AlphaManager: A Data-Driven-Robust-Control Approach to Corporate Finance*

Murillo Campello† Lin William Cong‡ Luofeng Zhou§

First draft: August 2022; this draft: August 2025.

Abstract

Corporate decision-making entails complex, high-dimensional, and non-linear stochastic control during which managers learn and adapt via dynamic interactions with the market environment. We propose a data-driven-robust-control (DDRC) framework to complement traditional theory, reduced-form models, and structural estimations in corporate finance research, emphasizing both empirical explanation and prediction of firm outcomes while delivering policy recommendations for a variety of business objectives. Specifically, we develop a predictive environment module using supervised deep learning and integrate a decision-making module based on generative deep reinforcement learning. By incorporating model ambiguity and robust control techniques, our framework not only better explains and predicts corporate outcomes in- and out-of-sample but also prescribes key managerial actions that significantly outperform historical ones. We document rich heterogeneity in model ambiguity, prediction performance, and policy efficacy in the cross section of U.S. public firms and over time. Importantly, DDRC helps delineate where theory and causal analysis should concentrate, integrate fragmented prior knowledge (e.g., via transfer learning), and reveal managerial preferences (through an extension involving inverse reinforcement learning).

Keywords: AI, Ambiguity, Big Data, Corporate Finance, Deep Learning, Generative Modeling, Inverse RL, Offline RL, Stochastic Control, Transfer Learning.

*We are grateful to Wei Jiang and Tom Sargent for their detailed comments and suggestions. We also thank Hui Chen (discussant), Serdar Dinc, Jianqing Fan, John Graham, Zhiguo He, Markus Pelger, Rik Sen (discussant), Lea Stern, conference and seminar participants at AI and Big Data in Finance Research (ABFR) Webinar, 8th Annual Global Quantitative and Macro Investment Conference (Wolfe Research), the 10th Annual Workshop of Business Financing and Banking Research Group(USyd), 5th Big Data Econometrics Theory and Applications Conference (CJoE), China Meeting of Econometric Society (CMES 2021, Shanghai), Chinese University of Hong Kong, City University of Hong Kong, Columbia Business School, Cornell SC Johnson, Cornell Tech, Duke University, 2024 EFMA Annual Meeting, 2023 GSU-RFS FinTech Conference, Harvest Fund 25th Anniversary Ceremony, INFORMS Annual Conference, 4th International FinTech Research Forum (RUC), Johns Hopkins University, 1st Annual MARC Conference at the University of Toronto, National University of Singapore, NYU Stern, Peking University Guanghua, Rutgers Business School, 6th Shanghai Financial Forefront Symposium, Tsinghua University SEM, UBC Sauder, 4th Workshop on Big Data Econometric Theory and Application, and 2023 XJTLU AI and Big Data in Accounting and Finance Conference for helpful comments and discussions. Matthew Wong, Logan Kraver, and Atharwa Pandey provided exceptional research assistance. Send correspondence to Cong at will.cong@cornell.edu.

†University of Florida Warrington College of Business and NBER.

‡Cornell SC Johnson College of Business (Johnson), ABFER, IC3, and NBER.

§New York University Stern School of Business

1 Introduction

Corporate finance research studies firm decision-making and outcomes using theoretical models and archival data. The literature is built on simplified and tractable representations of the corporate environment where rational agents optimize their utilities and interact. Notably, scholars have raised concerns about the low predictability and limited explainability of corporate outcomes, underscoring a systematic gap in our understanding of corporate decision-making (see, e.g., Graham, 2022, Presidential Address of the American Finance Association). Many theoretical models in corporate finance are static, analyze partial-equilibrium, or overlook the interactions with the environment, while empirical studies typically report evidence that is merely "consistent with" *ex-ante* a priori theories without providing a realistic alternative (Spiegel, 2023). In contrast, real-world managerial actions and firm outcomes are now recorded with exceptional granularity and timeliness. Our paper demonstrates that a data-driven perspective — leveraging abundant data, algorithm advances, and powerful computation — can reveal novel empirical patterns and deliver valuable economic insights to guide both academic research and managerial practice.

Implementing a data-driven approach to uncover empirical patterns and advise corporate decision-making presents several challenges. First, managers face a combination of complex decisions in a high-dimensional action space, often contingent on numerous state variables. Their interdependent actions produce highly nonlinear effects that go beyond the reach of conventional econometric models and low-dimensional causal inference. Second, managerial decisions interact dynamically with the economic environment, with market feedback further influencing subsequent choices (e.g., Bond et al., 2010, 2012; Edmans et al., 2012, 2015). Such feedback loops not only impose significant costs but also complicate empirical analyses.[1] Moreover, because optimal corporate decisions are rarely "labeled" in historical data, standard supervised learning techniques are of limited use. Third, unlike physical laws, financial

[1]More generally in social science research, real-time online interaction to generate new data is impractical, either because unfiltered, continuous data collection is expensive (e.g., in high-frequency trading), unethical (e.g., hiring and firing employees), or possibly dangerous (e.g., law enforcement). Even in domains where online interaction is feasible, we might still want to utilize previously collected data instead — for example, if the domain is complex and effective generalization requires large datasets. Therefore, as explained later, we pursue an offline reinforcement learning (RL) approach using historical data similar to Cong et al. (2020).

markets evolve rapidly; consequently, concerns about data distributional shifts become both relevant and pressing.[2]

Given that managerial decision-making is essentially robust control problems characterized by high dimensionality, nonlinearity, dynamic learning, and evolving complexity, existing models often fall short of providing practical policy recommendations for corporate executives. For example, reduced-form empirical models can illuminate the economic mechanisms underlying a particular policy for a selective group of agents; however, their focus is typically local and low-dimensional. As a result, they lack the capacity to explain broader empirical outcomes, generate comprehensive counterfactuals, or yield generalizable recommendations beyond isolated causal effects. Structural estimations excel at managing model complexity to generate counterfactuals by modeling environments in a holistic manner that retains interpretability and a clear economic rationale. However, these models confine themselves to analytically tractable theories and conventional Markov decision processes (MDPs) with pre-specified transition probabilities, largely neglecting dynamic feedback and continuous learning about the environment. Doing so limits their ability to accurately explain observed data, generate reliable out-of-sample predictions, or ultimately provide practical guidance for managerial decision-making.

To advance predictions of firm outcomes under various counterfactual managerial decisions and ultimately to guide corporate decision-making, we integrate deep learning, offline reinforcement learning, and robust control based on ambiguity, thereby framing corporate decision-making as a data-driven-robust-control (DDRC) problem. The framework we propose ("AlphaManager model") serves as a data-driven counterpart to structural estimations by searching a broader modeling space that lends theoretical insights with data-driven patterns. In doing so, it provides a comprehensive depiction of the economic system, explain corporate outcomes, and generate high-dimensional, effective recommendations for enterprise decisions. In our DDRC framework, managerial decision-making is modeled as a robust control problem in which managers maximize their utilities based on contemporaneous states. AlphaManager comprises two modules: (1) the predictive environment module (PEM) and

[2]Data shifts, also known as distributional shifts, occur when the joint distribution of inputs and outcomes differs between training and test samples.

(2) the decision-making module (DMM). In PEM, we leverage supervised deep learning to capture the nonlinear, high-dimensional features inherent in financial big data. In DMM, assuming PEM as given, we apply offline deep reinforcement learning (RL) to reduce search costs and incorporate of flexible managerial objectives along with dynamic feedback.[3] Finally, the robust control techniques such as ambiguity aversion are introduced as constraints, ensuring that AlphaManager remains conservative in the face of high model uncertainty (for instance, due to overfitting, data shifts, or endogeneity concerns), thus mitigating risks.

AlphaManager first constructs predictive environments to test counterfactual via PEM, which generates counterfactual forecasts that both explain and predict variations in corporate outcomes. This process circumvents the need for costly experimentation or explicit causal identification, while still capturing environmental feedback effects. Despite apparent complexity, the DDRC framework enhances traditional reduced-form and structural approaches in several important ways. First, by simulating outcomes from a range of counterfactual managerial decisions, PEM uncovers empirical patterns that can both challenge and refine existing theories — informing new theoretical research. Second, it signals scenarios in which historical data and theory alone are insufficient, underscoring the necessity of reduced-form or structural models for effective learning. To assess model uncertainty, PEM deploys an ensemble of deep neural networks with identical architectures but different initializations; disagreement among these networks indicates that the model is extrapolating beyond its training data, rendering counterfactual predictions less reliable. Finally, our framework integrates insights from *both* reduced-form and structural models through ambiguity-guided transfer learning.[4] In situations where ambiguity is high, empirical causal identification and theoret-

[3]Reinforcement learning (RL) is "learning how to map situations to actions so as to maximize a numerical reward signal." It is one of the three paradigms of modern machine learning together with supervised learning and unsupervised learning. In RL, an agent learns about states of its environment and takes actions that potentially affect states going forward as well as its objective function to maximize. RL is particularly well suited for this task because it learns optimal actions through sequential decision-making and iterative experience accumulation (Sutton and Barto, 1998). Unlike dynamic programming in structural estimation – which relies on known transition probabilities and fixed rewards structures — RL addresses a more general MDP where transition probabilities and rewards are unknown. RL does it either via a model-based approach (e.g., AlphaManager which learns the a model of the environment from data) or a model-free approach (e.g., Q-learning).

[4]Transfer learning has proven valuable across many domains; for instance, the pre-training used in large language models (LLMs) is a form of transfer learning. This approach can be tailored to economics and finance (see Chen et al. (2023)).

ical modeling play a joint, critical role in enriching the base for counterfactual predictions.

AlphaManager then incorporates a decision-making module (DMM) for managerial policy optimization, marking our work the first to combine RL with robust control in corporate finance. At its core, corporate finance problems can be viewed as involving a "system" (the firm and its environment) and a "controller" (the manager).[5] The controller's objective is to optimally manage the system, constrained by (1) the manager's knowledge of the system's state and environment — information that is gleaned via "system sensors" like accounting, auditing, and reporting — and (2) the limited set of parameters that can be directly controlled. In AlphaManager, these two aspects are operationalized by first training a deep supervised learning based PEM from historical data that allows effective counterfactual analysis and data generation. We then train DMM using RL to identify the most effective combination of managerial actions for a given business objective. RL generates the optimal control trajectory based on an exogenously defined reward structure (e.g., market capitalization appreciation) using unlabeled data and by interacting with the environment. In doing so, it offers normative recommendations for managerial decision-making. Finally, by integrating robust control theory, particularly through ambiguity aversion, we guide the training of AlphaManager to extract optimal decisions even under high model uncertainty (Hansen and Sargent, 2023).

We validate our DDRC methodology through systematic numerical experiments within the canonical moral hazard framework of Holmstrom and Tirole (1997). This controlled experimental environment enables precise benchmarking of our reinforcement learning approach against known theoretical optima while isolating the effects of distributional uncertainty on optimal financial contracting. Our validation proceeds through four sequential stages: first establishing baseline convergence properties under perfect information conditions, then introducing controlled distributional perturbations to examine algorithmic ro-

[5]Traditionally control theory is generally designed to solve linear (or linear quadratic) systems with well-defined objectives, law of motions for states, and constraints. However, real-world systems are nonlinear, and techniques that linearize these systems are limited to specific cases. Moreover, accurately modeling such system is challenging, which is why traditional model-driven control approaches — such as structural models in corporate finance — often exclude systems whose underlying dynamics are not fully known. The complexity of these systems increases when addressing hyper-scale issues, such as financial contagion or climate change responses, where unknown nonlinearities and unobserved environmental states render standard dynamic programming and simulation methods inadequate.

bustness under realistic data limitations, implementing our PEM framework to demonstrate robust learning of investor response functions, and finally integrating PEM-learned environments with ambiguity-adjusted reinforcement learning to establish performance bounds under environmental uncertainty.

The numerical results demonstrate substantial practical value creation across diverse experimental scenarios, with our DDRC methodology achieving correlation coefficients exceeding 93.8% with theoretical optima and generating positive welfare improvements in over 85% of evaluated contracts. Notably, performance under distributional shift conditions actually exceeds baseline benchmarks, with correlation reaching 96.13% and improvement ratios averaging 70.8%, validating the economic benefits of explicit uncertainty quantification in financial decision-making frameworks. These findings establish that ambiguity-averse optimization, properly calibrated through ensemble learning techniques, provides effective hedging against model misspecification risks while maintaining robust decision quality across the data limitations commonly encountered in financial contracting applications.

Our AlphaManager application is trained on Compustat, which contains a long panel of fundamental variables of US-listed firms. We study nearly 20 thousand distinct firms with over half-a-million unique firm-quarter observations, ranging from 1976 to 2023. We supplement our dataset with additional stock market data from CRSP, and incorporate macro-level data; e.g., the National Financial Conditions Index (NFCI), from the Chicago Fed. We define two sets of variables: state variables and managerial decision variables. State variables describe the state that a firm faces in a given period of time. Typical examples involve fundamentals of firms (internal states), and macroeconomic or market conditions (external states). Managerial decision variables mediate the future state dynamics and also influence the utility functions of managers.

In PEM, a deep neural network is trained to minimize the mean-squared error (MSE) between real future states and predicted ones, with the input of current states and current managerial decisions. Our PEM achieves high accuracy in predicting and explaining the evolution of firm outcomes (the state variables) with the help of information on managerial decisions, i.e., the managerial planning one period forward. For instance, AlphaManager

produces a 64.7% cross-sectional R^2 for book asset growth and 3.2% for market cap growth, both out of sample. To make it comparable to other empirical research in corporate finance, we also calculate the predictability without managerial planning information. The predictability and explanability remain high for many variables, but more interestingly, for outcomes such as book asset, market capitalization, and enterprise value, the out-of-sample R^2s become negative. This finding is consistent with empirical asset pricing research where the market environment is simply too noisy. Critically for our analysis, however, these results inform us which managerial decision variables matter most for firm dynamics going forward.

Based on PEM, we are able to analyze the heterogeneous outcomes and model ambiguity under counterfactual managerial decisions by state variable, sector, book-to-market decile, and macroeconomic regimes. PEM performs particularly well in trade and transportation, education and healthcare, and manufacturing; predictability is also higher during expansions. Firms with higher firm-level states in the cross section have lower MSE and ambiguity. Notably, model ambiguity is highest for high book-to-market firms.

In DMM, with an exogenously specified utility function of managers, we obtain optimal policies that generate managerial decisions that maximize utilities based on current states. When optimizing the short-term market cap growth, the quarterly outperformance of optimal decisions compared to real managerial decisions is 10.1%. When optimizing the long-term market cap growth, the quarterly outperformance is 8.7%. We find similar patterns when the objective is set as enterprise value growth, with quarterly outperformance of 4.4% and 2.7%, for short and long-term growth, respectively.

We contrast DMM-suggested policy under short-term versus long-term growth in firm value, as well as the term structure of this growth under DMM suggested optimal managerial policy. The contrast between short- and long-term oriented managerial decisions and their implications for firm value are receiving renewed interest in the corporate finance literature (see, e.g., Almeida et al., 2024). Short-term AlphaManager RL outperforms long-term RL in the long run but with higher ambiguity. After ambiguity adjustment, the performance of short-term RL is decreasing in the evaluation time horizon compared to the long-term RL. We further discuss the long-term implication of firm valuation under managerial short-

termism from the point of view of board members. Notably, this approach does not require any knowledge of real decisions, which is the key difference between RL and commonplace supervised machine learning algorithms.

Naturally, the action and information space of a manager is vast. One might wonder about the set of variables needed for a given application and about when should one stop collecting data and constructing new variables. Conveniently, our proposed approach is not hampered by these concerns. Among other things, most variables one uses in applied corporate finance research already convey information about many others. For example, the firm's capital structure already conveys information about its access to credit, risk, asset mix (collateral), and more. The same applies to firm size, or as to whether it is publicly traded, hence the amount of information available to investors. Data "confoundedness" is a plus — not a challenge — to our data-driven approach. The approach we propose is less prone to criticism regarding inconsistencies in variable selection made by researchers in existing studies (see Mitton, 2022) — inconsistencies leading to practices like P-hacking and *ex-post* theory-fitting.

Given the large counterfactual tests we do using the predictive environmental module, AlphaManager necessarily makes extrapolation and interpolation from historical data. Given that historical data are endogenously generated, potential model misspecification poses a particular challenge. The literature traditionally examines parameter stability using Chow-type tests (Andersen et al., 2015). One can also detect static misspecification's by testing moment conditions under a GMM framework (Pan, 2002). For dynamic misspecification tests, Jarrow and Kwok (2015) provide an intuitive exact calibration approach utilizing analytical theoretical models. Our use of deep neural networks goes beyond economic theory and is data-driven. While it explores a much larger modeling–functional space, model misspecification is generally dynamic and hard to detect with conventional econometric tools. Therefore, to guard against and mitigate misspecifications, we adapt the new approach of utilizing ambiguity aversion and entropy-based measures to a data-driven setting. The model ambiguity measure also informs us of scenarios where additional knowledge from theories and causal identifications is important and the ones where predictive models trained on historical observations suffice. It not only guides corporate finance researchers on the dimensions to focus on, but also

allows the integration of fragmented knowledge through ambiguity-guided transfer learning.

Our study contributes to the emerging literature on AI in finance. Machine learning and natural language processing have been widely applied in investment and asset pricing (e.g. Cong et al., 2020, 2021; Gu et al., 2020; Feng et al., 2020). They have also seen applications in studies on corporate finance or financial market risk (e.g., Cong et al., 2018; Li et al., 2020; Bellstam et al., 2020; Hanley and Hoberg, 2019; Campello et al., 2024); but their applications beyond creating or improving the measure of some explanatory variables are rather limited. Exceptions to this line of work include Erel et al. (2018) on predicting board director performance, Cao et al. (2023) on how machines and managers interact in the context of conference calls and earnings announcements, Cao et al. (2021) on how an AI analyst can provide additional insights on stock market forecasts to human analysts.

Existing applications of artificial intelligence methodologies in corporate finance research have predominantly employed conventional machine learning architectures, including standard neural networks, support vector machines, and ensemble methods, primarily within supervised learning frameworks that optimize predictive accuracy for specific outcome variables. While these approaches have generated valuable empirical insights, they typically operate within static optimization paradigms that do not incorporate dynamic sequential decision-making processes or systematic uncertainty quantification. Our contribution advances this literature by implementing deep reinforcement learning within a robust control framework, specifically addressing the dynamic, high-dimensional stochastic control problems inherent in corporate decision-making environments. To our knowledge, this represents the first systematic application of model-based offline reinforcement learning techniques to corporate finance research, offering a methodologically distinct alternative to traditional reduced-form empirical models and structural estimation approaches. The framework's integration of predictive environment modules with ambiguity-adjusted policy optimization addresses fundamental limitations in existing approaches by explicitly modeling environmental uncertainty while generating actionable policy recommendations across diverse business objectives.

Our paper also contributes to the literature on model uncertainty ("ambiguity") and ro-

bust control. Ambiguity represents a source of uncertainty where an economic agent is not confident in which prior models to choose for prediction tasks Hansen and Sargent (2023).[6] Its applications in finance are rare and mostly focus on asset pricing and investment (e.g., Wang, 2005; Mamaysky et al., 2007; Dicks and Fulghieri, 2021), with Dicks and Fulghieri (2019) (financial intermediation), Garlappi et al. (2017) (corporate investment), Izhakian et al. (2022) (capital structure), and Malenko and Tsoy (2020) (security design) as exceptions. Meanwhile, robust control has been put forth in Hansen and Sargent (2001) where a rational agent solves a stochastic control problem under ambiguity aversion. Studies about robust control are mostly theoretical, except for Barnett et al. (2020) which applies robust control to the context of climate change risks. Our work contributes to the literature on model uncertainty as the first empirical/methodological application of the ambiguity concept in corporate finance. We also add to robust control studies by estimating ambiguity with the help of deep learning and then approximate solutions to robust control problems using estimated ambiguity and offline RL.

Finally, our work can be placed in the context of the emerging computer science literature on offline RL (a.k.a, batched RL, e.g., Fujimoto et al., 2019; Kidambi et al., 2020).[7]

A number of papers have illustrated the power of such an approach in enabling data-driven learning of policies for dialogue (Jaques et al., 2019), robotic manipulation behaviors (Ebert et al., 2018; Kalashnikov et al., 2018), and robotic navigation skills (Kahn et al., 2021). While Cong et al. (2020) is the first finance paper that applies offline RL (with online updates) to portfolio management, the authors do not fully optimize the environment module to mimic the real environment. We instead follow Kidambi et al. (2020) to build and optimize an environment module to calculate the transition probability across states without requiring experimenting via a simulator or costly interactions with the actual corporate or market

[6]Hansen and Sargent (2023) defines three sources of uncertainty: risk (in-model innovation), ambiguity (model uncertainty), and misspecification (model class uncertainty). Campello and Kankanhalli (2024) provide a comprehensive review of the research of uncertainty in corporate finance.

