+1}$)	0.081	0.318	0.411	0.459	0.711
2015-2021					
Corr($\hat{v}_2, q^{tot}_{2,t-1}$)	0.502	0.837	0.798	0.888	0.807
Corr($\hat{v}_2, q^{tot}_{2,t+1}$)	0.474	0.844	0.809	0.887	0.823

but increases for expert 2 (Corr=0.51) compared to Table 5.

In column (2), we add tangibility and our financial features. Then in columns (3) to (6), we add the weighted-average ESG score, industry-adjusted ESG score and individual E, S, and G scores. We find positive coefficients of 0.011 for the weighted-average ESG score and 0.003 for the industry-adjusted ESG score, respectively. Economically, a one standard deviation (1-SD) change in the weighted-average ESG score (SD=0.99) changes the reward by 1.1% which is a small to moderate effect. A 1-SD change in the industry-adjusted ESG score (SD=2.15) changes the reward by 0.65% which is a small effect. Breaking up the ESG score into its individual components has a more significant impact. Interestingly, the environmental score by itself has no impact at all. We recover normalized coefficients of 0.009 on the

social score and 0.013 on the governance score. 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 S score (SD=1.58) changes the reward by 1.4%. A 1-SD change in the G score (SD=1.97) changes the reward by 2.6%. 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.

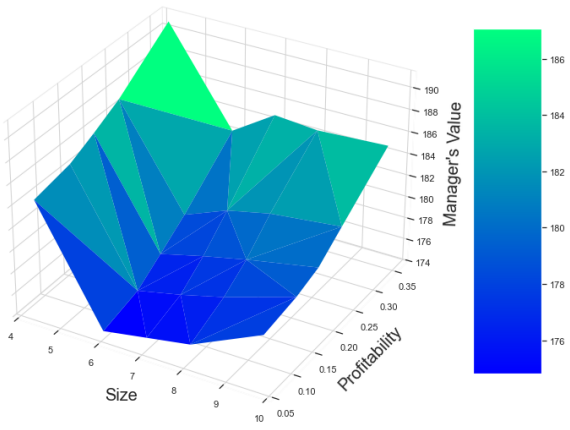
Including ESG scores improves the model's performance in terms of mean KL divergence, which decreases from 0.268/0.503 in column (1) to 0.258/0.482 in column (6) for experts 1/2, respectively. This improvement in model performance is moderate. From this perspective, ESG does not appear to be of large economic relevance for firms. However, this interpretation is incomplete.

Incorporating ESG scores into the manager's reward materially affects our interpretation of manager with shareholder alignment. The correlation between the recovered value and Total q increases substantially, from 0.204/0.509 in column (1) to 0.773/0.802 in column (6) for experts 1/2, respectively. This increase in alignment suggests that environmental, social, and governance considerations play a substantial role in firms' decision-making and suggests strong alignment of the reward function of the manager with the valuation of the firm in the market. Absent ESG considerations, manager incentives and market valuations exhibit a low correlation. However, incorporating ESG considerations increases the correlation to about 80% which is still less than unity but changes our interpretation whether manager incentives are aligned with shareholder/market valuations. Our results suggest that firms maximize shareholder value particularly because S and G considerations enter the firm's objective.

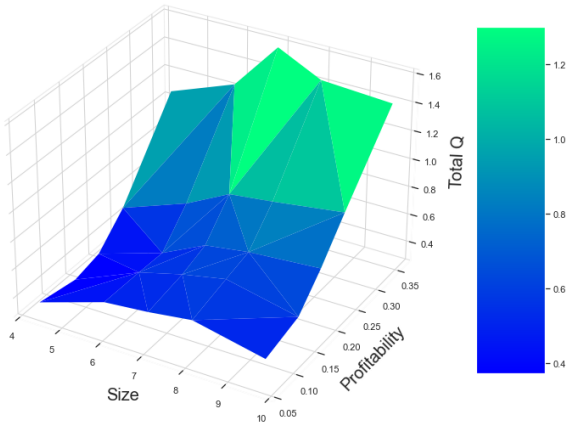
4.3 Misalignment

The misalignment between the manager's value and shareholders' valuation of the company 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.

Figure 2 illustrates the recovered value for the manager for each firm size-profitability combination in



(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 2: The manager's recovered value (left) and company valuation (right) for different firm sizes and profitability levels.

the left plot. 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 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, the manager's is U shaped in firm size. The manager's value is largest for small, highly profitable firms and lowest for unprofitable mid-sized firms. The large increase in manager value for the smallest firm size category is particularly astonishing. In contrast, Total q is inverse U shaped in firm size. Mid-sized firms have the largest q while the smallest and largest firms have the lowest q. Most notably, the manager's value and firm value are inversely related to each other in the firm size dimension.[16]

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 for small firms.

