Austrian Research Institue for Artificial Intelligence
Österreichisches Forschungsinstitut für Artificial Intelligence
http://www.oefai.at/oefai/oefai.html

Freyung 6/6, Vienna
A-1010, Austria
(+43-1) 5336112-0

ÖFAI Neural Computation Group Technical Report

December 15, 2003

OEFAI–TR–2003–33

A Simulation Study of Managerial Compensation

Brian Sallans
ÖFAI Neural Computation Group

Alexander Pfister
Vienna University of Economics and Business Administration

Georg Dorffner
ÖEFAI Neural Computation Group

Abstract

A computational economics model of managerial compensation is presented. Risk-averse managers are simulated, and shown to adopt more risk-taking under the influence of stock options. It is also shown that stock options can both help a new entrant compete in an established market; and can help the incumbent firm fight off competition by promoting new exploration and risk-taking. In the case of the incumbent, the stock options are shown to be most effective when introduced as a response to the arrival of a new entrant, rather than used as a standard part of the compensation package.

A Simulation Study of Managerial Compensation

Brian Sallans
ÖFAI Neural Computation Group

Alexander Pfister
Vienna University of Economics and Business Administration

Georg Dorffner
ÖEFAI Neural Computation Group

1 Introduction

Designing compensation for managers is an important aspect of firm governance and management. Understanding compensation contracts is therefore of great interest, both to owners and managers of a firm. Specifically, we focus on the role of stock options as part of a compensation contract.

Theoretical models of compensation address the problem of how to craft contracts which maximize firm value. Principal-Agency models have achieved some prominence as a model of contracts, including those between managers and owners of a company [Holmström 1979]. The key feature of principal-agency models is that the principal is limited in terms of the observability of the agent's actions, effort, results, or preferences. The principal therefore delegates decision making to the agent. Alternatively, compensation models can focus on how contracts or instruments are valued by managers as opposed to unrestricted traders (see for example Hall and Murphy [2002]).

Typically, the manager in a theoretical model of compensation is assumed to be a fully rational economic actor. She has complete knowledge and understands the consequences of her actions, at least probabilistically. She can therefore assess the riskiness of alternative actions, and weigh this against possible gains in her compensation. This assumption results in analytic models for which the contract can be found which optimizes the gain by the firm's owners [Bushman and Indjejikian 1993; Baiman and Verrecchia 1995; Choe 1998].

One limitation of principal-agency models is their restriction to rational agents. Empirical work seeks to address this limitation by studying the influence of compensation in actual firms. However, real firms are complicated. It is difficult to control for all variables in an empirical study. Empirical studies of compensation have led to conflicting and inconclusive results [Murphy 1999].

Beginning with the work of Herbert Simon [1982], there has been a realization that classical economic theory, based on full rationality, is limited. In reality, economic agents are bounded both in their knowledge and in their computational abilities. Recent work in simulation-based computational economics has sought to implement boundedly rational economic actors as learning agents, and to study the implications on the resultant economic systems. See Tesfatsion [2002] for a review of agent-based computational economics.

Computational economic models bridge the gap between theoretical and empirical economics. On one hand, a computational model can be used to test the predictions of theory under conditions which are too complex to be addressed analytically. On the other hand, computational models can be used to give insight into complex systems and suggest new hypotheses to be tested in empirical studies. Computational models offer an environment which is complex but controlled, where all assumptions are explicitly encoded in the model.

In this article we study management compensation using a discrete-time agent-based economic model. The model spans two markets: a consumer market and a financial equities market. The consumer market consists of production firms offering a good for sale, and customers who can purchase the good. The financial equities market consists of stock traders who can buy and sell shares in the production firms. This model is called the integrated markets model (IMM).

The IMM is intended to be a generic model of the interaction between financial and consumer markets. It has been shown to reproduce a large range of empirical "stylized facts", including learning-by-doing in the consumer market; low predictability, high kurtosis and volatility clustering in the financial market; and correlations between volatility and trading volume in the financial market. See Sallans, Pfister, Karatzoglou, and Dorffner [2003] for a detailed description of the model and experimental simulation and validation results. It has been used previously to simulate the influence of financial traders' expectations on the behavior of managers [Sallans, Dorffner, and Karatzoglou 2002].