[7]Unlike online RL where real-time interactions with the actual environment or environment simulators are possible and counterfactual statements can be evaluated directly, offline RL often works only on historical data without any online interactions with the environment to generate additional data for model training (e.g., Fu et al., 2020). In a pure data-driven scheme, offline RL enables researchers to explore fields that are considered infeasible by classical online RL algorithms, especially those closely related to human behaviors where environment interaction is costly, infeasible, or dangerous (Levine et al., 2020).

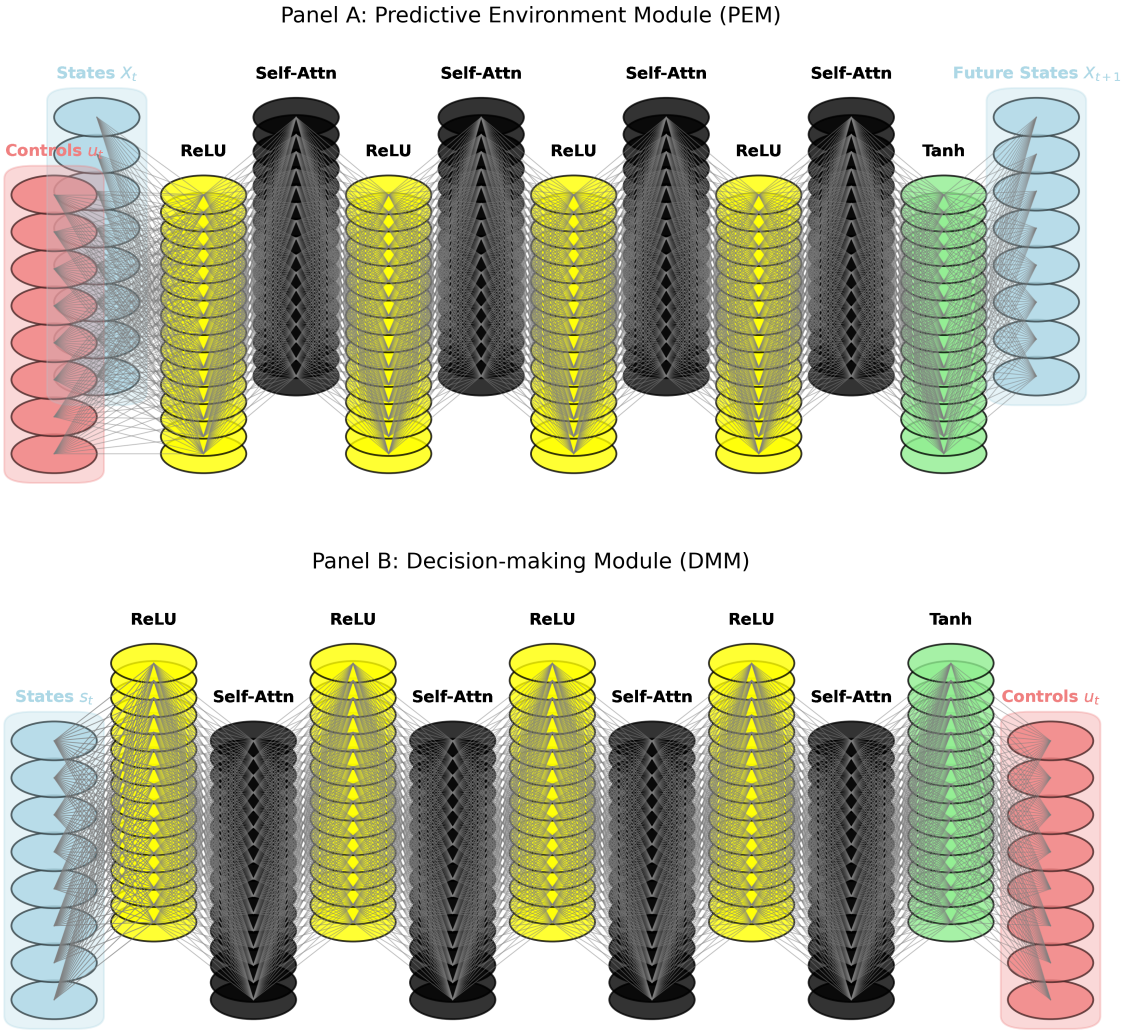
environment. Together with Chen et al. (2023), we are the first studies introducing transfer learning in finance. Transfer learning leverages knowledge from one task or domain and apply it to improve learning in a different but related task or domain. Our DDRC framework is compatible to incorporate inferences from existing empirical causalities and theoretical models to our PEM for better internal validity. Our innovation also lies in using ambiguity to guide the choice of transfer learning and the application in corporate finance. More broadly, our study adds to emerging studies utilizing AI for goal-oriented search, which involves both heuristic search using RL and greed search using panel trees (Cong et al., 2022, 2023).

The rest of the paper is organized as follows. Section 2 describes the DDRC framework. Section 3 validates our DDRC framework in a controlled environment of Holmstrom and Tirole (1997) and tests robustness of DDRC under potential data shifts. Section 4 details data and model training. Section 5 reports the functionality and empirical results from PEM, whereas Section 6 investigates the optimal managerial actions DMM recommends under various given objectives. Section 7 discusses novel research questions that DDRC is particularly suited for. Section 8 concludes.

2 The Data-Driven-Robust-Control (DDRC) Framework

The DDRC approach is comprised of two modules, one utilizes a collection of deep learners to describe the market environment and how the outcomes of interest respond to managerial actions, and the other involves a reinforcement learner offering an optimal policy while dynamically interacting with the market environment and incorporating feedback. The resulting AlphaManager architecture is illustrated in Figure 1. The uses of the rectifier or $ReLU$ (rectified linear unit) and $tanh$ activation functions are standard in deep learning; the self-attention mechanism is widely seen in many deep learning models, such as transformer models; the state variable vector X_t and managerial control vector u_t are introduced shortly.

Figure 1: AlphaManager (AM) Architecture



Note: This figure visualizes the AlphaManager (AM) architecture. Panel A displays the predictive environment module (PEM) with an input layer of a state vector $\mathbf{X_t}$ concatenated with a decision (control) vector $\mathbf{u_t}$. The state vector $\mathbf{X_t}$ (in blue color) includes firm-specific state variables such as firm fundamentals, macroeconomic variables, etc. The decision vector $\mathbf{u_t}$ (in red color) contains leverage, cash holdings, equity financing, etc. Appendix A contains a full list of firm-specific state variables and control variables. The hidden layers are fully-connected layers with rectified linear unit ($ReLU$) activation functions with a self-attention mechanism. After the hidden layers comes a *tanh* activation function which transforms the output spectrum to interval [-1, 1] and the last layer is the output layer with predictions of the system state evolution. Similarly, Panel B shows the structure of the decision-making module (DMM) mapping current state vector $\mathbf{X_t}$ to the optimal control $\mathbf{u_t}$ for a given managerial objective. When implemented, AM uses DMM to generate the optimal managerial actions and then predicts system state evolution using PEM.

2.1 Predictive Environment Module (PEM)

Counterfactual policy analysis in corporate finance necessitates comprehensive modeling of market environment dynamics that capture both immediate impulse responses and equilibrium effects of managerial decisions. Traditional causal identification strategies—including instrumental variable estimation, regression discontinuity designs, and natural experiments—face fundamental limitations in addressing the high-dimensional, sequential decision-making problems characteristic of corporate financial management. These approaches typically identify treatment effects for isolated interventions under restrictive local validity conditions, providing insufficient coverage of the multidimensional policy space relevant for practical decision-making.

Moreover, conventional causal inference methods encounter significant implementation constraints in corporate settings. Credible identification strategies require either naturally occurring quasi-experimental variation or costly randomized controlled trials, both of which prove scarce when analyzing complex strategic decisions. When available, these approaches illuminate causal relationships for specific treatment dimensions while holding other strategic variables constant, which poses a limitation that constrains their applicability to realistic managerial scenarios involving simultaneous optimization across multiple decision domains.

Our framework encompasses comprehensive analysis of both discrete corporate events, such as merger transactions, restructuring decisions, or bankruptcy proceedings, and continuous strategic adjustments involving combinations of capital allocation, financing choices, and operational policies. The latter category represents the predominant form of corporate decision-making, involving high-dimensional optimization problems that extend beyond the analytical scope of traditional causal identification frameworks.

The Predictive Environment Module (PEM) addresses these limitations through a model-based reinforcement learning approach that shares conceptual similarities with structural estimation methodologies while incorporating several distinctive features. Like structural models, our framework develops comprehensive representations of the economic environment that enable broad counterfactual analysis across diverse policy scenarios. However, rather than imposing theoretical restrictions through parametric assumptions and closed-form ana-

lytical solutions, the PEM employs data-driven learning algorithms to estimate environment dynamics directly from observed firm behavior and market responses. This approach trades the theoretical interpretability characteristic of structural models for enhanced flexibility in capturing complex, potentially nonlinear relationships that may not conform to standard theoretical specifications.

The PEM employs deep neural network architectures to model the dynamic relationships between managerial decision variables and subsequent evolution of both firm-specific fundamentals and broader market state variables. This modeling approach addresses the inherent complexity of corporate financial environments, where high-dimensional action spaces interact through potentially nonlinear transmission mechanisms that extend beyond the analytical tractability of conventional econometric specifications.

Neural network methodologies provide several critical advantages for environment modeling in corporate finance applications. First, these architectures demonstrate exceptional capacity for capturing nonlinear dependencies within high-dimensional variable spaces, a characteristic essential for representing the complex interdependencies among corporate decisions, firm fundamentals, and market responses. Second, the universal approximation theorem (Hornik et al., 1989) establishes that feedforward neural networks with sufficient width can approximate any continuous function to arbitrary precision, providing theoretical justification for their application to unknown functional relationships governing corporate environment dynamics.

The empirical motivation for this approach stems from the multidimensional nature of corporate decision-making, where managers simultaneously optimize across capital structure choices, investment allocation, dividend policies, operational strategies, and risk management decisions. These choices generate feedback effects that propagate through firm fundamentals, influence market perceptions, and ultimately affect future decision constraints through complex, potentially nonlinear channels. Traditional reduced-form approaches typically examine isolated relationships between specific decision variables and outcomes, while structural models impose parametric restrictions that may inadequately capture the full complexity of these interdependencies.

The PEM therefore serves as a comprehensive representation of the corporate financial ecosystem, learning environment dynamics directly from observed firm behavior while maintaining sufficient flexibility to accommodate the nonlinear interactions and feedback mechanisms that characterize modern corporate finance environments.

Our approach recognizes that corporate decisions involve the manager taking a sequence of high-dimensional actions to optimize some given economic objective, which makes them stochastic control problems. Suppose the dynamics of the state variables in the system follow:

$$\Delta X_{t+1} = f(X_t, u_t) + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim N(0, \Sigma), \tag{1}$$

where X_t denotes a firm's internal state variables (e.g., accounting fundamental) and external state variables (e.g., inflation and unemployment) which are out of the manager's control at time t, u_t stands for the vector of managerial actions that are made by managers at time t, f describes how the expected value of state change, ΔX_{t+1}, corresponds to current state and managerial actions, and ε_{t+1} is the normally distributed risk term which represents the in-model uncertainty (i.e., the uncertainty that has been captured by the model). We can then use a feedforward neural network to approximate the function f. Instead of using fully-connected layer design, we adopt a combination of a fully-connected layer and a self-attention layer, following the popular design of transformer encoder.

PEM is expected to generate reliable out-of-sample predictions for counterfactual evaluations — tasks that require both extrapolation and interpolation. This process is inherently problematic because historical data are generated by endogenous managerial decisions, leaving gaps in some regions of the action-state space. Similar to reduced-form models, two primary challenges arise: overfitting and data shift. In structural estimations, these concerns are mitigated by assuming that the theoretical model perfectly represents the underlying system. Analogously, if the state variables we include fully capture the system's evolution, the risks associated with extrapolation and interpolation would be minimized.

Various statistical procedures can mitigate these concerns without relying on such strong assumptions. For instance, to address overfitting, one common approach is to include L1-norm or L2-norm penalties in the loss function during the training of a supervised deep

neural network (e.g., Kaniel et al., 2023). Furthermore, by leveraging the virtue of complexity (e.g., Kelly et al., 2024), our over-parameterized PEM in the empirical implementation is designed to reduce overfitting and enhance out-of-sample performance. However, even when overfitting is controlled, PEM inherently introduces model ambiguity (Hansen and Sargent, 2023) because the data-generating process may shift, leading to uncertainty about whether a properly trained model will generalize to new datasets. In other words, neural network models with different parameters might perform similarly on the training data, yet yield divergent predictions due to a combination of residual overfitting and data shifts. As such, selecting the appropriate model *ex ante* adds an additional layer of uncertainty. In Section 2.3, we demonstrate how to empirically estimate ambiguity for any point in the variable space and address this issue by incorporating ambiguity aversion into the objective function.[8]

2.2 Decision-making Module and Offline Reinforcement Learning

Managers take a sequence of actions to solve the following stochastic control problem:

$$\max_{\{u_{t_0},\ldots,u_{t_0+T}\}} \mathbb{E}_{t_0} \sum_{t=t_0}^{t_0+T} r(X_t, u_t) \qquad s.t.\ \Delta X_{t+1} = f(X_t, u_t) + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim N(0, \Sigma), \tag{2}$$

where $r(X_t, u_t)$ is the instantaneous reward function given the state X_t and the managerial decision u_t, and T is the length of the optimizing time period associated with the managerial objective. An example of instantaneous reward function is the enterprise value growth for the next period given current state and managerial decisions.

The conventional approach to stochastic control problems involves deriving HJB equations to obtain closed-form solutions. When analytical solutions are unattainable, numerical algorithms can yield reliable, low-variance empirical approximations — provided the problem is low-dimensional. In our setting, the dynamics lack an analytical characterization, and the environment evolves in a high-dimensional, potentially nonlinear manner, which motivates our use of reinforcement learning.

[8]This approach essentially adds a penalty based on relative entropy (also known as Kullback-Leibler divergence in computer science literature).

In order to functionally represent managerial decisions of interest, we follow the canonical policy gradient algorithm in RL (Sutton and Barto, 1998) and assume that decisions are made contingent on the state variable X. Specifically,

$$u_t = g(X_t), \tag{3}$$

where the functional representation g is called decision-making module (DMM), serving as a decision-making device contingent on the state vector. By substituting (3) into (2), we can reformulate the stochastic control problem as:

$$\begin{aligned} \max_{g(\cdot)} \quad & \mathbb{E}_{t_0} \sum_{t=t_0}^{t_0+T} r(X_t, u_t) \\ s.t. \quad & \Delta X_{t+1} = f(X_t, u_t) + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, \Sigma) \\ & u_t = g(X_t). \end{aligned}$$

AlphaManager continuously learns about the environment through interactions and refines her understanding using the predictive environment module (PEM), which models the function $f(X_t, u_t)$. At the same time, she maximizes her expected cumulative payoff by optimizing over a broad space of dynamic policies, represented by $g(X_t)$, using the decision-making module (DMM).

2.3 Robust Control and Ambiguity Aversion

Neural networks usually have more free parameters than available training samples, leading to under-identified parameter estimates. Even when fewer parameters are used, endogeneity issues in the training data can undermine a model's out-of-sample predictive power because data-driven training does not distinguish causal relationships or separate supply-side shocks from demand-side shocks. Consequently, variations in parameter initialization or model architecture may produce multiple models that perform similarly in-sample yet yield different out-of-sample predictions. Traditional model selection focuses on validation set performance, but when the first moment of predictions is comparable across models, ad-

ditional metrics to guide model choices become necessary. Data shifts further compound the problem; the dispersion of predictions on the test set not only signals potential overfitting but also reflects model uncertainty arising from these shifts. This "across-model" uncertainty is referred to as *ambiguity* in robust control literature (Hansen and Sargent, 2023). The literature typically handles ambiguity problems in two steps. First, researchers estimate ambiguity through some metrics (e.g., relative entropy) to summarize the discrepancy among multiple models defined as statistical distributions indexed by different parameters. If this discrepancy for a given test sample is large, the ambiguity for this specific sample is high. Second, when modeling a decision-maker's problem, a punishment for ambiguity is added to the original objective function, making the agent ambiguity-averse.

To estimate ambiguity, we apply ensemble learning by training a set of neural networks with similar structures for PEM. These networks are trained under the same dataset and loss function but with different initial coefficients or slight variations in the model specification.[9] Under the supervised learning framework, these networks achieve comparable functionality in-sample, as prediction errors are consistently penalized during training. However, out-of-sample performance may diverge across models, particularly for unprecedented scenarios or unusual events not adequately represented in the training sample. The level of disagreement among these models therefore serves as a natural proxy for whether a given input lies within the empirical support of the training data. When ensemble disagreement is high, this indicates either that current firm states represent unusual market conditions or that proposed managerial decisions constitute uncommon strategic choices relative to historical patterns. Formally, we trained a set of environment modules $i = 1, 2, ..., I$ mapping (X_t, u_t) to $\hat{X}^i_{t+1}$. We define *boosting error*, motivated by the relative entropy which is a theoretical metric of ambiguity which captures ensemble disagreement, for (X_t, u_t) is calculated as:

$$\text{BoostingError}(X_t, u_t) = \frac{1}{D} \sum_{d=1}^{D} \left(\max_{i=1,2,...,I} \hat{X}^i_{t+1,d} - \min_{i=1,2,...,I} \hat{X}^i_{t+1,d} \right)^2, \tag{4}$$

[9]The choice of ambiguity metrics could vary, as shown in Hansen and Sargent (2023). We choose to use boosting error in our context given its simplicity of calculation and its intuitive interpretation as maximum dispersion of predictions among models with similar structure and parameter initialization under the same training process.

where D is the number of dimensions for state vector X and u_t is an arbitrary managerial control attempted. Note that there are alternative specifications of boosting errors one can use, which still capture the concept of ambiguity.

In addition to direct ambiguity penalization, we implement pessimistic optimization protocols commonly utilized in economics and computer science applications.[10] Under model uncertainty, the decision-maker maximizes the minimum expected reward generated across the ensemble of models. This pessimistic approach provides an additional mechanism, complementing explicit ambiguity aversion, to control for model uncertainty when generating policy recommendations.

To train DMM and solve the RL problem, we use fully connected neural networks with self-attention mechanisms to parameterize the decision-making device g, the mapping from state to control. We also incorporate the methodological approach of Kidambi et al. (2020) and Yu et al. (2020) for integrating pessimistic reward evaluation and ambiguity penalization within the training process. Specifically, during DMM training, for any given initial state X_{t_0}, the empirical objective function combines pessimistic reward maximization with explicit ambiguity control:

$$J(X_{t_0}) := \sum_{t=t_0}^{t_0+T} \min_{i=1,2,\ldots,I} r^i\left(X_t, g(X_t)\right) - \delta \cdot \max\left\{0, \max_{t=t_0,\ldots,t_0+T} \text{BoostingError}\left(X_t, g(X_t)\right) - \theta(X_{t_0})\right\} \tag{5}$$

In Eq (5), $r^i\left(X_t, g(X_t)\right)$ denotes the instantaneous reward function by environment module i, given state PEM-predicted time-t state, X_t, and the decision-making device g, and we take the lowest instantaneous reward across I different environment modules. The ambiguity punishment parameter, δ, is chosen to be very large so that we guarantee that the boosting error is always small enough. Empirically we pick $\delta = 10^7$. BoostingError$(X_t, g(X_t))$ is the boosting error at time t, given PEM-predicted time-t state, X_t, and the decision-making device g, calculated as Eq (4). The benchmark error, $\theta(X_{t_0})$, is calculated as a constant α plus β times the boosting error of X_{t_0} and the lagged real managerial decisions $\bar{u}_{t_0-1}$ which

[10] See Barnett et al. (2020) in economics and Jin et al. (2021) in computer science for modeling pessimistic agents solving max-min problems to mitigate ambiguity.

is known at time t:

$$\theta(X_{t_0}) = \alpha + \beta \cdot \text{BoostingError}(X_{t_0}, \bar{u}_{t_0-1}), \tag{6}$$

where $\beta > 0$ and $\bar{u}_{t_0-1}$ is the actual managerial decision in the previous period just a simple benchmark for how ambiguous the states correspond to. What we do here is essentially prioritizing the reduction in ambiguity over the baseline rewards, if the ambiguity exceeds some threshold. When considering different optimizing period T, given the monotonicity of $\max_{t=t_0,\ldots,t_0+T} \text{BoostingError}\left(X_t, g(X_t)\right)$, we choose higher values of α and β when optimizing under a higher T. The empirical choices of α and β under different optimizing period T are detailed in Appendix C.

3 Methodology Validation in Controlled Environment: Holmstrom and Tirole (1997)

3.1 Theoretical Foundation and Experimental Design

We validate our DDRC methodology within the canonical moral hazard framework of Holmstrom and Tirole (1997), which provides an analytically tractable environment for isolating the effects of distributional uncertainty on optimal financial contracting. This framework serves as an ideal controlled testing environment for several methodologically compelling reasons. First, the model's closed-form solutions enable precise benchmarking of our RL approximation performance against known theoretical optima. Second, the inherent information asymmetries between contracting parties naturally generate distributional misalignment scenarios that our DMM is designed to address.