[16]During the later period, the alignment has improved because the manager's value is now monotonic in firm size and profitability (using specification 5), as one would expect.

4.4 Extensions of the analysis and robustness

We consider several extensions of our analysis in Appendix D. 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. We therefore investigate first 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.

Table D.1 provides the results. Overall, our main results remain largely unchanged. Managers still act as if investment were less costly than prescribed by theory, with shadow cost coefficients ranging from 0.54 to 0.61. Adding managerial compensation, ESG scores and higher discount rates improves the model's explanatory performance. 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. Moreover, the coefficient on governance becomes negative, indicating that the manager would prefer bad governance when making leverage decisions in addition to investment decisions.

Next, we explore the impact of managerial preferences. Table D.2 shows that higher discount rates help explain managers' policies and align their recovered value with shareholders. As we add control and ESG-related features, *Alignment* increases to 0.675 (2007-2014) and 0.935 (2015-2021), respectively. Again, the results suggest that manager-shareholder alignment is imperfect, but it has increased over time.

Last, we explore the robustness of our results in Appendix E. In this section, we vary the estimation approach for choice and transition probabilities, the coarseness of the state and action variable grid, as well as rely on alternative definitions of managerial compensation. Overall, the results remain unchanged.

5 Conclusion

This paper uses a machine learning technique, inverse reinforcement learning, to recover the reward function(s) that managers use in deciding firms' investment and financial policies. We first derive conditions for a corporate recovery theorem that establishes when and under what conditions the recovery is unique. We then apply the method to a large dataset on U.S. public corporations including financial accounting information, intangible capital, managerial compensation, and ESG scores for the period 1975–2021.

We document several novel facts. First, managers' reward function can be recovered from the data

without imposing parametric assumptions. We find that managers act as if they underestimate the cost of investment. Managers' subjective cost of investment is about 45-65 cents on the dollar. Incorporating intangible capital improves the model performance but does not resolve this corporate recovery puzzle. Managerial compensation policies and ESG concerns enter the corporate objective. Managerial rewards significantly improve with social and governance scores. However, we find environmental factors do not materially affect managers' rewards.

The degree of alignment between managers' reward and shareholder-value maximization is reasonably high, with a correlation between manager's value function and Total q is in excess of 80%. The alignment is higher in the later than in the earlier period of our sample, suggesting that improvements in compensation policies and corporate governance have reached their intended goal. Social and governance considerations in the manager's reward are not in conflict with, but crucial ingredients for aligning incentives and maximizing shareholder value.

References

Abel, Andrew B. and Janice C. Eberly (1994) "A Unified Model of Investment Under Uncertainty," *The American Economic Review*, 84 (5).

Baron, David P. (2007) "Corporate Social Responsibility and Social Entrepreneurship," *Journal of Economics & Management Strategy*, 16 (3).

Bebchuk, Lucian A. and Roberto Tallarita (2020) "The Illusory Promise of Stakeholder Governance," *Cornell Law Review*, 106.

Ben-David, Itzhak and Alex Chinco (2024) "Modeling Managers As EPS Maximizers," Working paper.

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.

Bénabou, Roland and Jean Tirole (2010) "Individual and Corporate Social Responsibility," *Economica*, 77 (305).

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.

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.

Elhauge, Einer (2005) "Sacrificing corporate profits in the public interest," *New York University Law Review*, 80.

Ericson, Keith M. Marzilli (2024) "What Do Shareholders Want? Consumer Welfare and the Objective of the Firm," NBER Working Paper No. w32064.

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.

Graff Zivin, Joshua and Arthur Small (2005) "A Modigliani-Miller theory of altruistic corporate social responsibility," *The BE Journal of Economic Analysis & Policy*, 5 (1).

Graham, John R. (1996) "Debt and the marginal tax rate," *Journal of Financial Economics*, 41 (1).

——— (2022) "Presidential Address: Corporate Finance and Reality," *The Journal of Finance*, 77 (4).

Hart, Oliver and Luigi Zingales (2017) "Companies Should Maximize Shareholder Welfare Not Market Value," *Journal of Law, Finance, and Accounting*, 2 (2).

——— (2022) "The New Corporate Governance," *The University of Chicago Business Law Review*, 1 (1).

Hayashi, Fumio (1982) "Tobin's Marginal q and Average q: A Neoclassical Interpretation," *Econometrica*, 50 (1).

Hong, Harrison and Edward Shore (2023) "Corporate Social Responsibility," *Annual Review of Financial Economics*, 15 (1).

Jensen, Michael C. (1986) "Agency costs of free cash flow, corporate finance, and takeovers," *American Economic Review*, 76 (2), 323–329.

——— (2010) "Value Maximization, Stakeholder Theory, and the Corporate Objective Function," *Journal of Applied Corporate Finance*, 22 (1).

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).

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.