Our experimental design proceeds through four sequential validation stages to systematically examine DDRC framework performance under varying degrees of uncertainty. Initially, we establish baseline convergence properties under perfect information conditions, where the PEM observes investor response functions that precisely match the theoretical data-generating process. Subsequently, we introduce controlled perturbations to investor behavior to examine how traditional RL methods deteriorate under covariate and distributional

shifts, representing realistic deviations from theoretical predictions due to unobserved heterogeneity, market frictions, or temporal instability. In the third stage, we demonstrate how our PEM framework learns robust representations of investor response patterns under such uncertainty. Finally, we integrate PEM-learned environments with our DMM to establish performance bounds when contract designers operate under realistic information constraints.

The economic rationale underlying this methodological approach recognizes that real-world financial environments systematically deviate from theoretical assumptions. When observed investor behavior diverges from model predictions due to either institutional constraints, regulatory changes, or evolving market participant characteristics, managers employing traditional optimization frameworks may systematically misestimate optimal incentive provision. Our integrated DDRC framework addresses this fundamental challenge by explicitly modeling and adapting to such environmental distributional uncertainties.

3.2 Holmstrom and Tirole (1997) Moral Hazard Model

The Holmstrom and Tirole (1997) model examines optimal financial contracting between risk-neutral entrepreneurs (managers) and external investors under unobservable effort provision. This canonical framework captures the fundamental tension between efficient risk-sharing and incentive provision that characterizes principal-agent relationships in corporate finance settings.

The economic environment consists of entrepreneurs possessing initial wealth A, and access to profitable investment opportunities requiring unit capital investment without loss of generality. Investment projects generate deterministic returns R upon success and zero upon failure, where success probabilities depend critically on unobservable entrepreneurial effort choices. Low effort provision yields success probability p_L, while high effort increases this probability to $p_H = p_L + \Delta p$, where Δp represents the marginal productivity of entrepreneurial effort. The moral hazard problem arises because entrepreneurs forgo private benefits B when exerting high effort, creating the classic trade-off between effort incentives and efficient risk allocation.

Entrepreneurs, recognizing effort unobservability, must design contracts that simultane-

ously ensure external investor participation and entrepreneur profitability. The contracting space consists of state-contingent payment schedules: entrepreneurs receive R_b upon project success, while investors claim the residual return $R_l = R - R_b$. Critically, the investor's participation decision q depends on anticipated entrepreneurial behavior, which itself responds endogenously to contract terms, creating a complex equilibrium determination process that our PEM framework seeks to learn and predict.

In our DMM implementation, the environment represents investor response functions mapping contract offers to participation decisions, conditional on current market states. The PEM component learns these response patterns from observed investor behavior, enabling robust contract optimization even when investor behavior exhibits systematic deviations from theoretical predictions. The entrepreneur's optimization problem under this framework becomes:

$$\begin{aligned} \max_{R_b,\, e\in\{L,H\}} \quad & q \cdot \max\{A, (p_e \cdot R_b + \mathbf{1}_{e=L} \cdot B)\} + (1-q) \cdot A \\ \text{s.t.} \quad & R_b + R_l = R \\ & q = q(\Delta p, A, R_b; B, p_L) \\ & R_b \geq 0,\ R_l \geq 0,\ A < 1 < R \end{aligned}$$

3.3 Theoretical Solution and Economic Regime Characterization

The theoretical solution reveals three distinct economic regimes determined by the relative magnitudes of required external financing, project profitability, and moral hazard severity. These regimes provide benchmark cases for evaluating our RL methodology's performance under varying degrees of contracting complexity.

Regime 1: first-best implementation. When external financing requirements are minimal relative to expected project value under high effort provision, specifically when $1 - A < p_L \cdot R$, investors participate unconditionally ($q = 1$), effectively eliminating moral hazard constraints. This regime represents the first-best scenario where financial market

frictions do not distort real investment allocation decisions.

Regime 2: second-Best contracting under moral hazard. When financing requirements necessitate meaningful risk-sharing between entrepreneurs and investors, optimal contracts must balance incentive provision against participation constraints. The critical indifference condition $\Delta p \cdot R_b = B$ ensures that entrepreneurial effort incentives align with investor participation expectations, though at the cost of constrained efficiency due to moral hazard considerations.

Suppose $R_b > B/\Delta p$ and $p_H R_b > A$, the entrepreneur chooses to exert effort if the external investor chooses to take the contract and invest; otherwise, the entrepreneur chooses not to exert effort. So for the external investor thinking about whether to take the contract or not, if $R_b > B/\Delta p$ and $p_H R_b > A$ and $p_H R_l > 1 - A$, the investor will invest and the entrepreneur will exert effort; if $R_b < B/\Delta p$ and $p_L R_b > A$ and $p_L R_l > 1 - A$, the external investor will invest and the entrepreneur will not exert effort.

Regime 3: market breakdown. When external financing requirements exceed project value even under optimal effort provision, specifically when $1-A > p_H \cdot R_l$, no mutually beneficial contracts exist ($q = 0$), resulting in socially inefficient investment project abandonment despite positive net present value.

The formal characterization proceeds as follows: For entrepreneurs facing investor participation uncertainty, optimal contracting strategies depend on the intersection of effort incentive constraints and investor participation thresholds. When $R-(1-A)/p_H > \max\{B/\Delta p, A/p_H\}$ and contracts satisfy $R - (1 - A)/p_H > R_b > \max\{B/\Delta p, A/p_H\}$, entrepreneurs implement high-effort strategies earning $p_H R_b \leq p_H \cdot R - (1 - A)$ where $R_b^* = R - (1 - A)/p_H$. Alternatively, when $\min\{B/\Delta p, R - (1 - A)/p_L\} > A/p_L$ and contracts satisfy $\min\{B/\Delta p, R - (1 - A)/p_L\} > R_b > A/p_L$, entrepreneurs optimally choose low-effort provision, earning $p_L R_b + B \leq \min\{p_H \cdot B/\Delta p, p_L \cdot R - (1 - A) + B\}$ where $R_b^* = \min\{B/\Delta p, R - (1 - A)/p_L\}$. In all other parameter configurations, entrepreneurial effort becomes not enforceable.

This theoretical foundation establishes the controlled environment necessary for systematically evaluating our PEM-integrated RL methodology's capacity to learn optimal contracting strategies under various degrees of environmental uncertainty, as we demonstrate in

the subsequent empirical analysis.

3.4 Formal Specification of State and Control Spaces

Following the theoretical characterization established in the preceding analysis, we formally define the state and control spaces that constitute the foundation for our reinforcement learning implementation.

State Space Specification. The state vector $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^3$ is defined as $\mathbf{s} = (A, \Delta p, R)$, where each component captures fundamental economic parameters governing the contracting environment. Initial entrepreneurial wealth A determines the magnitude of required external financing $(1 - A)$, effort productivity differential Δp measures the economic significance of moral hazard concerns, and total project returns R establish the scale of potential value creation. We suppose private benefits and baseline success probability to $B = p_L = 0.1$ throughout the empirical analysis, ensuring parameter parsimony while preserving the essential economic trade-offs inherent in the theoretical framework.

Control Space Specification. The control vector $\mathbf{c} \in \mathcal{C} \subset \{L, H\} \times \mathbb{R}_+$ is defined as $\mathbf{c} = (e, R_b)$, encompassing both discrete effort provision decisions and continuous contract design choices. The effort choice $e \in \{L, H\}$ represents the entrepreneur's private action regarding productivity enhancement, while the contract payment $R_b \geq 0$ denotes the entrepreneur's state-contingent compensation upon project success. This control specification captures the dual optimization problem faced by entrepreneurs: simultaneous determination of optimal effort provision and contract negotiation strategies.

Environment Function Specification. The unknown environment function $q : \mathcal{S} \times \mathcal{C} \to \{0, 1\}$ maps state-control combinations to binary investor participation decisions, where $q(\mathbf{s}, \mathbf{c}) = 1$ indicates external investor willingness to provide required financing $(1 - A)$. This function represents the central object of learning in our framework, as its accurate estimation determines the feasibility of alternative contracting strategies. The PEM component specifically targets this function, learning robust representations of investor response patterns from observed market behavior while accounting for potential distributional uncertainty in the underlying data-generating process.

3.5 RL under Perfect Information

We initiate our empirical validation by examining reinforcement learning performance under idealized conditions where the environment function $q(\cdot)$ is perfectly observable. This benchmark establishes convergence properties and approximation accuracy before introducing realistic complications arising from environmental uncertainty.

Methodological framework. Our implementation employs a discrete approximation approach adapted to the relatively low-dimensional state-control space characteristic of the Holmstrom and Tirole (1997) framework. Rather than deploying complex neural network architectures that may introduce unnecessary computational complexity, we utilize exhaustive grid search techniques that leverage the model's analytical tractability while maintaining algorithmic transparency. This approach enables precise performance benchmarking against known theoretical solutions while providing clear intuition for the underlying optimization dynamics.

Experimental design. We generate a comprehensive training dataset consisting of 100,000 uniformly sampled state vectors $\mathbf{s}_i = (A_i, \Delta p_i, R_i)$. For each state realization, we construct dual control grids corresponding to low-effort ($e = L$) and high-effort ($e = H$) strategies, with contract payment schedules R_b spanning the economically feasible range determined by participation and limited liability constraints.

Optimization procedure. The reinforcement learning algorithm evaluates the entrepreneur's objective function across all grid points within each effort regime, utilizing the perfect environment function to determine investor participation probabilities for each state-control combination. The optimization procedure systematically compares expected payoffs across both effort strategies and selects the control vector $\mathbf{c}^*$ that maximizes entrepreneurial utility, accounting for both the direct effects of contract terms and the indirect effects operating through investor participation incentives.

Performance validation. Empirical results demonstrate remarkably precise convergence between reinforcement learning solutions and analytical optima derived from the theoretical framework. The algorithm successfully identifies optimal contracts across all three economic regimes, with numerical approximation errors consistently below 0.1% of theoretical values. This performance validation confirms that our RL methodology effectively captures the complex strategic interactions underlying optimal financial contracting when environmental uncertainty is absent.

Implications for data shifts. While these results establish the fundamental validity of our algorithmic approach, they simultaneously highlight the critical importance of accurate environment learning in realistic applications. The perfect information benchmark reveals that small perturbations in investor response functions can substantially alter optimal contracting strategies, motivating the development of robust learning mechanisms that we address through the DDRC framework in subsequent analysis. This finding underscores the broader methodological contribution of our approach: developing reinforcement learning algorithms capable of maintaining performance stability under realistic conditions of environmental distributional uncertainty.

3.6 Environment Learning and Uncertainty Quantification

In realistic applications, the investor response function $q(\cdot)$ remains unobservable, necessitating empirical estimation from available contracting data. We implement the PEM framework to address this fundamental identification challenge while simultaneously quantifying the distributional uncertainty inherent in environment learning under potentially incomplete or non-representative training data.

PEM architecture and ensemble learning. Our implementation employs Light Gradient Boosting Machine (LGBM) algorithms within an ensemble learning framework designed to capture both predictive accuracy and model uncertainty. Specifically, we train 11 distinct LGBM models with varying random initialization seeds, creating a heterogeneous ensemble

that enables robust uncertainty quantification through prediction variance. This ensemble approach provides principled estimates of epistemic uncertainty arising from limited training data while maintaining computational efficiency relative to more complex Bayesian alternatives.

Ambiguity quantification methodology. We measure ambiguity specified in Eq (4), with constraint parameters set to $\alpha = \beta = 0$ in Eq (6) for analytical parsimony. The resulting ambiguity estimates capture prediction disagreement across ensemble members, providing empirically grounded measures of environmental uncertainty that inform subsequent reinforcement learning optimization. To incorporate distributional robustness into the decision-making framework, we set the penalty parameter $\delta = 5$ in Eq (5), reflecting moderate ambiguity aversion.

Experimental validation across distributional regimes. We systematically evaluate PEM performance under three distinct data-generating scenarios that represent common challenges in empirical financial economics: representative sampling conditions, distributional shift, and covariate shift. These scenarios enable controlled assessment of our methodology's robustness to realistic departures from idealized learning environments.

Baseline performance under representative sampling. Under conditions where training data represents an unbiased random sample from the population distribution, our PEM framework demonstrates uniform ambiguity patterns across the state-control space. Training data encompasses both optimal and suboptimal contracting decisions, ensuring comprehensive coverage of economically relevant scenarios. The resulting ambiguity distribution exhibits systematic uniformity across state variables, indicating successful generalization capacity and confirming that the ensemble learning approach effectively captures the underlying investor response function without systematic bias or overfitting concerns.

Distributional shift analysis. We examine performance degradation under distributional shift by restricting training observations to cases where initial wealth A falls below the

80th percentile of the empirical distribution, while reserving higher-wealth observations for out-of-sample testing. This experimental design mimics realistic scenarios where available contracting data systematically under-represents certain market segments or regulatory environments. Empirical results, presented in Figure 2 (left panel), reveal systematic elevation in ambiguity levels for test observations relative to training data. The magnitude of uncertainty increases monotonically with the degree of extrapolation beyond the training distribution, confirming that our ambiguity quantification methodology successfully identifies regions of heightened prediction uncertainty. This pattern validates the economic intuition that investor behavior predictions become increasingly unreliable when extrapolating to market conditions not adequately represented in historical data.

Covariate shift analysis. We assess robustness to covariate shift by constructing training samples that systematically under-represent near-optimal contracting decisions. Specifically, we measure the Euclidean distance between empirical contracts and theoretical optima, restricting training data to observations below the 80th percentile of this optimality distance measure. This design captures scenarios where available data predominantly reflects suboptimal decision-making, potentially due to institutional constraints, behavioral biases, or transition periods following regulatory changes.

Results shown in Figure 2 (right panel) indicate minimal differences in ambiguity levels between training and test distributions under covariate shift conditions. This finding suggests that our PEM framework exhibits relative robustness to variation in the optimality characteristics of training data, likely reflecting the ensemble learning approach's capacity to extract generalizable patterns from diverse contracting strategies rather than overfitting to specific decision quality characteristics.

3.7 Empirical Performance Analysis: DDRC Implementation Results

We evaluate the practical effectiveness of our DDRC methodology through comprehensive performance analysis across the three experimental scenarios. Our assessment focuses

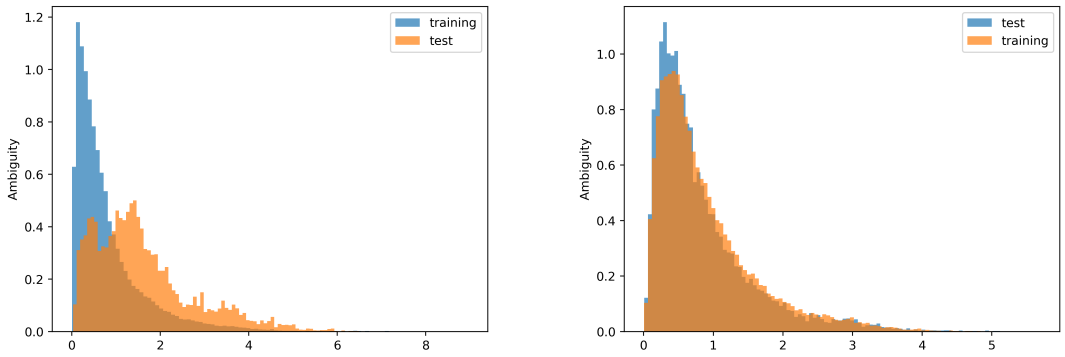


Figure 2: Distributional Uncertainty Analysis: Ambiguity Patterns under Distributional Shift (Left Panel) and Covariate Shift (Right Panel)

on two complementary metrics: correlation between DDRC-suggested contracts and theoretical optima, and economic improvement ratios measuring value enhancement relative to empirically observed baseline decisions.

Performance metrics and evaluation framework. We quantify algorithmic accuracy through Pearson correlation coefficients between DDRC-recommended contract payments R_b and theoretically optimal values R_b^*, providing direct measures of approximation quality across diverse economic environments. Additionally, we calculate improvement ratios measuring the percentage enhancement in entrepreneurial objective function values achieved by DDRC recommendations relative to empirically observed contracting decisions, enabling assessment of practical economic significance beyond pure approximation accuracy.

Baseline performance under representative conditions. Under representative sampling conditions, our DDRC methodology achieves a correlation of 93.8% between recommended and theoretically optimal contract payments, demonstrating robust approximation accuracy across diverse economic scenarios. Subsample analysis reveals consistent performance across varying levels of initial wealth and decision quality, indicating stable algorithmic behavior independent of specific economic conditions within the training distribution.

Economic improvement analysis reveals substantial practical value creation: 85.1% of observations exhibit positive improvement relative to baseline empirical decisions, with av-

erage and median improvement ratios of 38.9% and 65.0%, respectively. The divergence between mean and median improvements reflects the mechanical amplification of approximation errors when baseline decisions approach theoretical optima—a phenomenon with limited economic significance given the small absolute welfare effects in such cases.

Distributional robustness under distributional shift. Performance under distributional shift conditions exceeds baseline benchmarks, with DDRC achieving 96.13% correlation with theoretical optima and generating positive improvements in 96.8% of test observations. Average and median improvement ratios reach 70.8% and 84.0%, respectively, substantially exceeding representative sampling performance levels.

This counterintuitive improvement reflects the beneficial effects of ambiguity aversion in focusing algorithmic attention on regions of heightened uncertainty. By incorporating ambiguity aversion through the penalty parameter δ, our methodology exhibits enhanced conservatism that aligns with near-optimal decision-making in the specific experimental environment, suggesting potential benefits of robust optimization approaches even beyond their primary uncertainty management objectives.

Covariate shift resilience. Under covariate shift conditions, DDRC maintains superior performance with 96.19% correlation with theoretical optima and positive improvements in 92.97% of observations. Average and median improvement ratios of 63.1% and 82.5%, respectively, demonstrate sustained value creation capacity despite training data limitations.

These results validate the robustness of ambiguity-adjusted reinforcement learning approaches to variation in training data quality and coverage, confirming that explicit uncertainty quantification enables sustained performance across diverse empirical environments commonly encountered in financial economics applications.

3.8 Methodological Synthesis and Implications

The DDRC framework addresses fundamental challenges in applying machine learning methodologies to economic decision-making problems characterized by distributional uncer-

tainty and limited representative data. Our empirical validation demonstrates that explicit uncertainty quantification through ensemble learning, combined with ambiguity-averse optimization, enables robust performance across realistic data limitations commonly encountered in financial contracting applications.

Methodological contributions. The integration of predictive environment modules with distributionally robust optimization represents a significant advancement in addressing the persistent challenge of distributional shift in economic applications. Unlike conventional machine learning approaches that assume training and deployment distributions remain constant, our framework explicitly models and adapts to distributional uncertainty, providing principled mechanisms for maintaining decision quality under realistic data limitations.

Economic implications. Our findings suggest that incorporating distributional robustness into financial decision-making frameworks can generate substantial economic value even when underlying distributional assumptions prove incorrect. The consistent improvement in decision quality across diverse experimental scenarios indicates that ambiguity aversion, properly calibrated through empirical uncertainty quantification, serves as an effective hedge against model misspecification risks inherent in complex economic environments.

Practical applications. The demonstrated robustness to both distributional and covariate shifts positions the DDRC methodology as particularly suitable for financial institutions operating in dynamic regulatory environments, emerging markets with limited historical data, or transition periods following structural economic changes. The framework's capacity to maintain performance quality while explicitly acknowledging uncertainty provides a principled foundation for risk-aware financial decision-making in practice.

The controlled validation within the Holmstrom and Tirole (1997) framework establishes three critical foundations for our empirical implementation on U.S. public firms. First, the demonstrated convergence to theoretical optima under perfect information conditions validates the fundamental algorithmic architecture underlying both PEM and DMM components. Second, the robust performance under distributional and covariate shifts provides

empirical confidence that our ambiguity quantification methodology will effectively identify regions of heightened prediction uncertainty in real-world applications where theoretical benchmarks are unavailable. Third, the superior performance under distributional uncertainty suggests that explicit ambiguity aversion, properly calibrated, can generate economic value even when the specific sources of uncertainty differ from those encountered in practice. These validation results inform our confidence intervals and performance interpretation in the subsequent empirical analysis of corporate decision-making across diverse industries and market conditions.

4 Data and Training

4.1 Data and Variables

Our panel data is fairly standard in corporate finance research and is based on quarterly CRSP–Compustat merged database, where we are able to include 20,485 different firms ranging from 1976:Q1 to 2023:Q2, with 784,460 firm-quarter observations. The variables in this paper fall into two classes: state variables (X_t) and managers' decision variables (u_t). In each period, managers make their decisions and their decisions will impact state variable dynamics for the next period. Examples of state variables include fundamental variables that are not entirely determined by managers, such as the firm's market capitalization, interest expenses, sales growth, earnings forecasts, and Tobin's Q. They also include past managerial decisions. Decision variables u_t reflect managerial decisions directly (or at least largely so). Examples include corporate investment spending, cash savings, and capital structure.

In order to train our supervised learning-based PEM, we require the existence of observations for two consecutive fiscal quarters in standard Compustat dataset (with "datafmt" code to be "STD"). We only focus on domestic (with "popsrc" code to be "D") industrial companies (with "indfmt" code to be "INDL"), and we exclude financial and utility firms (with "naics2" code to be 22 or 52) from our sample. We also require book assets to be positive. AlphaManager has a rolling training for RL. We divide our dataset by fiscal quarter to have roughly 30% initial training set, which implies that the initial burn-in is until

Table 1: Summary Statistics for Firm-specific States and Managerial Decisions

variable	count	mean	median	std
Leverage	784,460	0.2449	0.2020	0.2333
Acquisition	784,460	0.0047	0.0000	0.0227
Investment	784,460	0.0144	0.0068	0.0227
Cash	784,460	0.2041	0.0932	0.2651
Dividend	784,460	0.0019	0.0000	0.0051
DebtIssue	784,460	0.9122	0.0000	1.8047
EquityIssue	784,460	0.5231	0.0010	1.0551
RDExp	784,460	0.0130	0.0000	0.0309
Repurchase	784,460	0.0026	0.0000	0.0095
Total Assets	784,460	5.2788	5.0987	2.2292
Current Assets	784,460	4.2719	4.2478	2.1785
Sales	784,460	3.7052	3.6178	2.2490
Payables	784,460	2.7043	2.3038	2.0113
COGS	784,460	3.3015	3.1079	2.1523
InterestExp	784,460	0.5467	0.0000	1.1486
Inventory	784,460	2.4818	2.1510	2.2778
CurrentLiability	784,460	3.5870	3.3725	2.1935
Receivables	784,460	3.1032	2.9350	2.1419
Revenue	784,460	3.6999	3.6127	2.2525
MarketCap	784,460	5.2181	5.0620	2.2858
EnterpriseValue	784,460	5.7726	5.6021	2.2511
Volume	784,460	10.3049	10.4611	2.6685
Return	784,460	-0.0140	0.0000	0.2872

Note: This table documents the summary statistics of our firm-level state variables and decision variables. Our sample starts from 1976:Q1 to 2023:Q2, and covers 784,460 unique firm-quarter observations. In the sample selection, we keep domestic industrial firms (with "popsrc" code to be "D" and "indfmt" code to be "INDL") from standard Compustat (with "datafmt" code to be "STD"). We exclude financial and utility firms (with "naics2" code to be 22 or 52) from the sample. We also require book assets to be positive. We winsorize our each variable at 1% and 99%. The detailed variable definitions and formulations are in Appendix A.

1991. We winsorize each variable at 1% and 99%, and subsequently normalize them to be bounded inside $[-1, 1]$ for training convenience. We also consider macroeconomic states for the same time period, using the Chicago Fed National Financial Condition Index (NFCI) subindices (risk, credit, financial leverage, and non-financial leverage) as macro covariates. The detailed variable selection is listed in Appendix A. Summary statistics for firm-level variables are reported in Table 1.

4.2 Training and Computation

We describe the training of the AlphaManager model here and list the hyperparameter choices in Appendix B. PEM has 268 inputs, including 12 firm fundamental variables (e.g., log book assets), 12 variables capturing their growth (first difference in variable values). We further incorporate the corresponding variables lagged 1 through 4 quarters to capture patterns in the time series. Involving lagged variables provides additional information on time dependence, and incorporating the changes in values helps neural networks better extract more information than typically learned from the levels of state variables alone. We additionally include lagged decisions and their growth, which are publicly observable at the current time point, as well as their 1-4 quarter lagged versions. We also have 2 stock market states and their current growth, together with the 1-4 quarter lagged values. For the 4 macro states, we only consider their current levels and their levels lagged 1 through 4 quarters. For managerial decisions, we have 9 decision variables (e.g., leverage) and their correspondingly changes (growth) from their previous values. The output of PEM has 14 dimensions, namely 12 next-quarter changes in fundamental and 2 next-quarter changes in return and volume.

For training PEM, we first initialize a neural network and train the model in mini-batch mode (with batch size to be 2048 observations) using data until 1991:Q4, which is ready to predict system states in 1992:Q1 "out-of-sample." In this stage, we train our neural network for 30 epochs. Next, we include 1992:Q1 data into our training set to update the environment module and use the updated model to predict the system states in 1992:Q2 for additional 5 epochs, and so on. We also train 10 auxiliary models with different realizations in the initialization phase and with slightly different model designs at the same time for ambiguity estimation. We use four hidden layers in the baseline PEM, with each layer having 512 neurons. In total, we have 1,186,304 parameters.

Our DMM has 250 inputs, namely the inputs of PEM excluding current decisions and their growth versions ($9 + 9 = 18$), and has 9 outputs, which are changes in current decisions compared to the last-quarter ones. For training DMM, we initialize a neural network and train the RL model using data until 1991 to recommend managerial decisions in 1991. We train the model for 64 epochs first, and then when the loss function (after ambiguity

punishment) turns positive, we continue to train the model for 5 epochs and move on to the next stage. Then, we focus on 1992 data to recommend 1992 managerial decisions, and follow the same training scheme going forward in time. As baseline specifications, we consider two objectives, market capitalization growth and enterprise value growth. We do this for two horizons: one quarter ahead ("short horizon") and eight quarters ahead ("long horizon"). We use a four-hidden layer network in training our DMM, with each layer having 256 neurons. In total, we have 265,993 parameters.

We train our neural networks on the Red Cloud platform, with 16 CPUs, 55G memory, and an A100 GPU.[11] The training process takes 30-35 hours for PEM (including the auxiliary models), 8 hours for the DMM with short-term objectives, and 75-80 hours for the DMM under long-term objectives. We use Adam optimizer (Kingma and Ba, 2014) in training neural networks.

5 The Economic Environment for U.S. Public Firms

5.1 Supervised PEM: A Corporate Finance "World Model"

When implemented on U.S. equities, PEM predicts next period system evolution, including reactions to current period managerial actions. In other words, environment module predicts X_{t+1} given X_t and u_t. As discussed earlier, a deep neural network representation could easily fit the training sample, but is not guaranteed to produce sensible out-of-sample counterfactuals, especially in the presence of distributional shifts. Let us discuss this in turn.

Suppose there is a rarely observed or an unseen event in the training sample, which is the general focus of event studies and causal identification strategies. One can use a dummy variable to denote whether this event occurs or not. In the training sample, this variable is almost always labeled 0 and only rarely 1, which makes it a phenomenon too "weak" to pick up. In this case, the ambiguity for our test sample is infinite, i.e., any models trained on our training set could provide unconstrained estimations on the treatment effect. To address

[11] Red Cloud is a subscription-based Infrastructure as a Service cloud that provides root access to virtual servers and storage on-demand under Cornell University Center for Advanced Computing (link).

this concern of ambiguity, we follow the idea of boosting and train a set of auxiliary neural networks to estimate the ambiguity by their dispersion in predictions.[12]

Acting as a metric of ambiguity, boosting error differentiates scenarios where all models coincide in predictions versus ones where model predictions are highly dispersed. By punishing the boosting error in the training of our decision-making module later on, the ambiguity averse agent avoids high-ambiguity actions which we have little reference from prior empirical data, and only explores the action domain with low ambiguity. Similar ideas are also seen in computer science literature related to offline reinforcement learning, such as Kidambi et al. (2020) and Yu et al. (2020).

5.2 Predicting Firm Fundamentals and Market Reactions

To train PEM, we use standard mean-squared error (MSE) as our loss function. We standardize input variables to be in the range $[-1, 1]$ to fit the $tanh(\cdot)$ function in the neural networks, rather than normalizing them to be with zero mean and unit variance.

To better understand the economic implication of training loss, we convert the technical loss to pseudo R^2 by dependent variables. The expression of pseudo R^2 for the d^{th} state for a given sample S_d with corresponding predictions $\hat{S}_d$ is:

$$\text{Pseudo } R^2(S_d, \hat{S}_d) = 1 - \frac{\sum_{X_d \in S_d}(X_d - \hat{X}_d)^2}{\sum_{X_d \in S_d}(X_d - \bar{X}_d)^2}, \tag{7}$$

where $X_d \in S_d$ is a sample of the d^{th} dimension of state vector, and $\hat{X}_d \in \hat{S}_d$ is its corresponding prediction from PEM; $\bar{X}_d$ is the initial training set (i.e., sample before 1991:Q4) average of X_d. It is not surprising that our environment module gains very high pseudo R^2 for most variables, as decision variables are used as network inputs. The pseudo R^2 for stock return growth is not as high as other variables, since it is driven by investors' demand — which incorporates factors like sentiment and liquidity — that is out of managers' control, at least before the control decisions are revealed to the investors in subsequent quarters.

[12]The main difference between our boosting method and commonly used algorithms such as AdaBoost lies in how we use the auxiliary predictors. These auxiliary networks share a similar structure with our neural network but with different initialization. Traditional applications combine all the auxiliary predictors to be a stronger predictor, while we use auxiliary predictors to jointly form a natural gauge of model ambiguity.

Table 2: Pseudo R^2 for the Growth of State Variables Without and With Control Information

	Ignoring Control Information		Including Control Information	
State Variable	Training R^2	Test R^2	Training R^2	Test R^2
Total Assets	–4.09%	–8.15%	55.44%	62.56%
Current Assets	–3.58%	–7.10%	44.49%	51.21%
Sales	29.54%	28.68%	31.33%	30.88%
Payables	21.46%	24.43%	24.40%	27.64%
COGS	25.68%	26.76%	27.00%	28.56%
InterestExp	73.26%	77.17%	73.36%	77.28%
Inventory	12.78%	13.71%	17.04%	18.92%
CurrentLiability	8.88%	7.72%	21.89%	22.69%
Receivabless	17.52%	18.77%	21.59%	23.20%
Revenue	29.51%	28.59%	31.31%	30.80%
MarketCap	1.32%	–3.33%	9.32%	7.07%
EnterpriseValue	–0.97%	–5.73%	14.61%	13.14%
Volume	12.81%	16.53%	15.77%	20.75%
Return	47.90%	45.27%	50.04%	48.19%

Note: This table documents the pseudo R^2 for the growth of state variables without and with the current control (u_t). PEM intends to predict future firm-specific state variables given current state and control. Since the current control vector is private information (for instance the manager's planning on leverage for the next period), it is not observable by outsiders such as investors or econometricians. In predicting the future, we first ignore all controls, by setting them to be 0, calculate PEM predictions, and then the pseudo R^2 (ignoring control) for each firm-specific state variable. Then, we incorporate control variables as inputs, calculate PEM predictions, and then the pseudo R^2 (with control) for each firm-specific state variable.

Table 2 shows the pseudo R^2 of future fundamentals ignoring or with managerial decision information for training sets (in-sample) and test sets (out-of-sample). With managerial decision information, we exploit the full capacity of PEM to generate counterfactuals for future firm-level states by feeding in current state variables as well as managerial decisions to be made. To evaluate the importance of managerial planning in predictions, we ignore the managerial decisions by setting them to be their corresponding last-quart values — as if firms stick with their latest decisions — and generate counterfactuals.

By comparing pseudo R^2s of future fundamentals ignoring or with managerial planning, we are able to identify pivotal state variables that managerial decisions tend to influence more. For outcomes such as gross revenue or net income, whether we consider the managerial action does not matter much in predictions. However, managerial actions seem to have a significant impact on state variables such as total asset and market capitalization. This

reveals that some corporate outcomes are influenced by managerial actions while others are not, at least not immediately. We consider both the case with and without controls because from outside investors' perspective, managerial controls are only disclosed with a delay and constitute "insider" information that outsiders may not use for trading. Another way to interpret this "controllability" result is to look at the outcome variables less affected by the manager. For example, growth in trading volume is likely driven by traders in the secondary market rather than directly by the manager of the firm.

Table 3 reports PEM's heterogeneous performance (measured by OOS MSE) for each state variable by its cross section ranking. We report results for three subsets of cross sections: pre-dot com bubble, between the dot com bubble and the great financial crisis (GFC), and post-GFC, where cutoff points of these two events are based on the NBER recession end months. For each state variable, we rank firms in each cross section as "low" and "high" by comparing the state variable level to its cross-sectional median, and calculate average and standard deviation of PEM MSE within each subsample. For most firm-level state variables, the lower half generally has higher prediction error, with the exceptions of book value of current asset, interest paid and book value of current liability. For macro state variables, the dot com to GFC period is the hardest to predict; the same applies to our two objectives (market capitalization and enterprise value). Table 4 reports PEM's heterogeneous performance by sector, defined as the first digit of North American Industry Classification System (NAICS) code. Trade and Transportation sector ("naics1" = 4), Education and Healthcare ("naics1" = 6), and Other Services sector ("naics1" = 8) always have the lowest prediction error across three episodes. Table 5 reports PEM's performance by book-to-market decile, where 1 indicates the lowest decile and 10 indicates the highest decile. The best performers with regard to prediction error of PEM oscillate among middle deciles and the distribution across deciles exhibits a U-shape MSE, while the bottom decile always has the worst performance.

Table 6 results reveal heterogeneous ambiguity for each state variable (measured by boosting error defined in the previous section), following the same logic. For most firm-level state variables, the lower half generally has lower ambiguity. The dot com to GFC period usually has the highest ambiguity, and the pre dot com period has the lowest in general. In

Table 7, we do the same exercise for sector. Manufacturing ("naics1" = 3), Education and Healthcare ("naics1" = 6) and Other Services ("naics1" = 8) achieve the lowest ambiguity across three episodes. Mining and Construction ("naics1" = 2) has low ambiguity in the pre dot com period, while the ambiguity for Education and Healthcare sector ("naics1" = 6) gradually goes up. Table 8 shows that the highest three deciles of book-to-market has the lowest ambiguity over the three episodes.

Recapitalization example. Even though PEM is high-dimensional, we can analyze low dimensional action combinations, which are the focus of conventional reduced-form models. Appendix D uses firm recapitalization as an example to showcase how to use PEM to analyze low-dimensional policy conterfactuals.

5.3 Predicting the Evolution of Macroeconomic States

Macroeconomic state variables are distinct from firm-specific fundamentals or stock market reactions. These variables operate on a broader scale, influence firm fundamentals, and are largely insulated from the direct influence of individual firm decisions. For example, changes in an individual firm's leverage — whether it issues more debt or equity — are unlikely to substantially affect macroeconomic indicators like aggregate non-financial leverage or credit risk. These macroeconomic variables tend to be co-integrated and exhibit minimal cross-sectional variability, making them difficult to predict with precision.

In this context, the National Financial Conditions Index (NFCI) subindices, including risk, credit, financial leverage, and non-financial leverage, provide an important summary of macroeconomic conditions. However, predicting these macroeconomic variables poses a significant challenge due to their overlapping information content and their strong interrelationships. To investigate whether neural networks can overcome this, we compare the predictive performance of neural networks and vector autoregressive (VAR) models, which have long been considered a gold standard in time series forecasting for macroeconomic variables.

Our neural network approach achieves much better performance (measured by out-of-sample pseudo R^2s) in all four dimensions, especially for risk and credit where VAR fails

Table 3: Heterogeneous PEM Performance for System States

Variable		Full Sample		Pre-Dotcom		Dotcom–GFC		Post-GFC	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std
Total Assets	high	2.16%	8.53%	2.57%	9.42%	2.23%	8.48%	1.76%	7.68%
	low	4.40%	13.33%	4.84%	13.77%	4.30%	12.80%	4.08%	13.26%
Current Assets	high	3.49%	10.23%	3.57%	10.30%	3.54%	10.15%	3.39%	10.22%
	low	5.52%	14.97%	5.88%	15.17%	5.21%	14.43%	5.39%	15.13%
Sales	high	3.28%	10.11%	3.53%	10.49%	3.19%	9.96%	3.13%	9.86%
	low	6.86%	16.78%	6.53%	15.50%	6.36%	15.57%	7.46%	18.49%
Payables	high	6.12%	15.03%	7.15%	16.14%	6.28%	15.31%	5.11%	13.73%
	low	6.85%	14.86%	5.73%	13.04%	6.68%	14.44%	7.92%	16.46%
COGS	high	2.84%	11.01%	3.30%	12.32%	2.74%	10.64%	2.50%	9.95%
	low	4.49%	13.75%	4.94%	14.67%	4.32%	13.33%	4.20%	13.15%
InterestExp	high	4.90%	15.19%	3.54%	11.93%	4.67%	14.68%	6.21%	17.69%
	low	3.53%	15.08%	2.81%	12.38%	3.02%	13.80%	4.50%	17.78%
Inventory	high	5.29%	13.43%	5.84%	14.23%	5.58%	13.86%	4.64%	12.36%
	low	4.88%	15.53%	4.89%	15.13%	4.62%	14.98%	5.03%	16.21%
CurrentLiability	high	5.48%	13.98%	6.12%	14.75%	5.59%	14.20%	4.84%	13.09%
	low	6.69%	15.66%	6.83%	15.42%	6.58%	15.47%	6.64%	15.98%
Receivables	high	3.72%	10.62%	4.00%	11.15%	3.80%	10.55%	3.41%	10.17%
	low	7.26%	17.24%	6.37%	15.90%	7.05%	16.46%	8.16%	18.73%
Revenue	high	3.27%	10.14%	3.53%	10.58%	3.18%	10.02%	3.11%	9.81%
	low	6.84%	16.78%	6.53%	15.56%	6.35%	15.63%	7.41%	18.40%
MarketCap	high	7.49%	15.51%	9.77%	18.45%	7.79%	15.74%	5.30%	11.83%
	low	12.42%	21.40%	12.88%	21.69%	13.42%	22.53%	11.40%	20.34%
EnterpriseValue	high	6.11%	13.87%	8.60%	17.90%	5.99%	12.75%	4.02%	9.47%
	low	10.37%	19.01%	11.55%	20.73%	11.34%	19.67%	8.73%	16.75%
Volume	high	4.17%	9.22%	5.39%	9.97%	3.64%	8.59%	3.45%	8.80%
	low	8.44%	15.77%	9.81%	16.08%	8.86%	16.30%	6.98%	15.01%
Return	high	5.31%	11.18%	6.17%	12.10%	5.36%	11.38%	4.53%	10.11%
	low	6.73%	14.30%	7.52%	15.05%	7.27%	15.20%	5.70%	12.91%
MacroRisk	high	4.81%	5.51%	6.90%	7.23%	6.24%	6.75%	5.23%	6.26%
	low	5.74%	6.57%	5.25%	5.93%	5.11%	5.92%	5.43%	6.62%
MacroCredit	high	4.80%	5.53%	6.90%	7.23%	6.48%	6.89%	5.45%	6.52%
	low	5.74%	6.57%	5.49%	6.15%	4.92%	5.75%	5.22%	6.34%
MacroFinLev	high	4.81%	5.51%	5.99%	6.64%	5.80%	6.53%	5.75%	6.70%
	low	5.75%	6.58%	6.16%	6.69%	5.56%	6.23%	4.89%	6.10%
MacroNonfinLev	high	5.10%	5.72%	6.82%	7.18%	5.71%	6.48%	5.46%	6.61%
	low	5.71%	6.55%	5.25%	5.93%	5.64%	6.28%	5.18%	6.23%

Note: This table shows the heterogeneous PEM performance for each state variable. For each cross section, we divide firms in low and high groups for each state variable, and we calculate mean and standard deviation of PEM MSE for the focal state variable within three subsamples: pre dot com, between dot com and GFC, and post GFC.

Table 4: Heterogeneous PEM Performance for Sectors

MSE	full sample		pre-dotcom		dotcom–GFC		post-GFC	
sector	mean	std	mean	std	mean	std	mean	std
agriculture	8.37%	8.06%	8.84%	8.70%	8.46%	7.78%	7.88%	7.57%
mining & construction	7.51%	7.51%	7.04%	7.06%	7.64%	7.52%	7.78%	7.80%
manufacturing	5.56%	6.49%	5.68%	6.37%	5.66%	6.47%	5.37%	6.61%
trade & transportation	5.06%	5.85%	5.41%	6.06%	4.94%	5.72%	4.83%	5.72%
information & professional services	5.69%	6.48%	6.95%	7.32%	5.63%	6.19%	4.73%	5.73%
education & healthcare	4.89%	5.78%	5.67%	6.28%	4.41%	5.11%	4.48%	5.63%
recreation & accommodation	5.50%	6.10%	5.99%	6.09%	5.31%	5.56%	5.14%	6.38%
other services	4.54%	5.32%	4.75%	5.45%	4.13%	4.88%	4.54%	5.46%
public administration	7.63%	7.94%	8.10%	8.17%	7.70%	8.02%	5.81%	6.59%

Note: This table shows the heterogeneous PEM performance for each sector, defined by the first digit of North American Industry Classification System (NAICS) code. For each firm-quarter observation, we calculate average MSE of PEM for state variables to get the average MSE of PEM for each firm-quarter observation, and then for each cross section, we calculate mean and standard deviation of PEM MSE within three subsamples: pre dot com, between dot com and GFC, and post GFC, and we take the average across state variables.