Magill, Michael, Martine Quinzii, and Jean-Charles Rochet (2015) "A Theory Of The Stakeholder Corporation," *Econometrica*, 83 (5), 1685–1725.

Malmendier, Ulrike and Geoffrey Tate (2005) "CEO overconfidence and corporate investment," *Journal of Finance*, 60 (6), 2661–2700.

Miles, Samantha (2012) "Stakeholder: Essentially Contested or Just Confused?" *Journal of Business Ethics*, 108.

Morgan, John and Justin Tumlinson (2019) "Corporate Provision of Public Goods," *Management Science*, 65 (10).

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, Stuart Russell et al. (2000) "Algorithms for inverse reinforcement learning.," in *International Conference on Machine Learning (ICML)*, 1.

Peters, Ryan H. and Lucian A. Taylor (2017) "Intangible capital and the investment-q relation," *Journal of Financial Economics*, 123 (2).

Pástor, Luboš, Robert F. Stambaugh, and Lucian A. Taylor (2022) "Dissecting Green Returns," *Journal of Financial Economics*, 146 (2).

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 Schuerhoff, Boris Nikolov, and Volkan Cevher (2022) "Identifiability and generalizability from multiple experts in Inverse Reinforcement Learning," in *Advances in Neural Information Processing Systems*, 36.

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).

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).

Tobin, James (1969) "A General Equilibrium Approach To Monetary Theory," *Journal of Money, Credit and Banking*, 1 (1).

APPENDIX

Appendix A Proofs

Define X and Ω as in (14)-(15), where A and F are square invertible and C is of conformable dimension and defined by (16). Let $B \equiv -X_j^\top X_i$ and $D \equiv X_j^\top X_j$. 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

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

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

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

One obtains

$$A^{-1} = \begin{pmatrix} (A/D)^{-1} & -(A/D)^{-1} B^T D^{-1} \\ -D^{-1}B\,(A/D)^{-1} & D^{-1} + D^{-1}B(A/D)^{-1}B^\top D^{-1} \end{pmatrix}. \tag{A.3}$$

Define the projection matrix under the norm defined by D as

$$\begin{array}{ll} \mathbf{P}_{ij} & = X_i\,(A/D)^{-1}\,X_i^\top, \\ \mathbf{M}_{ij} & = I - X_i\,(A/D)^{-1}\,X_i^\top. \end{array} \tag{A.4}$$

The Schur complement Ω/A then equals

$$\begin{array}{rcl} \Omega/A & = & (f^\top f) - \begin{pmatrix} f^\top X_i & 0 \end{pmatrix} A^{-1} \begin{pmatrix} X_i^\top f \\ 0 \end{pmatrix} \\ & = & (f^\top f) - (f^\top X_i\,(A/D)^{-1}\,X_i^\top f) \\ & = & f^\top (I - X_i\,(A/D)^{-1}\,X_i^\top) f \\ & = & f^\top \mathbf{M}_{ij} f, \end{array} \tag{A.5}$$

with inverse

$$(\Omega/A)^{-1} = \left(f^\top \mathbf{M}_{ij} f\right)^{-1}. \tag{A.6}$$

The Schur complement Ω/F equals

$$\begin{array}{rcl} \Omega/F & = & \begin{pmatrix} 2X_i^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix} - \begin{pmatrix} X_i^\top f \\ 0 \end{pmatrix} (f^\top f)^{-1} \begin{pmatrix} f^\top X_i & 0 \end{pmatrix} \\ & = & \begin{pmatrix} 2X_i^\top X_i - X_i^\top f (f^\top f)^{-1} f^\top X_i & -X_i^\top X_j \\ -X_j^\top X_i & X_j^\top X_j \end{pmatrix}, \end{array} \tag{A.7}$$

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

$$\begin{array}{rcl} (\Omega/F)/D & = & 2X_i^\top X_i - X_i^\top f (f^\top f)^{-1} f^\top X_i - X_i^\top X_j (X_j^\top X_j)^{-1} X_j^\top X_i \\ & = & X_i^\top \left(I - f(f^\top f)^{-1} f^\top\right) X_i + X_i^\top \left(I - X_j (X_j^\top X_j)^{-1} X_j^\top\right) X_i \\ & = & X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i. \end{array} \tag{A.8}$$

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

$$= \begin{pmatrix} ((\Omega/F)/D)^{-1} & -((\Omega/F)/D)^{-1} B^\top D^{-1} \\ -D^{-1}B((\Omega/F)/D)^{-1} & D^{-1} + D^{-1}B((\Omega/F)/D)^{-1}B^\top D^{-1} \end{pmatrix} = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \tag{A.9}$$

with

$$\begin{array}{ll} E_{11} & = \left(X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i\right)^{-1}, \\ E_{12} & = \left(X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i\right)^{-1} X_i^\top X_j (X_j^\top X_j)^{-1}, \\ E_{21} & = (X_j^\top X_j)^{-1} X_j^\top X_i \left(X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i\right)^{-1}, \\ E_{22} & = (X_j^\top X_j)^{-1} + (X_j^\top X_j)^{-1} X_j^\top X_i \left(X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i\right)^{-1} X_i^\top X_j (X_j^\top X_j)^{-1}. \end{array} \tag{A.10}$$