Table 5: Heterogeneous PEM Performance for "Value" Deciles

MSE	full sample		pre-dotcom		dotcom–GFC		post-GFC	
book-to-market decile	mean	std	mean	std	mean	std	mean	std
1	6.76%	7.36%	7.92%	7.87%	6.47%	7.03%	5.68%	6.81%
2	5.74%	6.58%	6.68%	7.18%	5.58%	6.20%	4.81%	6.01%
3	5.28%	6.23%	6.03%	6.72%	5.23%	6.06%	4.45%	5.66%
4	5.12%	6.05%	5.64%	6.38%	5.21%	6.00%	4.45%	5.62%
5	4.95%	5.85%	5.35%	6.09%	5.12%	5.91%	4.35%	5.46%
6	5.02%	5.85%	5.36%	6.08%	5.27%	6.05%	4.41%	5.34%
7	5.24%	6.09%	5.49%	6.20%	5.45%	6.31%	4.77%	5.72%
8	5.59%	6.25%	5.55%	5.92%	5.82%	6.40%	5.46%	6.46%
9	6.08%	6.44%	5.75%	6.01%	6.41%	6.68%	6.17%	6.66%
10	7.86%	7.32%	7.13%	6.63%	8.37%	7.48%	8.22%	7.78%

Note: This table shows the heterogeneous PEM performance for each book-to-market decile. For each firm-quarter observation, we calculate average MSE of PEM for state variables to get the average MSE of PEM for each firm-quarter observation, and then for each cross section, we calculate mean and standard deviation of PEM MSE within three subsamples: pre dot com, between dot com and GFC, and post GFC, and we take the average across state variables.

Table 6: PEM: Heterogeneous Ambiguity for System States

variable		full sample		pre-dotcom		dotcom–GFC		post-GFC	
		mean	std	mean	std	mean	std	mean	std
Total Assets	high	6.93%	3.72%	7.18%	3.74%	7.02%	3.65%	6.64%	3.72%
	low	6.82%	3.79%	6.90%	3.84%	6.84%	3.69%	6.72%	3.82%
Current Assets	high	6.41%	3.00%	6.20%	3.06%	6.54%	2.95%	6.50%	2.96%
	low	6.41%	3.17%	6.13%	3.17%	6.57%	3.09%	6.54%	3.21%
Sales	high	6.31%	3.18%	5.91%	2.85%	6.43%	3.22%	6.58%	3.38%
	low	5.74%	2.90%	5.20%	2.60%	5.76%	2.79%	6.21%	3.12%
Payables	high	5.91%	2.78%	5.97%	2.73%	6.02%	2.79%	5.79%	2.81%
	low	5.53%	2.76%	5.30%	2.56%	5.44%	2.70%	5.78%	2.94%
COGS	high	5.50%	2.70%	5.46%	2.40%	5.68%	2.68%	5.42%	2.93%
	low	5.21%	2.67%	4.97%	2.33%	5.42%	2.70%	5.30%	2.89%
InterestExp	high	3.20%	3.46%	2.99%	3.30%	3.18%	3.34%	3.41%	3.66%
	low	2.38%	2.50%	2.63%	2.50%	2.32%	2.34%	2.20%	2.58%
Inventory	high	5.38%	2.50%	5.29%	2.35%	5.35%	2.50%	5.48%	2.61%
	low	5.15%	2.64%	5.23%	2.44%	5.03%	2.52%	5.15%	2.88%
CurrentLiability	high	7.30%	3.67%	7.21%	3.52%	7.44%	3.66%	7.29%	3.79%
	low	6.84%	3.55%	6.57%	3.49%	6.91%	3.46%	7.02%	3.63%
Receivables	high	5.16%	2.44%	4.94%	2.42%	5.42%	2.47%	5.19%	2.43%
	low	5.05%	2.51%	4.63%	2.38%	5.23%	2.46%	5.31%	2.61%
Revenue	high	6.29%	3.17%	5.88%	2.82%	6.37%	3.21%	6.58%	3.39%
	low	5.73%	2.89%	5.20%	2.59%	5.72%	2.78%	6.20%	3.11%
MarketCap	high	4.68%	2.78%	4.51%	2.40%	5.54%	3.59%	4.28%	2.35%
	low	5.27%	3.20%	4.49%	2.31%	6.03%	3.83%	5.47%	3.29%
EnterpriseValue	high	5.85%	2.67%	5.70%	2.50%	6.57%	2.98%	5.54%	2.53%
	low	6.11%	2.91%	5.45%	2.52%	6.78%	3.15%	6.27%	2.94%
Volume	high	5.26%	2.29%	5.73%	2.57%	5.47%	2.14%	4.72%	2.00%
	low	5.15%	2.33%	5.22%	2.46%	5.44%	2.34%	4.91%	2.18%
Return	high	9.57%	4.29%	9.14%	3.91%	10.10%	4.34%	9.60%	4.53%
	low	9.95%	4.35%	9.52%	4.15%	10.27%	4.35%	10.12%	4.50%
MacroRisk	high	5.63%	1.82%	5.98%	1.91%	6.33%	1.86%	5.93%	1.92%
	low	5.91%	1.94%	5.42%	1.84%	5.87%	1.87%	5.94%	2.04%
MacroCredit	high	5.64%	1.83%	5.88%	1.88%	6.27%	1.87%	6.00%	1.96%
	low	5.91%	1.94%	5.57%	1.90%	5.94%	1.87%	5.88%	1.99%
MacroFinLev	high	5.63%	1.83%	5.67%	1.89%	6.25%	1.92%	6.01%	1.95%
	low	5.92%	1.94%	5.72%	1.90%	5.97%	1.83%	5.86%	2.01%
MacroNonfinLev	high	5.78%	1.83%	5.99%	1.91%	6.37%	1.89%	5.91%	1.98%
	low	5.90%	1.94%	5.37%	1.83%	5.86%	1.83%	5.96%	1.98%

Note: This table shows the heterogeneous ambiguity for each state variable measured by the greatest different in predictions among PEM and its auxiliary models for that specific state variable. For each cross section, we divide firms in low and high groups for each state variable, and we calculate mean and standard deviation of ambiguity for the focal state variable within three subsamples: pre dot com, between dot com and GFC, and post GFC.

Table 7: PEM: Heterogeneous Ambiguity for Sectors

Ambiguity	full sample		pre-dotcom		dotcom–GFC		post-GFC	
sector	mean	std	mean	std	mean	std	mean	std
agriculture	6.71%	2.11%	6.57%	2.14%	7.07%	2.26%	6.63%	1.97%
mining & construction	6.35%	1.97%	6.01%	1.91%	6.76%	1.94%	6.39%	1.99%
manufacturing	5.74%	1.92%	5.50%	1.84%	5.96%	1.86%	5.82%	1.99%
trade & transportation	5.96%	1.92%	5.69%	1.83%	6.17%	1.91%	6.08%	1.97%
information & professional services	6.00%	1.93%	5.98%	1.98%	6.16%	1.86%	5.91%	1.93%
education & healthcare	5.75%	1.90%	5.64%	1.94%	5.88%	1.72%	5.76%	1.96%
recreation & accommodation	6.09%	1.86%	5.93%	1.82%	6.31%	1.80%	6.10%	1.91%
other services	5.58%	1.89%	5.27%	1.76%	5.92%	1.92%	5.80%	1.99%
public administration	6.09%	2.00%	6.00%	1.97%	6.27%	1.98%	6.02%	2.13%

Note: This table shows the heterogeneous ambiguity for each sector, defined by the first digit of North American Industry Classification System (NAICS) code. The ambiguity is measured by the greatest different in predictions among PEM and its auxiliary models for each state variable. For each firm-quarter observation, we calculate average ambiguity of PEM for state variables to get the average ambiguity of PEM for each firm-quarter observation, and then for each cross section, we calculate mean and standard deviation of PEM ambiguity within three subsamples: pre dot com, between dot com and GFC, and post GFC, and we take the average across state variables.

Table 8: PEM: Heterogeneous Ambiguity for "Value" Deciles

Ambiguity	full sample		pre-dotcom		dotcom-GFC		post-GFC	
book-to-market decile	mean	std	mean	std	mean	std	mean	std
1	6.42%	1.98%	6.32%	1.97%	6.56%	1.90%	6.41%	2.04%
2	6.04%	1.88%	5.91%	1.87%	6.24%	1.85%	6.02%	1.92%
3	5.83%	1.88%	5.72%	1.87%	6.03%	1.82%	5.80%	1.91%
4	5.73%	1.89%	5.60%	1.91%	5.99%	1.83%	5.67%	1.88%
5	5.66%	1.89%	5.55%	1.91%	5.94%	1.84%	5.57%	1.88%
6	5.65%	1.91%	5.46%	1.88%	5.93%	1.87%	5.63%	1.94%
7	5.64%	1.88%	5.45%	1.83%	5.93%	1.89%	5.60%	1.89%
8	5.71%	1.89%	5.49%	1.83%	6.00%	1.89%	5.73%	1.93%
9	5.82%	1.90%	5.46%	1.77%	6.14%	1.92%	5.93%	1.95%
10	6.26%	1.87%	5.96%	1.70%	6.49%	1.94%	6.40%	1.94%

Note: This table shows the heterogeneous ambiguity for each book-to-market decile. The ambiguity is measured by the greatest different in predictions among PEM and its auxiliary models for each state variable. For each firm-quarter observation, we calculate average ambiguity of PEM for state variables to get the average ambiguity of PEM for each firm-quarter observation, and then for each cross section, we calculate mean and standard deviation of PEM ambiguity within three subsamples: pre dot com, between dot com and GFC, and post GFC, and we take the average across state variables.

to attain positive pseudo R^2s while our neural network gets 11.4% and 27.3% out-of-sample pseudo R^2s respectively. We detail the time series results in Appendix E.

6 Optimal Policy Recommendations to Managers

6.1 Reinforcement Learning and the Decision-Making Module

The decision-making module, built on the predictive environment module, intends to seek optimal managerial decision-making processes given a certain utility function (i.e., objective) of the manager. Managers make decisions contingent on the current state of the firm including the current macroeconomic status, a neural network could projection the decision-making process as a function of current states: $u_t = g(X_t)$. The neural network is trained to maximize the expected objective value predicted by PEM. To mitigate the uncertainty of predictive environment module under unusual decisions, we punish directly the boosting error defined in Equation (4) and train a model based on the discrete version of utility function in Equation (5).

We start from a randomly initialized neural networks whose inputs are current states and outputs are current decisions to be made. We optimize an exogenously specified objective function using RL. The value of the objective function is given by the counterfactual statements from PEM. For each epoch, the decision-making module first gives out a set of decisions to be made, then together with state variables, these decisions are fed into PEM to generate counterfactual market reactions in the next quarter. Finally, the difference between counterfactual market capitalization (alternatively, enterprise value) and current market capitalization (enterprise value) is treated as the objective function value for this trait. After the feed-forward process is complete, backward propagation takes place and the parameters of the DMM are updated, given the loss function to be negative objective value.

We consider a set of reasonable objective functions and also a spectrum of horizons for managers' objectives. Specifically, we consider both short-term (1 quarter) and long-term (2 years) market capitalization and enterprise value as our objectives. We ignore agency problems and take managers as maximizing shareholders' value or overall investors'

value. The DDRC framework is flexible to allow other objectives and these two are useful benchmarks to consider.

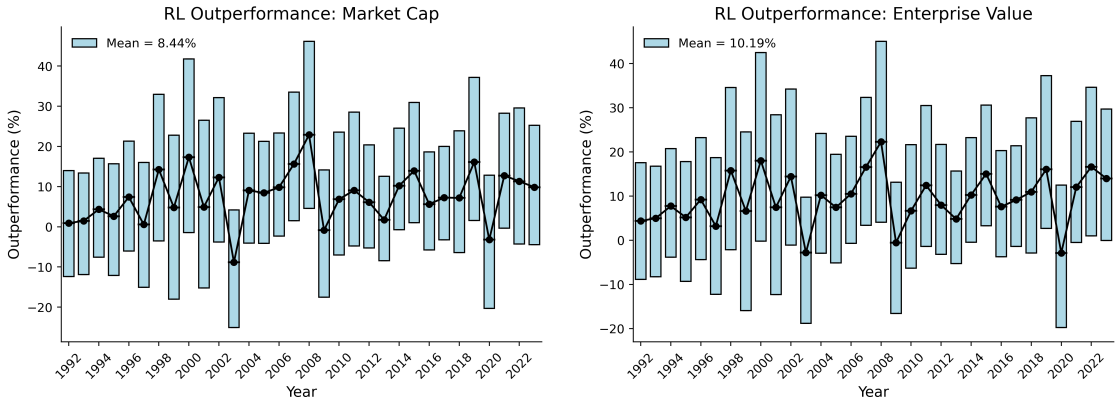
6.2 Performance Under Various Managerial Objectives

Long-term objectives maximizing valuation growth in 2 years. To outperform in the longer run, AlphaManager needs to have a longer horizon in its objective. We therefore examine how AlphaManager performs if the object is to maximize increases in the market capitalization or enterprise value over the subsequent 8 quarters instead of just over the next quarter. For long-term model with capitalization growth as the objective function, the average quarterly return is 8.73% while the first quarter performance is on average only 3.09%. When using enterprise value growth as the objective function, these two numbers become 4.43% and 1.28%, respectively.

Short-term objectives centering on next quarter valuation growth. We then consider a somewhat myopic objective of increasing market equity valuation over the next quarter. We estimate the corresponding short-term ambiguity constraint parameters under optimal long-term AM agents to make the results comparable. The overall average out-performance is 8.44% quarterly in the short-term when using market capitalization growth as the objective. When the enterprise value growth serves as the objective, the out-performance becomes 10.19%. The cross-sectional dispersion is non-trivial, as seen in the 75th and 25th percentile firms in Figure 3.

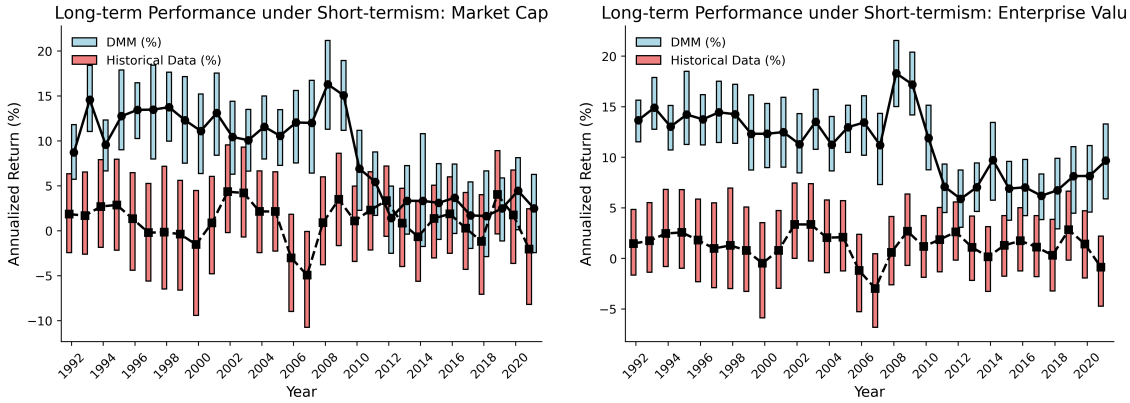
Figure 4 depicts what myopic behavior (a focus on next quarter outcomes) implies for the long-term (the next two-year) performance of the firms. This relationship is tracked over 30 years of data. *Ex-ante*, several outcomes could appear. On the one hand, one could argue that managers historically may not be myopic, and as such, historical actions may generate better long-term performance if short-termism hurts firm fundamentals in the long run. On the other hand, even with potentially longer-term goals, managers could have taken very suboptimal actions historically. The 1-quarter-market-cap-focused myopic AlphaManager can end up perform much better. In the data, we observe that though AlphaManager outperforms

Figure 3: Out-performance of AlphaManager with Short-term Objectives



Note: This figure shows the out-performance of AlphaManager trained with short-term market cap growth (left) and enterprise value growth (right) as objective functions. The solid line shows the median performance each year, the bars show the 75th and 25th percentiles respectively. The y-axis is in percentage of the change in objective for the next quarter predicted by the PEM.

Figure 4: Long-Term Performance Under Short-Termism

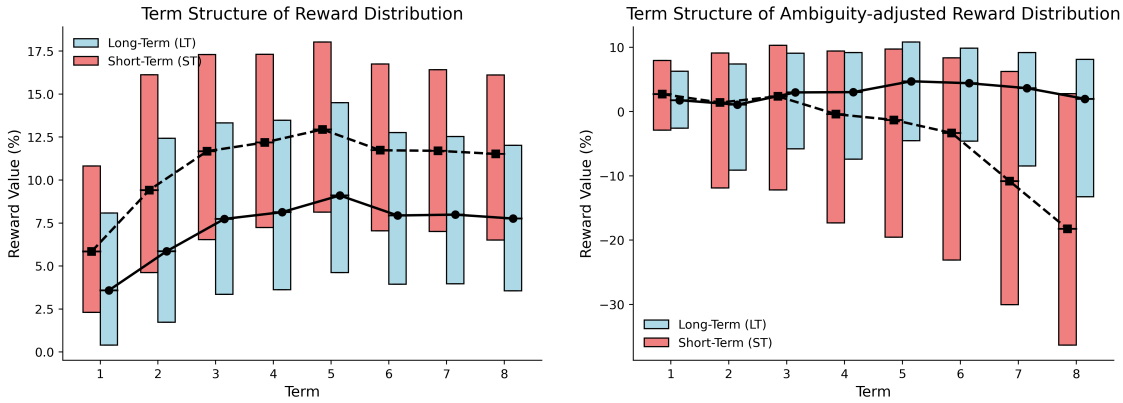


Note: This figure shows the long-term performance of AlphaManager trained with short-term market cap growth (left) and enterprise value growth (right) as objective functions. The dashed line shows the median average performance under historical managerial actions, while the solid line shows the median performance of actions suggested by the DMM module.

under the specified objective, over the long term (i.e., for a different objective with longer horizon) it underperforms. Algorithmic predictions do not capture cross-sectional "risk" (in-model volatility), so the algorithmic predictions (black lines) have much smaller variations.

Figure 5 shows the term structure for short-term and long-term AM under the objective of market cap growth. In the left panel, the short-term AM is out performing the long-term AM in longer terms' rewards, but this out performance is at the cost of high ambiguity. In

Figure 5: Term Structure of Reward and Ambiguity-adjusted Reward



Note: This figure shows the term structure for AlphaManager trained with short-term market cap growth with reward (left) and ambiguity-adjusted reward (right). The ambiguity adjustment follows Eq (8). The solid line and blue bars show the median performance of long-term AM, while the dashed line and red bars show the median performance of short-term AM. Lines represent the median value and bars represent the 25th to 75th percentiles.

the right panel, we adjust reward (i.e., market cap growth) using the formula:

$$\text{Adj. Reward}_t = \frac{1 + \text{Reward}_t}{\sqrt{\max\{1, \frac{\text{BoostingError}_t}{\text{BoostingError}_{t_0}}\}}} - 1, \tag{8}$$

where Reward_t is the reward that AM achieves under the PEM at time t, BoostingError_t is the boosting error for the AM-suggested managerial decisions at time t, and $\text{BoostingError}_{t_0}$ is the benchmark boosting error at time t_0. When the ambiguity is as low as the benchmark ambiguity, there is no punishment on the reward; the adjusted reward is decreasing in ambiguity of AM. After the adjustment, even though the short-term AM outperforms in the short-term, the long-run performance is worse than the long-term AM, and further into the future, the gap between adjusted rewards from short-term AM and long-term AM is also wider.

Table 9 reports heterogeneous predicted performance under the objective of 2-year enterprise value growth for the RL-recommended decisions by each state's cross-sectional ranking. It is worth noting that AM's impressive performance is driven by small and illiquid firms during high-risk episodes. The pre-dot com period enjoys the highest performance of AM. Table 10 documents the heterogeneous performance of RL. Manufacturing sector has high performance across three episodes. Table 8 shows the heterogeneity by book-to-market deciles.

The best performers are gradually shifting down as time goes by.

The analysis shows that AlphaManager again outperforms historical actions and there is heterogeneity in the cross-section. The average outperformance is much lower, reflecting that over the longer horizon, we just cannot predict much. Notably, these patterns also emerge in papers based on CEO surveys (e.g., Graham, 2022).

6.3 AI-Recommended Actions

We analyze the managerial actions recommended by AlphaManager under the four specified objective functions. Our analysis compares actual historical decisions (e.g., leverage levels) with model-generated recommendations for changes in these variables (e.g., leverage growth rates). When model-implied decisions exceed feasible bounds (such as negative leverage), we apply boundary constraints (setting them to zero, for instance).

First, with the objective of next-quarter market capitalization maximization (Figure 6), AlphaManager's recommendations diverge from historical practices in several key aspects: (1) pursuing more aggressive acquisition strategies, (2) maintaining higher cash reserves, (3) adopting a more equity-heavy capital structure while reducing debt (thereby decreasing leverage), (4) increasing dividend payouts, and (5) boosting investment expenditures, particularly in R&D. The model further suggests managers should tolerate greater investment variability and implement more share repurchases during economic downturns.