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

$$\begin{pmatrix} (\Omega/F)^{-1} & -(\Omega/F)^{-1} C^\top F^{-1} \\ -(\Omega/A)^{-1} CA^{-1} & (\Omega/A)^{-1} \end{pmatrix} \tag{A.11}$$

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

We get the analytical expressions

$$\begin{pmatrix} v_1 \\ v_2 \\ \theta \end{pmatrix} = \Omega^{-1} X^\top \begin{pmatrix} y_i - y_j \\ y_i \end{pmatrix} = \Omega^{-1} \begin{pmatrix} X_i^\top (2y_i - y_j) \\ -X_j^\top (y_i - y_j) \\ f^\top y_i \end{pmatrix} = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix}, \tag{A.12}$$

with

$$\begin{array}{ll} E_1 & = ((\Omega/F)/D)^{-1} X_i^\top (2y_i - y_j) + ((\Omega/F)/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) \\ & \quad -((\Omega/F)/D)^{-1} X_i^\top f \left(f^\top f\right)^{-1} f^\top y_i, \\ E_2 & = -D^{-1}B((\Omega/F)/D)^{-1} X_i^\top (2y_i - y_j) - D^{-1} X_j^\top (y_i - y_j) \\ & \quad -D^{-1}B((\Omega/F)/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) + D^{-1}B((\Omega/F)/D)^{-1} X_i^\top f \left(f^\top f\right)^{-1} f^\top y_i, \\ E_3 & = -(\Omega/A)^{-1} f^\top X_i (A/D)^{-1} X_i^\top (2y_i - y_j) \\ & \quad -(\Omega/A)^{-1} f^\top X_i (A/D)^{-1} B^\top D^{-1} X_j^\top (y_i - y_j) + (\Omega/A)^{-1} f^\top y_i. \end{array} \tag{A.13}$$

We now obtain our main results:

$$\begin{array}{rcl} v_1 & = & ((\Omega/F)/D)^{-1} X_i^\top \left(I - X_j (X_j^\top X_j)^{-1} X_j^\top\right) y_i \\ & & \quad +((\Omega/F)/D)^{-1} X_i^\top \left(I - f \left(f^\top f\right)^{-1} f^\top\right) y_i \\ & & \quad -((\Omega/F)/D)^{-1} X_i^\top \left(I + X_j (X_j^\top X_j)^{-1} X_j^\top\right) y_j \\ & = & \left(X_i^\top \left(\mathbf{M}_f + \mathbf{M}_j\right) X_i\right)^{-1} X_i^\top \left[\left(\mathbf{M}_f + \mathbf{M}_j\right) y_i - \mathbf{M}_j y_j\right], \end{array} \tag{A.14}$$

and

$$\begin{array}{rcl} v_2 & = & -D^{-1}B((\Omega/F)/D)^{-1} X_i^\top \left(I - X_j (X_j^\top X_j)^{-1} X_j^\top\right) (y_i - y_j) - D^{-1} X_j^\top (y_i - y_j) \\ & & \quad -D^{-1}B((\Omega/F)/D)^{-1} X_i^\top \left(I - f \left(f^\top f\right)^{-1} f^\top\right) y_i \\ & = & (X_j^\top X_j)^{-1} X_j^\top \left[\mathbf{Q}_{ij} \mathbf{M}_f y_i - (I - \mathbf{Q}_{ij} \mathbf{M}_j) (y_i - y_j)\right], \end{array} \tag{A.15}$$

with projection matrix

$$\mathbf{Q}_{ij} = X_i ((\Omega/F)/D)^{-1} X_i^\top, \tag{A.16}$$

and

$$\begin{array}{rcl} \theta & = & (\Omega/A)^{-1} f^\top \left(I - X_i (A/D)^{-1} X_i^\top\right) y_i \\ & & \quad -(\Omega/A)^{-1} f^\top X_i (A/D)^{-1} X_i^\top \left(I + X_j (X_j^\top X_j)^{-1} X_j^\top\right) (y_i - y_j) \\ & = & (\Omega/A)^{-1} f^\top \mathbf{M}_{ij} y_i + (\Omega/A)^{-1} f^\top \mathbf{P}_{ij} \mathbf{M}_j (y_j - y_i). \end{array} \tag{A.17}$$

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 $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, 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 $P(z' \mid z)$. 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 $P_i(z' \mid z)$. These, in turn, lead to different state transition probabilities $P_i(s' \mid s, a)$, which we collect in a three-dimensional array $T_i(s' \mid s, a)$. 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 $t = 1, \ldots, n_{steps}$, 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

$$\hat{\pi}_{i,N}(a|s) = \frac{\#\left(a_{i,t} = a \text{ and } s_{i,t} = s \; ; t \leq N\right)}{\#\left(s_{i,t} = s \; ; t \leq N\right)} \to \pi_i(a|s) \text{ a.s. as } N \to \infty \tag{B.18}$$

and

$$\hat{T}_{i,N}(s'|s, a) = \frac{\#\left(s_{i,t+1} = s \text{ and } s_{i,t} = s \text{ and } a_{i,t} = a \; ; t \leq N\right)}{\#\left(s_{i,t} = s \text{ and } a_{i,t} = a \; ; t \leq N\right)} \to T_i(s'|s, a) \text{ a.s. as } N \to \infty \tag{B.19}$$

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.

Appendix C Data Definitions

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 liabibilites (`dlc`) plus long-term debt (`dltt`) minus cash and cash equivalents (`che`); divided by EBITDA (`ebitda`) if EBITDA is positive. Treated as missing if EBITDA is negative.	$\frac{dlc_t+dltt_t-che_t}{ebitda_t}$
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`) $\succ$ liquidation value (`pstkl`) $\succ$ carrying value (`pstk`) $\succ$ 0). Negative values are treated as missing.	$seq_t + txditc_t - coalesce(pstkrv, pstkl, pstk, 0)$
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 liabibilites (`dlc`) plus long-term debt (`dltt`); divided by total assets (`at`). Negative values are treated as mssing.	$\frac{dlc_t+dltt_t}{at_t}$
Net Book Leverage	Debt in current liabibilites (`dlc`) plus long-term debt (`dltt`) minus cash and cash equivalents (`che`); divided by total assets (`at`)	$\frac{dlc_t+dltt_t-che_t}{at_t}$
Execucomp		
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 (`shrown_excl_opts`) owned by the CEO; divided by common shares outstanding (`csho`)	$\frac{shrown_excl_opts_t}{csho_t}$

(Table C.1 continued)

Variable Name	Definition	Formula
CEO Own. & Options	CEO Ownership plus unexercised exercisable options (`opt_unex_exer_num`) owned by the CEO; divided by common shares outstanding (`csho`)	$CEOOwn._t + \frac{opt_unex_exer_num_t}{csho_t}$
CEO Own. & Opt. 2	CEO Ownership & Options plus unexercised unexercisable options (`opt_unex_unexer_num`) owned by the CEO; divided by common shares outstanding (`csho`)	$CEOOwn.\&Opt._t + \frac{opt_unex_unexer_num_t}{csho_t}$
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 (`shrown_excl_opts`) owned by the top 5 managers; divded by number of shares outstanding (`csho`)	$\frac{\sum_{i=1}^{5} shrown_excl_opts_{i,t}}{csho_t}$
Mana. Own. & Opt.	Managerial Ownership plus the sum of unexercised exercisable options (`opt_unex_exer_num`) owned by the top 5 managers; divided by common shares outstanding (`csho`)	$Mana.Own._t + \frac{\sum_{i=1}^{5} opt_unex_exer_num_{i,t}}{csho_t}$
Mana. Own. & Opt. 2	Managerial Ownership & Options plus the sum of unexercised unexercisable options (`opt_unex_unexer_num`) owned by the top 5 managers; divided by common shares outstanding (`csho`)	$Mana.Own.\&Opt._t + \frac{\sum_{i=1}^{5} opt_unex_unexer_num_{i,t}}{csho_t}$
Peters & Taylor (2017) Data (Downloaded from WRDS)		
Total Capital (K_{tot})	Gross property, plant and equipment (`ppegt`) plus intangible capital (`K_int`)	$ppegt_t + K_{int,t}$
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).	$\frac{ib_t + dp_t}{K_{tot,t-1}} + (1 - \kappa) I_{int,t}$

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.

Appendix D Extensions of the Analysis

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

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 D.1 shows that our results remain largely unchanged. Managers still act as if investment were less costly than prescribed by theory, with coefficients ranging from -0.613 in column (1) to -0.542 in column (6). Adding managerial compensation, ESG scores and higher discount rates improves the model's explanatory performance and the alignment between the managers' recovered value and Total q. The KL divergence decreases from 0.673 and 0.879 in column (1) to 0.669 and 0.848 in column (6) for experts 1 and 2 respectively. The correlation between the recovered value and Total q increases from 0.416 and 0.312 in column (1) to 0.640 and 0.580 in column (6). Comparing column (5) of Table D.1 to column (6) of Table 8, 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. Moreover, the coefficient on governance becomes negative, indicating that the manager would prefer bad governance when making leverage decisions in addition to 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