In Figure 7, when optimizing for short-term enterprise value maximization, AlphaManager generates recommendations largely consistent with the market capitalization objective, with one notable exception: our DMM suggests maintaining or even moderately increasing debt issuance rather than reducing it. This recommendation reflects the general benefits of leverage, particularly during the pre-financial crisis period, where higher leverage appears to have been value-enhancing.

In Figure 8, under the objective of maximizing long-term market capitalization growth (over eight quarters), AlphaManager recommends several strategic adjustments relative to historical practices: (1) pursuing more acquisitions, particularly for small firms; (2) maintaining higher cash reserves; (3) moderately increasing debt issuance while also issuing more

Table 9: Heterogeneous Performance of AlphaManager Across Firms

variable		full sample		pre-dotcom		dotcom-GFC		post-GFC	
		mean	std	mean	std	mean	std	mean	std
Total Assets	high	7.04%	6.26%	5.99%	5.90%	7.69%	7.48%	7.54%	5.57%
	low	6.93%	3.72%	6.17%	8.22%	9.95%	12.16%	13.75%	10.69%
Current Assets	high	7.51%	6.87%	6.07%	6.20%	8.28%	8.13%	8.28%	6.33%
	low	6.41%	3.00%	6.08%	7.99%	9.35%	11.82%	13.02%	10.64%
Sales	high	7.41%	6.17%	6.44%	5.61%	8.00%	7.26%	7.87%	5.76%
	low	6.31%	3.18%	5.71%	8.40%	9.64%	12.34%	13.42%	10.77%
Payables	high	7.36%	6.51%	6.26%	5.86%	7.94%	7.68%	7.96%	6.12%
	low	5.91%	2.78%	5.90%	8.24%	9.69%	12.08%	13.33%	10.62%
COGS	high	7.34%	6.24%	6.45%	5.69%	7.85%	7.36%	7.79%	5.81%
	low	5.50%	2.70%	5.71%	8.35%	9.78%	12.26%	13.51%	10.70%
InterestExp	high	8.26%	8.08%	6.57%	6.80%	8.39%	9.29%	9.63%	8.04%
	low	3.20%	3.46%	5.61%	7.44%	9.19%	10.85%	11.64%	9.87%
Inventory	high	7.86%	7.12%	6.32%	6.03%	8.56%	8.29%	8.76%	6.97%
	low	5.38%	2.50%	5.84%	8.11%	9.08%	11.72%	12.53%	10.44%
CurrentLiability	high	7.43%	6.55%	6.27%	5.91%	8.04%	7.78%	8.05%	6.07%
	low	7.30%	3.67%	5.88%	8.20%	9.59%	12.02%	13.25%	10.69%
Receivables	high	7.46%	6.46%	6.32%	5.79%	8.07%	7.68%	8.09%	6.02%
	low	5.16%	2.44%	5.84%	8.29%	9.57%	12.09%	13.21%	10.73%
Revenue	high	7.41%	6.17%	6.44%	5.61%	8.00%	7.26%	7.87%	5.76%
	low	6.29%	3.17%	5.71%	8.40%	9.63%	12.34%	13.42%	10.77%
MarketCap	high	6.96%	6.98%	5.32%	6.59%	7.69%	8.22%	7.93%	6.14%
	low	4.68%	2.78%	6.83%	7.60%	9.95%	11.67%	13.36%	10.59%
EnterpriseValue	high	6.88%	6.62%	5.52%	6.34%	7.54%	7.85%	7.64%	5.78%
	low	5.85%	2.67%	6.64%	7.84%	10.10%	11.90%	13.65%	10.63%
Volume	high	7.55%	8.86%	4.91%	7.47%	7.96%	9.98%	9.60%	8.64%
	low	5.26%	2.29%	7.25%	6.62%	9.68%	10.26%	11.69%	9.37%
Return	high	7.64%	8.64%	4.92%	7.06%	7.64%	9.78%	10.01%	8.42%
	low	9.57%	4.29%	7.23%	7.05%	10.00%	10.40%	11.29%	9.64%
MacroRisk	high	8.02%	8.20%	7.54%	8.02%	8.98%	10.99%	10.44%	9.19%
	low	8.63%	9.03%	4.60%	5.79%	8.66%	9.24%	10.91%	8.92%
MacroCredit	high	7.73%	8.09%	7.27%	8.04%	7.70%	10.59%	10.80%	9.42%
	low	8.66%	9.04%	5.21%	6.29%	9.86%	9.62%	10.52%	8.77%
MacroFinLev	high	8.48%	8.10%	6.43%	7.26%	9.95%	10.33%	10.92%	9.49%
	low	8.60%	9.05%	5.78%	7.05%	7.81%	9.89%	10.38%	8.62%
MacroNonfinLev	high	8.60%	8.58%	7.73%	7.96%	11.04%	10.63%	11.38%	9.56%
	low	8.59%	9.00%	4.22%	5.57%	6.88%	9.30%	9.92%	8.50%

Note: This table shows the heterogeneous performance under RL-recommended decisions by each state variable. The objective is chosen as 1-quarter market cap change. For each cross section, we divide firms in low and high groups for each state variable, and we calculate mean and standard deviation of the performance of RL-recommended decisions within three subsamples: pre dot com, between dot com and GFC, and post GFC.

Table 10: Heterogeneous Performance of AlphaManager Across Industries

RL performance (quarterly avg)	full sample		pre-dotcom		dotcom-GFC		post-GFC	
sector	mean	std	mean	std	mean	std	mean	std
agriculture	6.23%	7.49%	4.33%	6.12%	5.69%	7.12%	8.31%	8.30%
mining & construction	7.28%	7.58%	4.99%	6.62%	7.15%	8.47%	8.95%	7.34%
manufacturing	9.01%	9.71%	5.90%	7.38%	9.48%	10.84%	11.54%	10.01%
trade & transportation	7.86%	6.89%	6.80%	6.05%	8.32%	7.70%	8.51%	6.99%
information & professional services	8.81%	9.12%	6.27%	7.48%	8.55%	10.50%	11.04%	8.75%
education & healthcare	8.75%	8.06%	6.57%	6.56%	8.54%	8.89%	10.84%	8.19%
recreation & accommodation	7.35%	6.79%	6.64%	6.30%	7.09%	7.56%	8.19%	6.64%
other services	6.92%	6.30%	6.78%	6.55%	6.76%	6.16%	7.31%	5.98%
public administration	6.72%	9.36%	5.28%	8.02%	7.96%	10.37%	9.46%	10.71%

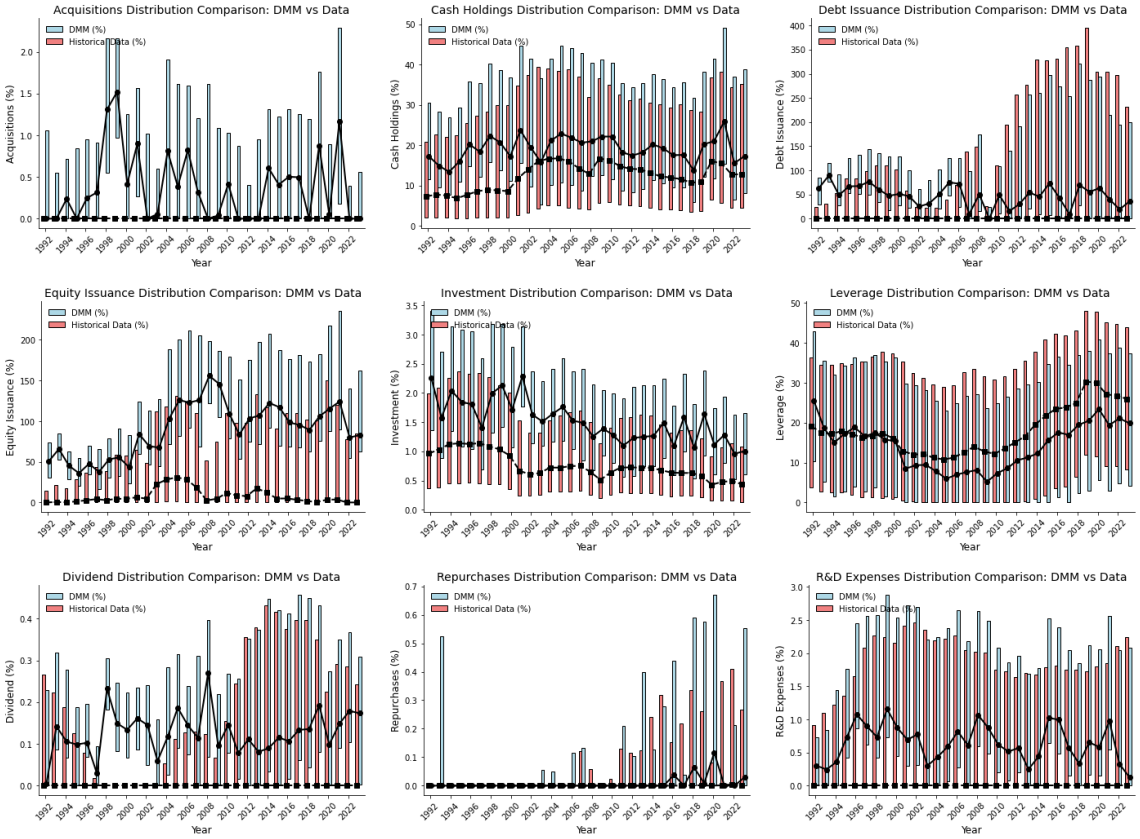
Note: This table shows the heterogeneous performance under RL-recommended decisions by each sector, defined by the first digit of North American Industry Classification System (NAICS) code. For each cross section, we calculate mean and standard deviation of the performance of RL-recommended decisions within three subsamples: pre dot com, between dot com and GFC, and post GFC.

Table 11: Heterogeneous Performance of AlphaManager Across Value Deciles

RL performance (quarterly avg)	full sample		pre-dotcom		dotcom-GFC		post-GFC	
book-to-market decile	mean	std	mean	std	mean	std	mean	std
1	6.62%	10.92%	2.91%	9.15%	6.66%	12.65%	10.79%	9.71%
2	7.91%	9.74%	4.78%	7.89%	8.48%	11.11%	11.01%	9.37%
3	8.31%	9.02%	5.67%	6.92%	8.96%	10.41%	10.79%	9.13%
4	8.61%	8.60%	6.34%	6.62%	9.25%	9.90%	10.67%	8.86%
5	8.82%	8.35%	6.75%	6.35%	9.43%	9.52%	10.64%	8.79%
6	8.84%	8.11%	7.17%	6.31%	9.36%	9.24%	10.31%	8.54%
7	9.07%	8.17%	7.61%	6.33%	9.24%	9.04%	10.59%	8.92%
8	9.13%	8.16%	7.52%	6.14%	9.45%	9.15%	10.63%	8.92%
9	8.83%	8.04%	6.86%	5.95%	9.34%	9.19%	10.36%	8.52%
10	8.32%	7.84%	6.95%	6.44%	7.75%	8.35%	9.98%	8.40%

Note: This table shows the heterogeneous performance under RL-recommended decisions by each book-to-market decile. For each cross section, we calculate mean and standard deviation of the performance of RL-recommended decisions within three subsamples: pre dot com, between dot com and GFC, and post GFC.

Figure 6: Empirical and Optimal Decisions under 1-QTR Market Cap Growth

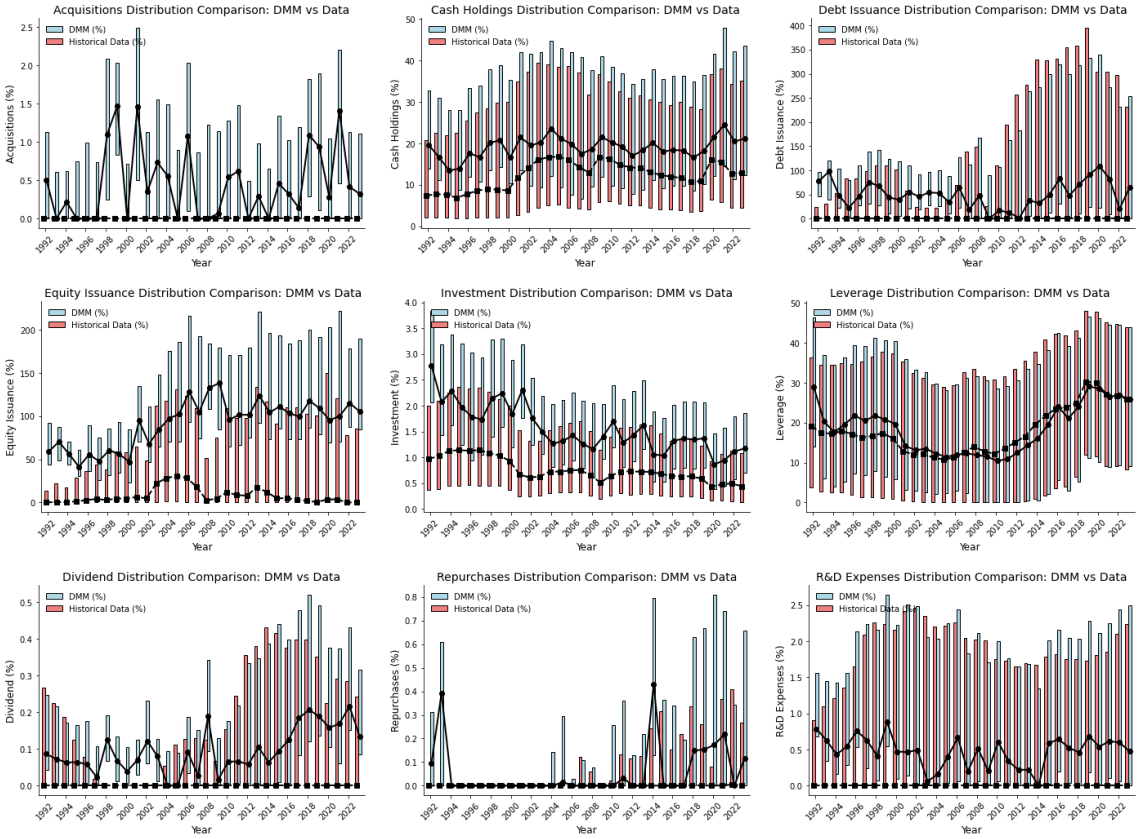


Note: This figure shows the empirical (red) and optimal decisions (blue) suggested by AM under 1-quarter market cap growth as the objective to maximize. We plot median to represent cross-sectional variation. The decisions are acquisitions (Faqc_fund), cash holdings (Fcash_hold), debt issuance (Fdebt_issue_log), equity issuance (Fequity_issue_log), investment ratio (Finv_ratio2), leverage (Flev_market), dividend (Fpayout_div), repurchase (Fpayout_repurchase), and R&D expenses (Frd_exp), with formal definitions detailed in Appendix A.

equity (with the latter being especially beneficial for small firms); and (4) reallocating investment by reducing R&D expenditures while increasing capital expenditures, dividend payouts, and share repurchases. Additionally, the model suggests a net reduction in leverage despite the slight increase in debt, as the equity issuance more than offsets it.

Finally, in Figure 9, for the objective of maximizing long-term enterprise value (over eight quarters), AlphaManager recommends pursuing more aggressive acquisition strategies while maintaining moderately higher cash reserves. The model suggests increasing both debt and equity issuance, resulting in slightly higher leverage during normal economic conditions but reduced leverage following the onset of the GFC. Additional recommendations include

Figure 7: Empirical and Optimal Decisions under 1-QTR Enterprise Value Growth



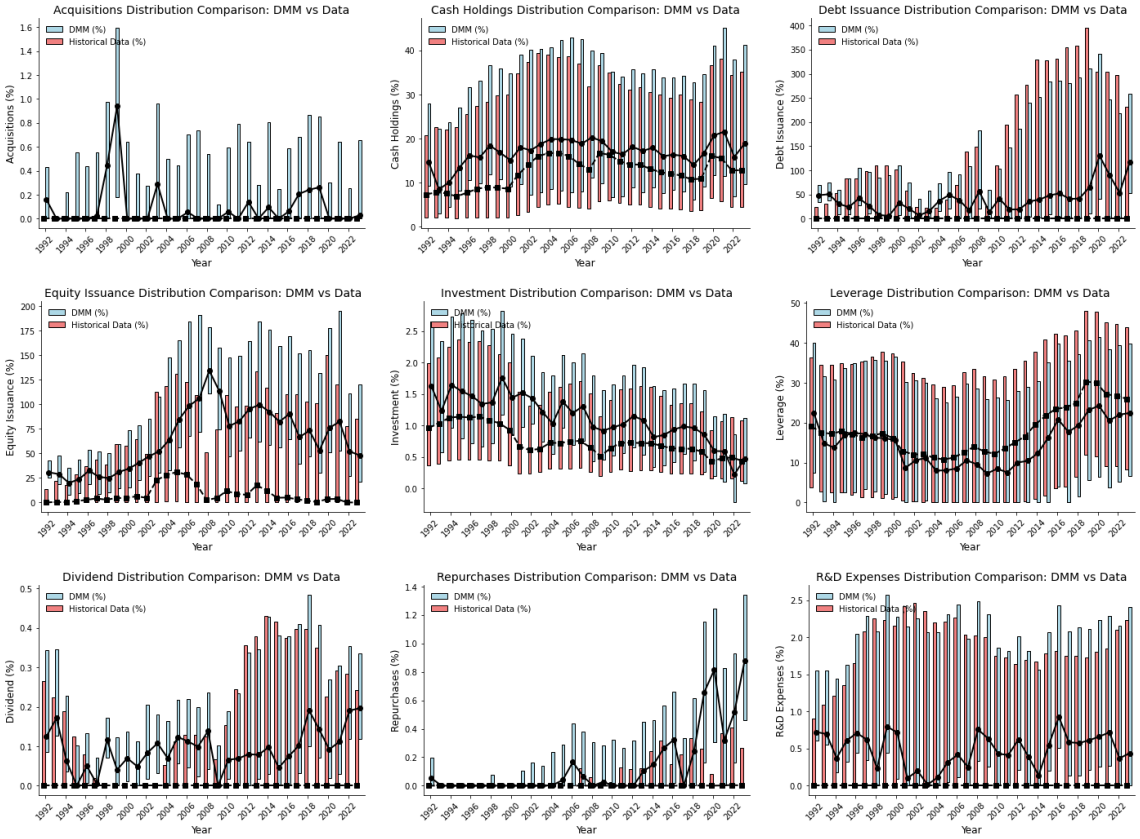
Note: This figure shows the empirical (red) and optimal decisions (blue) suggested by AM under 1-quarter enterprise value growth as the objective to maximize. We plot median to represent cross-sectional variation. The decisions are acquisitions (Faqc_fund), cash holdings (Fcash_hold), debt issuance (Fdebt_issue_log), equity issuance (Fequity_issue_log), investment ratio (Finv_ratio2), leverage (Flev_market), dividend (Fpayout_div), repurchase (Fpayout_repurchase), and R&D expenses (Frd_exp), with formal definitions detailed in Appendix A.

moderately decreasing dividend payouts, expanding share repurchase programs, and reducing R&D expenditures. These strategic adjustments collectively aim to optimize capital structure while reallocating resources toward growth-oriented activities.

7 New Research Questions in Corporate Finance

Our DDRC approach focuses more on producing reliable empirical predictions and policy recommendations. It does not serve as a substitute for theory, reduced-form models, or structural estimations, because it does not offer insights on the underlying economic mechanisms.

Figure 8: Empirical and Optimal Decisions under 8-QTR Market Cap Growth

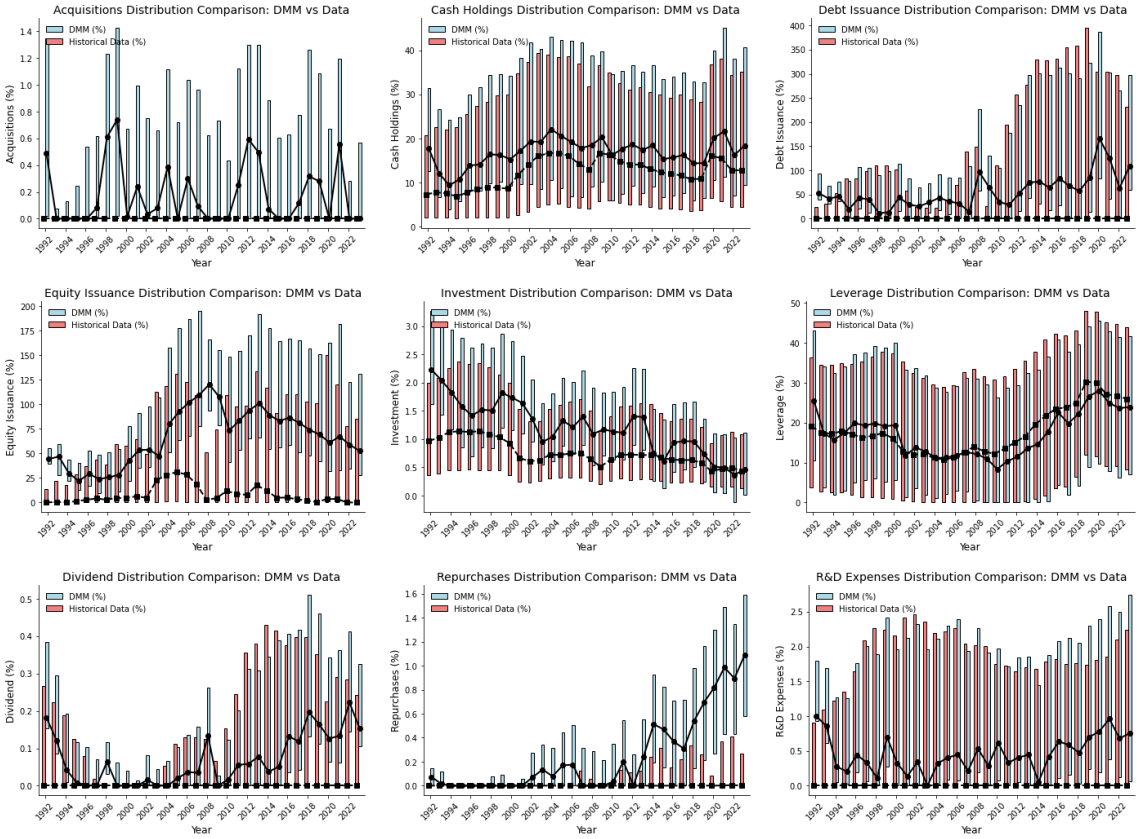


Note: This figure shows the empirical (red) and optimal decisions (blue) suggested by AM under 8-quarter market cap growth as the objective to maximize. We plot median to represent cross-sectional variation. The decisions are acquisitions (Faqc_fund), cash holdings (Fcash_hold), debt issuance (Fdebt_issue_log), equity issuance (Fequity_issue_log), investment ratio (Finv_ratio2), leverage (Flev_market), dividend (Fpayout_div), repurchase (Fpayout_repurchase), and R&D expenses (Frd_exp), with formal definitions detailed in Appendix A.

The associated cost of the many connections AlphaManager establishes in the data is that they can be computationally expensive in real time, are sometimes difficult to rationalize, lacking the usual notion of *ex-ante* theory predictions or causality statements.

However, both problems can be mitigated by incorporating theoretical guidance (multiple concurrent models). DDRC is still a useful approach for understanding the underlying economic mechanisms by complementing other paradigms in corporate finance research. In a way, DDRC is a data-driven version of the structural approach that allows the incorporation of knowledge from theory and reduced-form models, as well as from financial big data. To wit, DDRC is not at odds nor compete with reduced causal exercises and theory-driven

Figure 9: Empirical and Optimal Decisions under 8-QTR Enterprise Value Growth



Note: This figure shows the empirical (red) and optimal decisions (blue) suggested by AM under 8-quarter enterprise value growth as the objective to maximize. We plot median to represent cross-sectional variation. The decisions are acquisitions (Faqc_fund), cash holdings (Fcash_hold), debt issuance (Fdebt_issue_log), equity issuance (Fequity_issue_log), investment ratio (Finv_ratio2), leverage (Flev_market), dividend (Fpayout_div), repurchase (Fpayout_repurchase), and R&D expenses (Frd_exp), with formal definitions detailed in Appendix A.

structural estimations. Instead, they give us guidance as to when we should use those other approaches to data analysis.

We next discuss how the measure of model ambiguity in AlphaManager reveals the types of corporate decisions, firms, and market environment that would benefit more from theoretical guidance or causal identifications, as well as the types for which big data and powerful algorithms suffice. We then discuss how to incorporate the insights from other approaches into DDRC through ambiguity-guided transfer learning. Finally, we demonstrate how DDRC enables us to study managerial preferences.

7.1 Ambiguity and the Need for Theory/Reduced-Form Models

Our measure of ambiguity or relative entropy among the boosted PEMs informs how much we can rely on the data-driven approach. First, when the dispersion in predictions is high, one has to go back to theory or reduced-form models (with or without causal identification) in order to fully explain or predict firm outcomes, or to make counterfactual recommendations. Second, robust control makes DMM avoid exploring policies with more dispersed responses from PEMs. AlphaManager is thus conservative in the offline learning at the expense of missing out learning more profitable policies, the knowledge of which has to be derived from other conventional approaches. Figure 10 plots the estimated average ambiguity time series for each one of our state variables. For example, at the turns of macroeconomic regimes, ambiguity is estimated to be high.

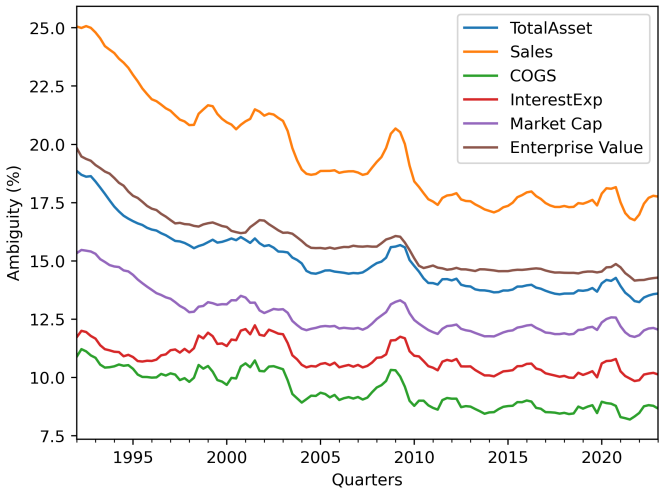
This apparent limitation of AlphaManager is its strength in disguise. The corporate finance literature has documented and studied a large collection of phenomena. Yet findings are typically siloed without a unified framework. DDRC, in addition to guiding managers in practice, provides insights about what areas in corporate finance research deserve more emphasis and attention going forward. This is so because along the dimensions that AlphaManager has greater ambiguity, theoretical insights and causal identification would be particularly informative and thus deserve more attention.

7.2 Ambiguity-Guided Transfer Learning

The identification of specific domains, firm characteristics, and macroeconomic conditions exhibiting heightened model ambiguity naturally leads to a fundamental methodological question: how can theoretical insights from established models be systematically integrated within the DDRC framework to reduce epistemic uncertainty? We address this challenge through ambiguity-guided transfer learning, whereby theoretical knowledge supplements empirical training data in regions where ensemble disagreement indicates insufficient observational density.

Transfer learning methodology leverages knowledge acquired from related analytical domains to enhance predictive performance in target applications, particularly when con-

Figure 10: Average Ambiguity Time Series by State Variables



Note: This figure shows average estimated ambiguity time series for key state variables—total assets, sales, COGS, interest expenses, market cap, and enterprise value. Ambiguity is calculated as follows: we first compute the square-rooted boosting error for each observation, aggregate these values at the state-quarter level, and then apply a 4-quarter moving average to the aggregated series.

fronting data limitations or computational constraints. Within corporate finance applications, this approach enables the systematic incorporation of theoretical predictions, empirical relationships derived from causal identification strategies, and counterfactual insights from structural models into the PEM training process.

Within our analytical framework, researchers seeking to incorporate theoretical predictions (e.g., Hennessy and Whited, 2005; Bolton et al., 2011), empirical relationships identified through reduced-form methodologies, or counterfactual predictions derived from structural estimations can leverage these insights to generate synthetic training data for integration into the PEM training sample. However, direct integration of historical data from reduced-form empirical studies proves methodologically inadequate. Causal identification strategies typically focus on localized treatment effects or firm-specific heterogeneity, producing insufficient observational variation for effective neural network training. Consequently, theoretical integration requires systematic generation of synthetic observations that extend established relationships across broader parameter ranges while maintaining theoretical consistency.

Traditional transfer learning applications typically involve adaptation to a single related

dataset. However, as previously discussed, corporate finance research encompasses numerous disparate insights across multiple theoretical and empirical domains. This raises critical questions regarding optimal integration strategies and appropriate boundary conditions for incorporating diverse knowledge sources. The ambiguity metric provides a principled approach to addressing these challenges by guiding the weighting scheme applied to different data sources within the transfer learning framework. Specifically, when PEM exhibits greater uncertainty or provides less precise characterization of particular patterns, researchers should systematically increase the relative weights assigned to data corresponding to those patterns within the transfer learning architecture.

In this paper, we demonstrate the practical implementation of theoretical knowledge integration within the DDRC framework by employing the investment-based model developed by Liu et al. (2009) as our example. This framework provides analytically tractable representations of corporate investment decisions under financial constraints and convex adjustment costs, making it particularly suitable for synthetic data generation. Given that market capitalization and enterprise value constitute primary optimization objectives in our empirical analysis, reducing prediction ambiguity for these variables enables more aggressive policy optimization across broader strategic domains.

Our implementation focuses on the model's core counterfactual prediction equation (i.e., equation (3) in Liu et al. (2009)) for next-period investment returns, $r^I_{i,t+1}$, which relates the marginal benefit of future investment to current marginal costs (with more details discussed in Appendix F):

$$r^I_{i,t+1} = \frac{(1-\tau_{i,t+1})[\alpha_i \frac{Y_{i,t+1}}{K_{i,t+1}} + \frac{a}{2}\left(\frac{I_{i,t+1}}{K_{i,t+1}}\right)^2] + \tau_{i,t+1}\cdot\delta_{i,t+1} + (1-\delta_{i,t+1})\cdot[1+(1-\tau_{i,t+1})a\frac{I_{i,t+1}}{K_{i,t+1}}]}{1+(1-\tau_{i,t}\cdot a(I_{i,t}/K_{i,t}))}, \tag{9}$$

where $\tau_{i,t+1}$ represents the effective tax rate, α_i denotes firm-specific total factor productivity, a captures adjustment cost parameters, $Y_{i,t+1}$ represents gross income, $K_{i,t+1}$ denotes the capital stock, $I_{i,t+1}$ represents investment flows, and $\delta_{i,t+1}$ indicates the depreciation rate.

Our transfer learning protocol integrates theoretical insights with empirical data through targeted PEM retraining. For each quarter, we load the baseline PEM trained exclusively

on historical data, augment the training sample with theoretically-derived synthetic observations, and conduct five additional training epochs to incorporate the expanded information set. This approach preserves the model's empirical foundation while systematically reducing ambiguity in regions where theoretical knowledge provides reliable guidance.

For each observation in our empirical dataset, we generate counterfactual scenarios by systematically varying investment levels while maintaining consistency with the theoretical framework. Specifically, we construct two alternative investment paths: one increasing actual investment by 0.1% and another decreasing actual investment by 0.1%. These perturbations generate sufficient variation to inform model training while remaining within economically plausible bounds.

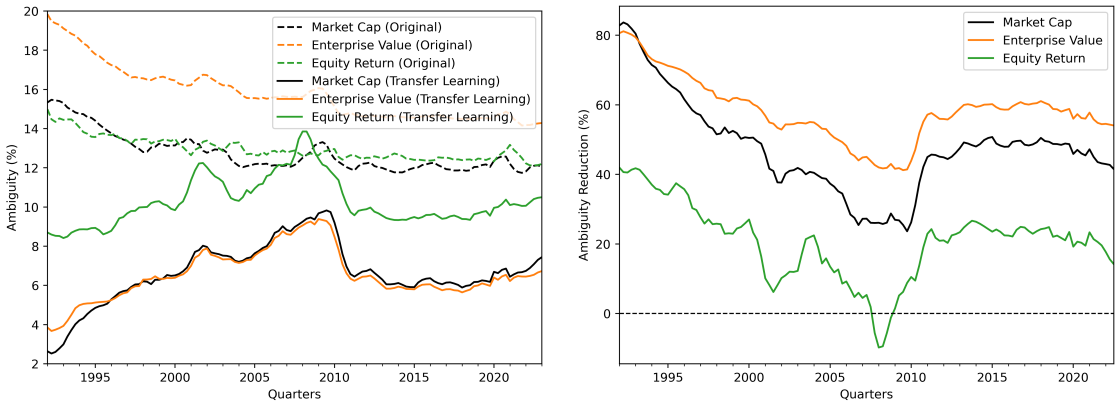
The counterfactual generation process proceeds through three sequential steps. First, we calculate implied investment returns using the Liu et al. (2009) equation for each investment scenario. Second, we derive corresponding equity returns through standard Weighted Average Cost of Capital (WACC) decomposition, maintaining consistency between investment profitability and equity valuation. Third, we compute counterfactual market capitalization and enterprise value evolution, holding all other state variables constant at their observed values.

Figure 11 presents comprehensive evidence of transfer learning effectiveness across our three primary state variables of interest. The left panel compares ambiguity levels between baseline PEM (dashed lines) and transfer learning-enhanced PEM (solid lines) for market capitalization, enterprise value, and equity returns. The right panel quantifies ambiguity reduction magnitudes achieved through theoretical integration.

The results demonstrate substantial and systematic ambiguity reduction across all target variables. Market capitalization ambiguity decreases by approximately 40% on average, with particularly pronounced improvements during periods of heightened market volatility. Enterprise value ambiguity exhibits similar patterns, with average reductions of 35-45% depending on macroeconomic conditions. Equity return predictions show the most dramatic improvement, with ambiguity reductions exceeding 60% in certain periods.

Temporal analysis reveals that transfer learning benefits vary systematically with mar-

Figure 11: Ambiguity after Transfer Learning



Note: This figure left panel shows the ambiguity for market cap, enterprise value, and equity return for the original PEM and the PEM after transfer learning. The right panel shows the ambiguity reduction for these three states after transfer learning.

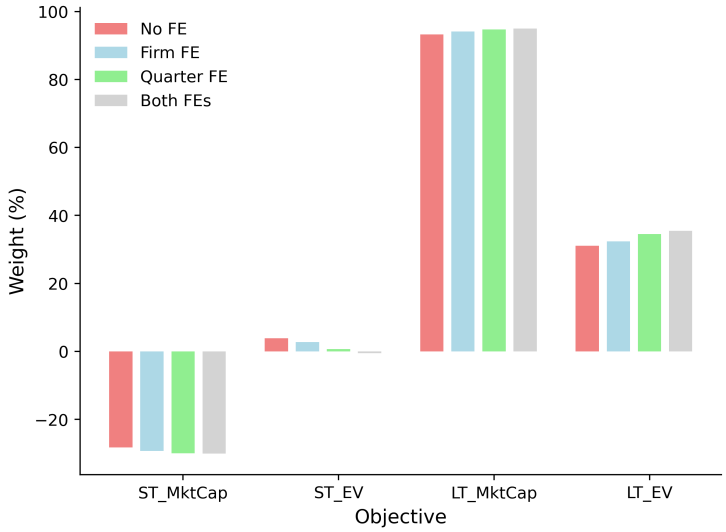
ket conditions. During relatively stable economic periods, theoretical augmentation provides modest improvements as empirical data prove sufficient for reliable prediction. However, during periods of structural change or unprecedented market conditions—such as the dot-com crisis or financial crisis periods—theoretical guidance becomes critically important, generating ambiguity reductions of 50-70%.

This implementation demonstrates the feasibility of systematic theoretical integration within data-driven frameworks while highlighting several important methodological considerations. First, the effectiveness of transfer learning depends critically on the quality and applicability of underlying theoretical models. The Liu et al. (2009) framework proves particularly suitable due to its empirical validation and parameter transparency, but integration of less well-established theories requires careful validation.

Second, our implementation represents a proof-of-concept demonstration rather than comprehensive implementation across multiple theoretical domains. Future research should systematically evaluate integration strategies across diverse theoretical frameworks, develop optimal weighting schemes for potentially conflicting theoretical predictions, and establish principled criteria for determining when theoretical augmentation justifies additional computational complexity.

Finally, while our results suggest substantial promise for ambiguity-guided theoretical

Figure 12: Revealed Weights on Four Objective Candidates



Note: This figure shows projected weights on four objective candidates (the growth of short-term market cap, short-term enterprise value, long-term market cap, and long-term enterprise value) in DMM training. We compare results without fixed effect, with firm fixed effect, with quarter fixed effect, and with both fixed effects.

integration, the current implementation focuses on a single theoretical model and limited set of outcome variables. Broader validation across diverse corporate finance applications remains necessary before claiming general applicability of this methodological approach.

7.3 Revealed Managerial Preferences

In our DMM, we derive optimal decisions by optimizing exogenous objective functions. However, empirical managers often consider additional factors beyond our model's specifications, such as ESG commitments or various forms of compensation. This divergence in objective functions naturally leads to differences between the theoretically optimal decisions suggested by our model and actual managerial decisions observed in practice.

Leveraging existing results from DMM with four different objectives considered above, a simple projection approach with cross-equation constraint helps us understand the weights for each objective which best fit the observed managerial decisions. Specifically, we solve the

following constrained optimizing problem:

$$\begin{aligned} \min_{\beta_k} \quad & \left(\sum_{k=1}^{4} \beta_k \cdot u_{j,t,k} - u^*_{j,t} \right)^2 \\ \text{s.t.} \quad & \sum_{k=1}^{4} \beta_k = 1, \end{aligned}$$

where $u^*_{j,t}$ is the empirical decisions made by firm j at time t, while $u_{j,t,k}$ is the suggested decisions by DMM under the k-th objective. This constrained optimization problem can be viewed as a stacked least square exercise with cross-equation constraint, as the parameter β_k is common across all dimensions of managerial decisions for each objective function. We also consider adding firm and/or quarter FEs to absorb firm and/or quarter specific unobserved features that could potentially affect corporate decisions. Empirically, we subtract an optimal decision from a given objective from the empirical as well as other three optimal decisions, estimate the stacked regression model with and without FE(s), and back out the β for the subtracted objective. Figure 12 shows the empirical results. The empirical evidence is consistent with a representative long-term manager focusing on maximizing long-term market cap growth (and some long-term enterprise value) who is willing to sacrifice short-term market cap growth. However, there is little variation in R^2s in these four specifications, ranging from 10.04% to 10.68%, indicating nontrivial potential incentive misalignment to these four objective candidates.

8 Concluding Remarks

For any given managerial objective, corporate decision-making constitutes a high-dimensional, nonlinear stochastic control problem where managers learn about and interact to the evolving economic environment while formulating optimal dynamic policies. Conventional approaches in corporate finance and stochastic control often fall short of explaining or predicting empirical firm outcomes and have limited practical adoption. We introduce an AI-assisted, data-driven framework that prioritizes empirical performance and the generation of the useful

counterfactuals to deliver effective policy recommendations for diverse business objectives.

Our approach first constructs a predictive environment module using supervised deep learning with Transformers and then integrates a decision-making module based on deep reinforcement learning. This dual-module structure goes beyond mere hypothesis testing on historical data or simulations by incorporating model ambiguity and robust control techniques. As a result, our framework not only improves in-sample and out-of-sample prediction of corporate outcomes but also identifies critical managerial decisions and offers dynamic policies that adapt to market evolution and feedback. Moreover, it distinguishes scenarios where theoretical insights and causal identification are essential from those where data-driven predictive models suffice. The framework's flexibility is further enhanced by its ability to incorporate insights from theory, reduced-form, and structural models through ambiguity-guided transfer learning, making it a promising unified approach for corporate finance research.

We believe that the DDRC approach opens the door to investigating a range of important, yet previously unexplored, topics. For instance, while we assume a given managerial objective in this paper, techniques exist to learn historical managerial objectives from data. Additionally, our framework provides new insights into the heterogeneity of managerial actions, firm controlability, and their interactions with macroeconomic conditions – areas that warrant further study.

References

Almeida, Heitor, Nuri Ersahin, Vyacheslav Fos, Rustom M Irani, and Mathias Kronlund, "How Do Short-Term Incentives Affect Long-Term Productivity?," *The Review of Financial Studies*, 2024, p. hhae064.

Andersen, Torben G, Nicola Fusari, and Viktor Todorov, "Parametric inference and dynamic state recovery from option panels," *Econometrica*, 2015, *83* (3), 1081–1145.

Barnett, Michael, William Brock, and Lars Peter Hansen, "Pricing uncertainty induced by climate change," *The Review of Financial Studies*, 2020, *33* (3), 1024–1066.

Bellstam, Gustaf, Sanjai Bhagat, and J Anthony Cookson, "A text-based analysis of corporate innovation," *Management Science*, 2020.

Bolton, Patrick, Hui Chen, and Neng Wang, "A unified theory of Tobin's q, corporate investment, financing, and risk management," *The journal of Finance*, 2011, *66* (5), 1545–1578.

Bond, Philip, Alex Edmans, and Itay Goldstein, "The real effects of financial markets," *Annu. Rev. Financ. Econ.*, 2012, *4* (1), 339–360.

_ , Itay Goldstein, and Edward Simpson Prescott, "Market-based corrective actions," *The Review of Financial Studies*, 2010, *23* (2), 781–820.

Campello, Murillo and Gaurav Kankanhalli, "Corporate decision-making under uncertainty: review and future research directions," *Handbook of Corporate Finance*, 2024, pp. 548–590.