$$\gamma_i = \frac{1}{1 + \bar{r}_{f,i} + c} \text{ for } i = 1, 2, \tag{D.20}$$

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 D.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 0.259 and 0.455 in column (1) for expert 1 and 2 respectively (compared to 0.268 and 0.503 in column (1) of Table 8). As we add control and ESG-related features, it further drops to 0.249 and 0.447 in column (4) (compared to 0.258 and 0.482 in column (5) of Table 8). This result is largely unchanged when we add compensation features in column (5). The correlation between the recovered value and Total q increases to 0.499 and 0.721 in column (1) for expert 1 and 2 respectively (compared to 0.204 and 0.509 in column (1) of Table 8). As we add control and ESG-related features, it further increases to 0.675 and 0.935 in column (4) for expert 1 and 2 respectively (compared to 0.773 and 0.802 in column (5) of Table 8).

Table D.1: Reward recovery for joint investment and capital structure problem.

The table presents the results of recovering the reward function from corporate investment and capital structure data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, CEO compensation, and ESG scores. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the two action variable 5 total investment rate bins, and 5 net debt to EBITDA bins. The discount factors are $\gamma_1 = 0.971$ and $\gamma_2 = 0.980$ for specifications (1)-(5) and $\gamma_1 = 0.917$ and $\gamma_2 = 0.926$ for specification (6).

Panel A: Feature importance in reward recovery specification $u = f\theta$						
	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	4.782	7.342	7.360	7.322	7.345	7.598
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-2.932	-4.505	-4.534	-4.387	-4.418	-4.115
	(-0.613)	(-0.614)	(-0.616)	(-0.599)	(-0.602)	(-0.542)
Tangibility		0.043	0.043	0.019	0.023	0.031
		(0.006)	(0.006)	(0.003)	(0.003)	(0.004)
Net Book Leverage		0.788	0.786	0.809	0.807	0.818
		(0.107)	(0.107)	(0.110)	(0.110)	(0.108)
CEO Bonus			-0.994		-0.969	-1.084
			(-0.135)		(-0.132)	(-0.143)
CEO Own. & Opt.			-0.632		-0.814	-0.950
			(-0.086)		(-0.111)	(-0.125)
Environmental Score				-0.040	-0.040	-0.039
				(-0.005)	(-0.005)	(-0.005)
Social Score				0.072	0.074	0.076
				(0.010)	(0.010)	(0.010)
Governance Score				-0.002	-0.004	-0.002
				(-0.000)	(-0.001)	(-0.000)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.673	0.672	0.671	0.668	0.667	0.669
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.879	0.849	0.849	0.846	0.845	0.848
N_1	18,518	18,518	18,518	18,518	18,518	18,518
N_2	10,575	10,575	10,575	10,575	10,575	10,575
Panel B: Alignment between recovered value $\hat{v}$ and Total q						
Corr$(\hat{v}_1, q_{1,t}^{tot})$	0.416	0.471	0.477	0.489	0.495	0.640
Corr$(\hat{v}_2, q_{2,t}^{tot})$	0.312	0.395	0.396	0.407	0.408	0.580

Table D.2: Reward recovery with managerial risk aversion or impatience.

The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, CEO compensation, and ESG scores. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the action variable 5 total investment rate bins. The discount factors are $\gamma_1 = 0.917$ and $\gamma_2 = 0.926$.

Panel A: Feature importance in reward recovery specification $u = f\theta$					
	Recovered feature importance θ ($\theta_{\text{normalized}}$)				
	(1)	(2)	(3)	(4)	(5)
Alt. Profitability	23.728	23.886	23.192	22.861	22.484
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-13.906	-13.930	-13.793	-13.643	-13.497
	(-0.586)	(-0.583)	(-0.595)	(-0.597)	(-0.600)
Tangibility		0.020	-0.025	0.106	0.074
		(0.001)	(-0.001)	(0.005)	(0.003)
Net Book Leverage		0.095	-0.175	0.212	0.12
		(0.004)	(-0.008)	(0.009)	(0.005)
CEO Bonus			-46.124		-42.976
			(-1.989)		(-1.911)
CEO Own. & Opt.			-3.732		-6.58
			(-0.161)		(-0.293)
Environmental Score				-0.016	-0.035
				(-0.001)	(-0.002)
Social Score				0.168	0.209
				(0.007)	(0.009)
Governance Score				0.275	0.241
				(0.012)	(0.011)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.259	0.259	0.256	0.249	0.249
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.455	0.455	0.452	0.447	0.446
N_1	18,518	18,518	18,518	18,518	18,518
N_2	10,575	10,575	10,575	10,575	10,575
Panel B: Alignment Between the Recovered Value ($\hat{v}$) and Total q					
$\text{Corr}(\hat{v}_1, q_{1,t}^{tot})$	0.499	0.514	0.562	0.675	0.780
$\text{Corr}(\hat{v}_2, q_{2,t}^{tot})$	0.721	0.873	0.904	0.935	0.929

Appendix E Robustness

In this section, we check the robustness of our results to a different estimation approach for choice and transition probabilities, as well as a finer state and action variable grid.

Estimation of choice and transition probabilities. In our main analysis, we have estimated choice and transition probabilities using a multinomial logit model. This model essentially gives us 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 our method of estimating the empirical choice and transition probabilities, we now repeat our analysis using an alternative count-based estimator. As the name suggests, this estimator simply measures how frequently a given choice (transition) occurred in the data. Since we need to avoid zero-probability events and cannot guarantee ex ante that we will observe very possible (state, action)-pair and (state, action, next state)-triplet in the data, we perform so *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

$$\hat{\pi}_i(a|s) = \frac{\#\left(a_{i,t} = a \text{ and } s_{i,t} = s\right) + \alpha}{\#\left(s_{i,t} = s\right) + \alpha|\mathcal{A}|)}, \tag{E.21}$$

and the transition probabilities based on

$$\hat{T}_i(s'|s,a) = \frac{\#\left(s_{i,t+1} = s \text{ and } s_{i,t} = s \text{ and } a_{i,t} = a\right) + \alpha}{\#\left(s_{i,t} = s \text{ and } a_{i,t} = a\right) + \alpha|\mathcal{S}|}, \tag{E.22}$$

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

Table E.1 shows the results when we use our 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 (for the joint investment and leverage problem) actions. We now repeat our analysis using finer grids, with up to 100 possible states and actions.

Table E.2 shows the results when increase the number of state and action variable bins in the joint investment and leverage problem with managerial compensation, ESG scores and higher discount rates. We still find that managers underestimate the cost of investment. However, some coefficients change sign, the explanatory performance in terms of KL divergence decreases and so does the alignment between the recovered value and Total q. We believe that this is 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 disproportionally while the number of observations used to estimate these parameters stays constant.

Alternative definitions of managerial compensation. In our analysis, we have opted to measure managerial compensation as the compensation received by the CEO. We now repeat our analysis with managerial compensation, defined as the compensation received by the top 5 managers by total current compensation in a given year.

Table E.3 shows that our results are not materially affected when we use managerial compensation instead of CEO compensation.

Table E.1: Robustness of reward recovery for joint investment and leverage problem to the estimation of probabilities.

The table presents the results of recovering the reward function from corporate investment data when we change our estimation approach. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, and ESG scores. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the two action variables 5 total investment rate bins, and 5 net debt to EBITDA bins. The discount factors are $\gamma_1 = 0.971$ and $\gamma_2 = 0.98$ for columns (1) to (5). In column (6), we change the discount factors to $\gamma_1 = 0.917$ and $\gamma_2 = 0.926$.

Panel A: Feature importance in reward recovery specification $u = f\theta$						
	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	2.170	1.313	1.269	1.322	1.280	1.52
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-1.161	-1.080	-1.051	-1.053	-1.021	-1.051
	(-0.535)	(-0.822)	(-0.828)	(-0.797)	(-0.798)	(-0.692)
Tangibility		-0.512	-0.529	-0.518	-0.532	-0.495
		(-0.390)	(-0.417)	(-0.392)	(-0.416)	(-0.326)
Net Book Leverage		0.001	-0.005	0.015	0.009	0.037
		(0.000)	(-0.004)	(0.011)	(0.007)	(0.025)
CEO Bonus			-1.909		-1.894	-1.923
			(-1.504)		(-1.480)	(-1.265)
CEO Own. & Opt.			-0.401		-0.545	-0.652
			(-0.316)		(-0.426)	(-0.429)
Environmental Score				-0.019	-0.019	-0.016
				(-0.015)	(-0.015)	(-0.011)
Social Score				0.038	0.040	0.042
				(0.029)	(0.032)	(0.027)
Governance Score				-0.005	-0.007	-0.006
				(-0.004)	(-0.005)	(-0.004)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.631	0.634	0.630	0.635	0.631	0.629
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.606	0.605	0.604	0.605	0.603	0.602
N_1	18,518	18,518	18,518	18,518	18,518	18,518
N_2	10,575	10,575	10,575	10,575	10,575	10,575
Panel B: Alignment Between the Recovered Value ($\hat{v}$) and Total q						
$\text{Corr}(\hat{v}_1, q_{1,t}^{tot})$	0.606	0.603	0.605	0.607	0.608	0.633
$\text{Corr}(\hat{v}_2, q_{2,t}^{tot})$	0.389	0.531	0.526	0.534	0.528	0.543

Table E.2: Robustness of reward recovery for joint investment and leverage problem to number of state and action variable bins. The table presents the results of recovering the reward function from corporate investment data when we change our binning approach. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, and ESG scores. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. 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. The discount factors are $\gamma_1 = 0.917$ and $\gamma_2 = 0.926$.