_ , _ , and Pradeep Muthukrishnan, "Corporate hiring under Covid-19: Financial constraints and the nature of new jobs," *Journal of Financial and Quantitative Analysis*, 2024, *Forthcoming.*

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L Zhang, "How to talk when a machine is listening: Corporate disclosure in the age of AI," *The Review of Financial Studies*, 2023, *36* (9), 3603–3642.

_ , _ , Junbo L Wang, and Baozhong Yang, "From man vs. machine to man+ machine: The art and AI of stock analyses," *Columbia Business School Research Paper*, 2021.

Chen, Hui, Yuhan Cheng, Yanchu Liu, and Ke Tang, "Teaching Economics to the Machines," *Available at SSRN 4642167*, 2023.

Cong, Lin William, Guanhao Feng, Jingyu He, and Junye Li, "Sparse Modeling Under Grouped Heterogeneity with an Application to Asset Pricing," 2023.

_ , _ , _ , and Xin He, "Growing the Efficient Frontier on Panel Trees," *NBER Working Paper*, 2022, (w30805).

_ , Ke Tang, Jingyuan Wang, and Yang Zhang, "AlphaPortfolio: Direct Construction Through Reinforcement Learning and Interpretable AI," *Working Paper*, 2020.

_ , _ , _ , and _ , "Deep Sequence Modeling: Development and Applications in Asset Pricing," *The Journal of Financial Data Science*, 2021, *3* (1), 28–42.

_ , Tengyuan Liang, and Xiao Zhang, "Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information," Technical Report, resubmission requested. 2018.

Dicks, David and Paolo Fulghieri, "Uncertainty, investor sentiment, and innovation," *The Review of Financial Studies*, 2021, *34* (3), 1236–1279.

Dicks, David L and Paolo Fulghieri, "Uncertainty aversion and systemic risk," *Journal of Political Economy*, 2019, *127* (3), 1118–1155.

Ebert, Frederik, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

Edmans, Alex, Itay Goldstein, and Wei Jiang, "The real effects of financial markets: The impact of prices on takeovers," *The Journal of Finance*, 2012, *67* (3), 933–971.

_ , _ , **and** _ , "Feedback effects, asymmetric trading, and the limits to arbitrage," *American Economic Review*, 2015, *105* (12), 3766–3797.

Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach, "Selecting directors using machine learning," Technical Report, National Bureau of Economic Research 2018.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, "Taming the Factor Zoo: A Test of New Factors," *The Journal of Finance*, 2020, *75* (3), 1327–1370.

Fu, Justin, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine, "D4rl: Datasets for deep data-driven reinforcement learning," *arXiv preprint arXiv:2004.07219*, 2020.

Fujimoto, Scott, David Meger, and Doina Precup, "Off-policy deep reinforcement learning without exploration," in "International Conference on Machine Learning" PMLR 2019, pp. 2052–2062.

Garlappi, Lorenzo, Ron Giammarino, and Ali Lazrak, "Ambiguity and the corporation: Group disagreement and underinvestment," *Journal of Financial Economics*, 2017, *125* (3), 417–433.

Graham, John R, "Presidential address: Corporate finance and reality," *The Journal of Finance*, 2022, *77* (4), 1975–2049.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Hanley, Kathleen Weiss and Gerard Hoberg, "Dynamic interpretation of emerging risks in the financial sector," *The Review of Financial Studies*, 2019, *32* (12), 4543–4603.

Hansen, Lars Peter and Thomas J Sargent, "Robust control and model uncertainty," *American Economic Review*, 2001, *91* (2), 60–66.

_ **and Thomas J. Sargent**, "Risk, ambiguity, and misspecification: Decision theory, robust control, and statistics," *Journal of Applied Econometrics*, 2023, *n/a* (n/a).

Hennessy, Christopher A and Toni M Whited, "Debt dynamics," *The journal of finance*, 2005, *60* (3), 1129–1165.

Holmstrom, Bengt and Jean Tirole, "Financial intermediation, loanable funds, and the real sector," *the Quarterly Journal of economics*, 1997, *112* (3), 663–691.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, 1989, *2* (5), 359–366.

Izhakian, Yehuda, David Yermack, and Jaime F Zender, "Ambiguity and the tradeoff theory of capital structure," *Management Science*, 2022, *68* (6), 4090–4111.

Jaques, Natasha, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard, "Way off-policy batch deep reinforcement learning of implicit human preferences in dialog," *arXiv preprint arXiv:1907.00456*, 2019.

Jarrow, Robert and Simon Sai Man Kwok, "Specification tests of calibrated option pricing models," *Journal of Econometrics*, 2015, *189* (2), 397–414.

Jin, Ying, Zhuoran Yang, and Zhaoran Wang, "Is pessimism provably efficient for offline rl?," in "International Conference on Machine Learning" PMLR 2021, pp. 5084–5096.

Kahn, Gregory, Pieter Abbeel, and Sergey Levine, "Badgr: An autonomous self-supervised learning-based navigation system," *IEEE Robotics and Automation Letters*, 2021, *6* (2), 1312–1319.

Kalashnikov, Dmitry, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke et al., "Scalable deep reinforcement learning for vision-based robotic manipulation," in "Conference on Robot Learning" PMLR 2018, pp. 651–673.

Kaniel, Ron, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh, "Machine-learning the skill of mutual fund managers," *Journal of Financial Economics*, 2023, *150* (1), 94–138.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, "The virtue of complexity in return prediction," *The Journal of Finance*, 2024, *79* (1), 459–503.

Kidambi, Rahul, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims, "MOReL: Model-based offline reinforcement learning," in "NeuIPS 2020" 2020.

Kingma, Diederik P and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan, "Measuring corporate culture using machine learning," *The Review of Financial Studies*, 2020.

Liu, Laura Xiaolei, Toni M Whited, and Lu Zhang, "Investment-based expected stock returns," *Journal of Political Economy*, 2009, *117* (6), 1105–1139.

Malenko, Andrey and Anton Tsoy, "Asymmetric information and security design under Knightian uncertainty," *Available at SSRN 3100285*, 2020.

Mamaysky, Harry, Matthew Spiegel, and Hong Zhang, "Improved forecasting of mutual fund alphas and betas," *Review of Finance*, 2007, *11* (3), 359–400.

Mitton, Todd, "Methodological variation in empirical corporate finance," *The Review of Financial Studies*, 2022, *35* (2), 527–575.

Pan, Jun, "The jump-risk premia implicit in options: Evidence from an integrated time-series study," *Journal of financial economics*, 2002, *63* (1), 3–50.

Spiegel, Matthew, "For corporate finance to truly advance we need more genuinely testable models," *Financial Review*, 2023.

Sutton, Richard S and Andrew G Barto, *Introduction to reinforcement learning*, Vol. 135, MIT press Cambridge, 1998.

Wang, Zhenyu, "A shrinkage approach to model uncertainty and asset allocation," *Review of Financial Studies*, 2005, pp. 673–705.

Yu, Tianhe, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma, "Mopo: Model-based offline policy optimization," *arXiv preprint arXiv:2005.13239*, 2020.

Appendix A - Variable Selection and Definitions

Table 12 shows how we construct our firm-specific state and decision variables using Compustat and CRSP.

Table 12: Variable Selection and Definitions

variable	explanation	category	formula	source
lev	leverage	Decision	(dtllq + dlcq) / L.atq	Compustat
aqc	acquisition	Decision	aqcq / L.atq	Compustat
inv_ratio	investment	Decision	capxq / L.atq	Compustat
cash_hold	cash holding	Decision	cheq / L.atq	Compustat
div_payout	dividend	Decision	devq / L.atq	Compustat
log_debt_issue	debt issuance	Decision	log(1 + dltisq)	Compustat
log_eq_issue	equity issuance	Decision	log(1 + sstkq)	Compustat
rd_exp	R&D expenses	Decision	xrdq / L.atq	Compustat
repurchase	repurchases	Decision	prstkcq / L.atq	Compustat
log_at	book asset	State	log(1 + atq)	Compustat
log_act	current asset	State	log(1 + actq)	Compustat
log_sale	gross revenue	State	log(1 + saleq)	Compustat
log_ap	accounts Payables	State	log(1 + apq)	Compustat
log_cogs	cost of good sold	State	log(1 + cogsq)	Compustat
log_intpn	net interest paid	State	log(1 + intpnq)	Compustat
log_invt	inventories	State	log(1 + invtq)	Compustat
log_lct	current liabilities	State	log(1 + lctq)	Compustat
log_rect	Receivabless	State	log(1 + rectq)	Compustat
log_revt	net revenue	State	log(1 + revtq)	Compustat
log_market_cap	market cap	State	log(1 + csho * prccq)	Compustat
log_enterprise_val	enterprise value	State	log(1 + atq + csho * prccq - ceqq - txdbq)	Compustat
log_vol	trading volume	State	log(1 + VOL)	CRSP
log_ret	equity return	State	log(1 + RET)	CRSP
macro1	risk	Macro	risk	ChicagoFed
macro2	credit	Macro	credit	ChicagoFed
macro3	leverage	Macro	leverage	ChicagoFed
macro4	non-financial leverage	Macro	nonfinancial_leverage	ChicagoFed

Appendix B - Technical Details

B.1 Predictive Environment Module Hyper-Parameters

Table 13: Predictive Environment Module: Neural Network Hyperparameters

Hyperparameter	Value (main)	Value (macro)
optimizer	Adam	Adam
learning rate	0.000003	0.0005
batch size	2048	full
epoch (first)	40	50
epoch (rolling)	12	3
depth (main & aux)	4	2
neurons (main)	(512, 512, 512, 512)	(300, 300)
aux network num	10	-
dropout	0.3	-
l2 norm	0.00001	0.01

B.2 Decision-making Module Hyperparameters

Table 14: Decision-making Module: Neural Network Hyperparameters

Hyperparameter	Value
optimizer	Adam
periods in obj. func.	{1, 8}
learning rate	0.00003
batch size	full
epoch	64
depth	4
neurons	(256, 256, 256, 256)

Appendix C - Ambiguity Constraint Parameters

We first estimate parameters in the ambiguity constraints, namely α and β in Eq (6), for the long-term models. We follow an iterative process to search for proper constraints that could give out positive and reasonable long-term performance. For each cross section, we start from a tight lower-bound for these parameters, with 0 intercept and 1 slope. Essentially, this constraint requires AM managerial decisions as certain as the last set of observable decisions (the benchmark). Instead of search for optimal decisions which maximize long-term firm values, AM in this stage tries to minimize ambiguity of suggested managerial decisions. After 10 training epochs, we use a dynamic programming algorithm to estimate α and β so that the constraint never binds and the area under the curve is minimized. In this way, the ambiguity punishment for the current AM equals to 0, and the ambiguity constraint is tight enough that AM performance is not likely to be irrationally high. After training for 5 more epochs, we examine whether the reward after ambiguity punishment is positive on average. If yes, we continue to train AM for 5 epochs and move on to the next cross section; if not, the current ambiguity constraint is likely too tight, so that AM is too ambiguity-averse and does not have sufficient incentive to explore the environment and reach for higher firm value. In this case, we re-estimate α and β to loose the ambiguity constraint again and try to train the model for 5 more epochs, until it has a positive reward after ambiguity punishment.

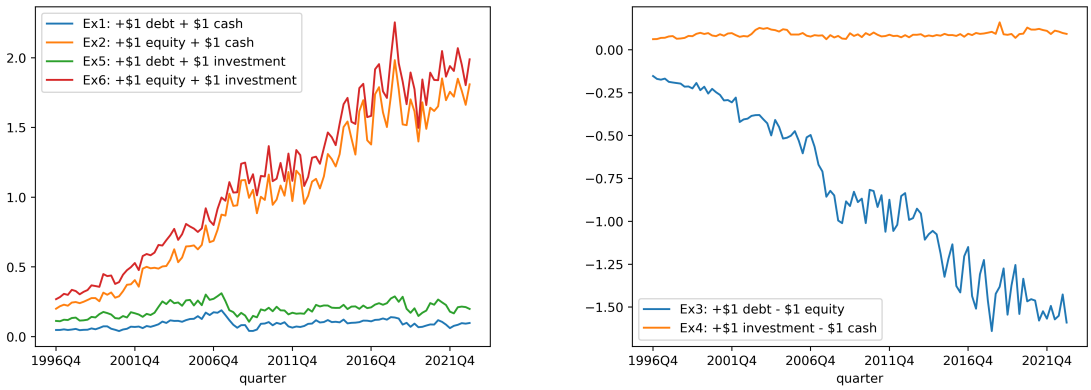
We then estimate α and β for the short-term models. To make short-term models comparable to the corresponding long-term models in each cross section, we first calculate the long-term AM's short-term boosting error, and then we use the dynamic programming algorithm to estimate α and β. Throughout the training of the short-term AM for that cross section, we stick with this ambiguity constraint and try to optimize short-term objective. The short-term AM is expected to behave better than the long-term AM, because under the long-term AM, the suggestion managerial decisions fall inside the constraint, while short-term objective values are suboptimal.

Appendix D - An Illustration of Firm Recapitalization

Even though PEM is high-dimensional, we can analyze low dimensional action combinations, which are the focus of conventional reduced-form models. Firm recapitalization that shows up in every corporate finance textbook serves as an excellent illustration. Specifically, we ask how the enterprise value changes when a manager (firm):

(1) raises $1 in debt and put that $1 into its cash savings
(2) raises $1 in equity and put that $1 into its cash savings
(3) raises $1 in debt and $1 less equity
(4) puts $1 cash into investment
(5) raises $1 in debt and put that $1 into investment
(6) raises $1 in equity and put that $1 into investment

Figure 13: Recapitalization Exercises: Capital Structure and Financing Investments



Note: This figure shows the recapitalization exercises where counterfactuals are generated by PEM. The x-axis is the calendar year, and the y-axis is dollar change in enterprise value. Solid lines plot the median value of the change in enterprise value for each cross section, and their legends explain the content of the recapitalization exercises.

Exercises 1–3 primarily address capital structure adjustments, whereas Exercises 4–6 focus on financing investments. Overall, Exercise 6 emerges as the optimal recapitalization strategy for increasing enterprise value. Notably, the dispersion in performance among the six strategies has grown over time.

We observe several patterns. First, Exercise 4 corresponds to the return on asset investment and is only marginally above zero. Second, in Exercise 3, altering the capital structure while keeping enterprise value constant results in a reduction in enterprise value — suggesting that firms generally carry excessive leverage and should decrease their debt proportion.

Third, Exercises 1, 2, 5, and 6 show that debt financing is less effective in boosting enterprise value compared to equity financing.

Appendix E - Detailed Results on PEM for Macroeconomic Predictions

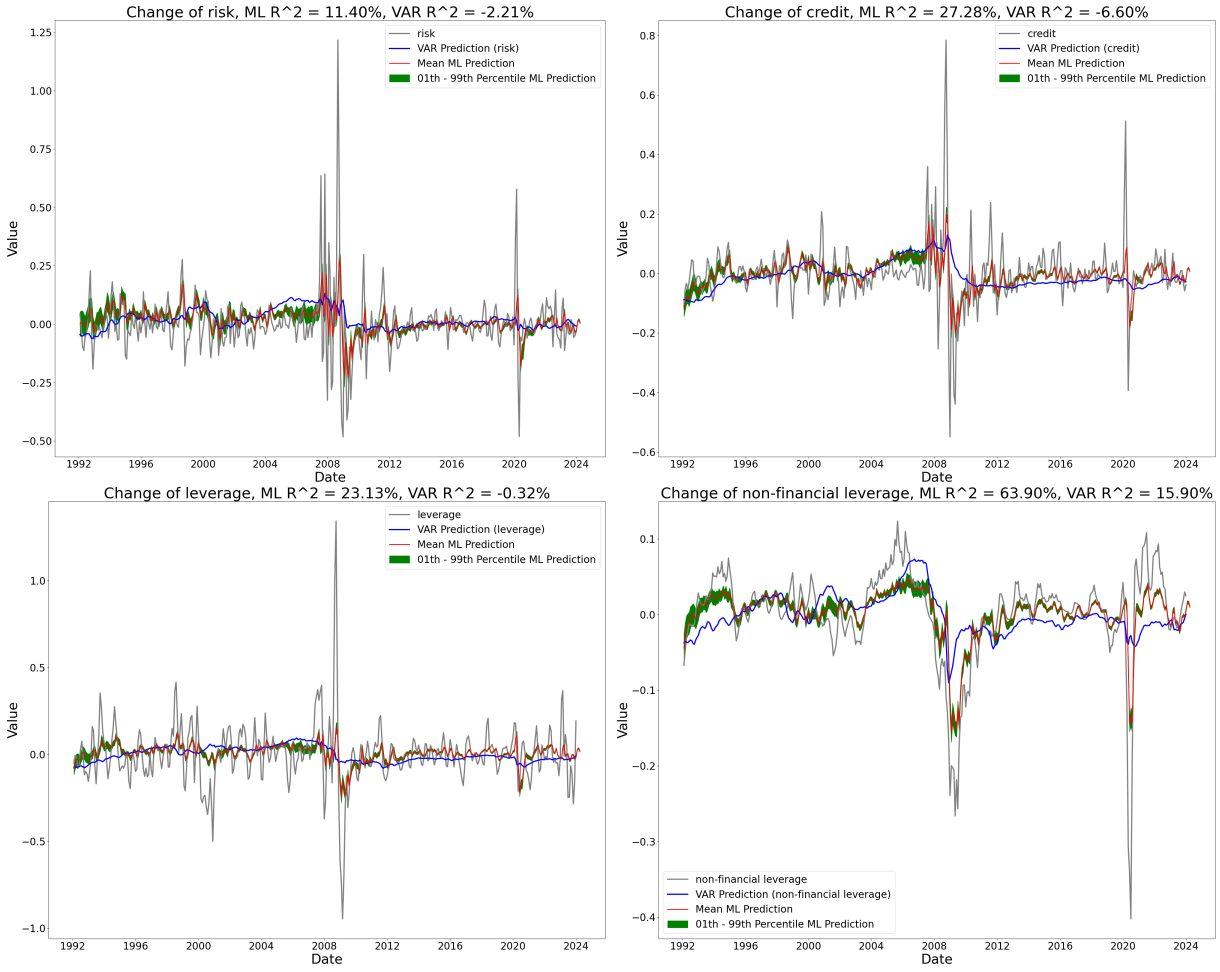


Figure 14: Predicting macroeconomic variables using neural network and VAR

Figure 14 shows the predicted change in next month for macroeconomic state variables using current and one-month lagged macroeconomic states generated by a neural network model and a VAR model. Neural networks, with their capability to handle collinear variables, are enhanced by adding current growth rates of macroeconomic state variables as additional inputs, along with their historical data. VAR models are generally used in time series modelling, which are widely used to model macroeconomic variables. In our setting, the neural network model is comparable to the VAR model because they are using the same information set as well as the same training paradigm when predicting the macroeconomic growth. Consistent with PEM, we update the VAR model every year using all available

observations till that year. We find that using simple VAR models to predict these four time series generates much lower (or even negative) out-of- sample pseudo R^2 than PEM.

Risk Prediction The neural network achieves an R^2 of 11.40% when predicting changes in risk, demonstrating moderate predictive power. The model captures significant risk fluctuations, particularly during periods of financial distress like the 2008 crisis, though it still leaves some noise in the actual values. On the other hand, the VAR model performs poorly, with a negative R^2 of -2.21%, indicating that its predictions are worse than simply using the mean of the data. The VAR model struggles to capture key peaks and troughs, especially in volatile periods.

Credit Prediction The neural network fares better in predicting credit changes, achieving an R^2 of 27.28%, with the model capturing general trends during financial disruptions. The VAR model, however, performs significantly worse with an R^2 of -6.59%, highlighting its inability to accurately model credit risk in this context.

Leverage Prediction For leverage, the neural network attains an R^2 of 23.13%, demonstrating reasonable predictive capability. The model captures key variations in leverage, especially during volatile periods like 2008, though it struggles during calmer periods. The VAR model continues to underperform, with an R^2 of -0.32%, reflecting its failure to model significant leverage shifts accurately.

Non-Financial Leverage Prediction The neural network excels in predicting non-financial leverage, achieving an R^2 of 63.90%, the highest across all variables. The model effectively tracks the changes in non-financial leverage across economic expansions and contractions, particularly during the financial crisis and the COVID-19 pandemic. The VAR model performs better than for other variables, achieving a positive R^2 of 15.90%, though it still significantly underperforms relative to the neural network.

Overall, the neural network model consistently outperforms the VAR model in predicting macroeconomic conditions across all variables. The VAR model, while traditionally used for macroeconomic forecasting, struggles to capture the complex non-linear relationships inherent in these time series data, especially during periods of economic volatility. The neural network, with its ability to process both current and historical growth rates and

handle collinear variables, demonstrates a clear advantage in forecasting macroeconomic variables, particularly non-financial leverage.

Appendix F - Details on Transfer Learning and Liu et al. (2009)

where $\tau_{i,t+1}$ is the effective tax rate for company i, α_i is the total factor productivity for company i, a is the adjustment cost parameter, $Y_{i,t+1}$ is the gross income for company i in the next period, $K_{i,t+1}$ is the fixed asset for company i in the next period, $I_{i,t+1}$ is the investment for company i in the next period,