Panel A: Feature importance in reward recovery specification $u = f\theta$						
	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	7.598	4.699	5.831	4.216	4.684	3.602
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-4.115	-2.644	-3.141	-2.601	-2.648	-1.357
	(-0.542)	(-0.563)	(-0.539)	(-0.617)	(-0.565)	(-0.377)
Tangibility	0.031	-0.074	0.101	-0.346	0.233	0.855
	(0.004)	(-0.016)	(0.017)	(-0.082)	(0.050)	(0.237)
Net Book Leverage	0.818	0.329	0.500	0.339	0.364	-0.060
	(0.108)	(0.070)	(0.086)	(0.080)	(0.078)	(-0.017)
CEO Bonus	-1.084	0.134	-0.615	-1.971	-0.647	1.853
	(-0.143)	(0.028)	(-0.105)	(-0.467)	(-0.138)	(0.514)
CEO Own. & Opt.	-0.950	-0.987	-0.707	-1.550	-1.730	1.423
	(-0.125)	(-0.210)	(-0.121)	(-0.368)	(-0.369)	(0.395)
Environmental Score	-0.039	-0.027	-0.030	-0.047	-0.005	0.044
	(-0.005)	(-0.006)	(-0.005)	(-0.011)	(-0.001)	(0.012)
Social Score	0.076	0.037	0.042	0.060	0.041	0.002
	(0.010)	(0.008)	(0.007)	(0.014)	(0.009)	(0.000)
Governance Score	-0.002	-0.02	-0.008	0.014	0.003	0.024
	(-0.000)	(-0.004)	(-0.001)	(0.003)	(0.001)	(0.007)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.669	0.702	0.804	0.742	0.758	0.893
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.848	0.908	1.022	0.909	0.868	1.081
N_1	18,518	18,518	18,518	18,518	18,518	18,518
N_2	10,575	10,575	10,575	10,575	10,575	10,575
Panel B: Alignment Between the Recovered Value ($\hat{v}$) and Total q						
Corr($\hat{v}_1, q_{1,t}^{tot}$)	0.640	0.618	0.709	0.535	0.477	0.443
Corr($\hat{v}_2, q_{2,t}^{tot}$)	0.580	0.480	0.637	0.485	0.425	0.373
Panel C: Number of Bins for States and Action Variables						
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

Table E.3: Reward recovery including managerial compensation.

The table presents the results of recovering the reward function from corporate investment data. It displays the recovered feature importance weights, denoted by θ, for six different specifications. The reward features considered include alternative profitability, total investment rate, tangibility, net book leverage, and managerial compensation. Additionally, performance measures such as the Kullback-Leibler divergence between the estimated and true policies are provided. Model specifications include for the two state variables 5 ln(Total Capital) (BoY) bins, 5 alternative profitability (BoY) bins, and for the action variable 5 total investment rate bins. The discount factors are $\gamma_1 = 0.947$ and $\gamma_2 = 0.975$.

Panel A: Feature importance in reward recovery specification $u = f\theta$						
	Recovered feature importance θ ($\theta_{\text{normalized}}$)					
	(1)	(2)	(3)	(4)	(5)	(6)
Alt. Profitability	26.408	24.033	24.38	24.076	24.207	24.658
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
Total Investment Rate	-14.847	-14.303	-14.611	-14.367	-14.344	-14.688
	(-0.562)	(-0.595)	(-0.599)	(-0.597)	(-0.593)	(-0.596)
Tangibility		0.232	0.487	0.119	0.222	0.485
		(0.010)	(0.020)	(0.005)	(0.009)	(0.020)
Net Book Leverage		-2.460	-2.894	-2.51	-2.408	-2.839
		(-0.102)	(-0.119)	(-0.104)	(-0.099)	(-0.115)
Managerial Bonus			-2.095			-2.205
			(-0.086)			(-0.089)
Managerial Ownership				5.174		
				(0.215)		
Managerial Own. & Opt.					1.695	2.525
					(0.070)	(0.102)
$\overline{KLD}(\hat{\pi}_1, \pi_1)$	0.379	0.382	0.370	0.382	0.383	0.370
$\overline{KLD}(\hat{\pi}_2, \pi_2)$	0.562	0.606	0.600	0.606	0.607	0.601
N_1	39,305	39,305	39,305	39,305	39,305	39,305
N_2	29,093	29,093	29,093	29,093	29,093	29,093
Panel B: Alignment Between the Recovered Value ($\hat{v}$) and Total q						
$\text{Corr}(\hat{v}_1, q_{1,t}^{tot})$	0.350	0.205	0.306	0.131	0.156	0.223
$\text{Corr}(\hat{v}_2, q_{2,t}^{tot})$	0.350	0.453	0.765	0.236	0.351	0.570