The managers of the firms in the IMM try to optimize their own compensation. Depending on their contract, they might do this by increasing profits, or by taking actions which they predict will more directly boost the value of their stock-based compensation. Thus, as in principal-agency models, the problem of moral hazard still exists in the IMM. However, each manager explicitly implements a boundedly-rational agent which learns from experience, and itself has limited knowledge and computational power. We supplement the classical principal-agency framework by considering issues of limited knowledge, learning from experience, exploration versus exploitation of existing knowledge, and asymmetry between different sources of information for the decision-making agent. We use artificial intelligence techniques to implement learning and decision-making in the firm. Specifically, we use a reinforcment learning technique. Although reinforcement learning has its origins in computer science and optimization, it has been shown to be both a good descriptive and predictive model of human decision making in competitive "game-like" situations [Erev and Roth 1998; Flache and Macy 2002]. In particular, reinforcment learning models are good models of human decision making when the learner uses both limited memory and stochastic experimentation. We incorporate both of these features into our model. Further, competitive markets in which reinforcement learning is used has been shown to exhibit many phenomena seen in real markets, such as price wars, implicit price collusion, and business cycles [Tesauro 1999].

Under the IMM, the role of the compensation contract is somewhat different than under a principal-agency framework. The manager does not start with intimate knowledge of the consumer or financial market. Rather, it must learn what the markets want through experience. The manager receives feedback in terms of profits and movements in stock price. These two measures give the manager two different views of firm performance. Because these two estimators are generated by two different populations of boundedly-rational agents (consumers and stock traders respectively), they do not necessarily agree.

One appeal of using the IMM to investigate compensation is that it was not explicitly built for this purpose. It is a system consisting of firms, consumers and stock traders. In the IMM there is no explicit link between, for example, stock option value and risk-taking. A priori, it is not clear to what degree a manager can or will influence the volatility of their firm's stock. Any link between granting stock options and risk-taking is not built in to the model. If the predictions of principal-agency theory are born out, it will demonstrate that these predictions are quite robust to the typical assumptions of analytic models, such as rationality and prior knowledge.

In this paper we use the IMM first to test predictions of theoretical models of contracts, and second to generate new hypotheses that can be tested empirically. Specifically, we test the effect of stock options on managerial learning and behavior. In the first set of experiments, we test the effect on risk-taking and performance of stock option grants. The goal is to see how robust theoretical predictions are to the situation of incomplete knowledge and learning; and to confirm that the model of learning and decision making in the firm acts in a reasonable way. As a byproduct, we will suggest alternative mechanisms by which these "risk-taking" effects might occur. In the second set of experiments, we investigate the effect of option-granting on learning and new knowledge acquisition.

Our results will be of interest both to those interested in theoretical and empirical studies of compensation; and to decision makers responsible for crafting compensation contracts. In particular, we supplement theoretical models of compensation by suggesting new mechanisms which might promote enhanced risk-taking and improve firm performance, and suggest when it might be most advantageous to include stock options in compensation contracts.

For completeness, we give a description of the basic model in the next sections. This is followed

by a description of the different compensation schemes examined, and simulation results. We conclude with a discussion of the simulation results and their implications on managerial compensation.

2 The Integrated Markets Model

In this section we give a description of the integrated markets model. The reader is directed to Sallans, Pfister, Karatzoglou, and Dorffner [2003] for a detailed description of the model and validation results.

The model consists of two markets: a consumer market and a financial equities market. The consumer market simulates the manufacture of a product by *production firms*, and the purchase of the product by *consumers*. The financial market simulates trading of shares. The shares are traded by *financial traders*. The two markets are coupled: The financial traders buy and sell shares in the production firms, and the managers of firms may be concerned with their share price. The traders can use the performance of a firm in the consumer market in order to make trading decisions. Similarly, the production firms can potentially use positioning in product space and pricing to influence the decisions of financial traders (see figure 1).

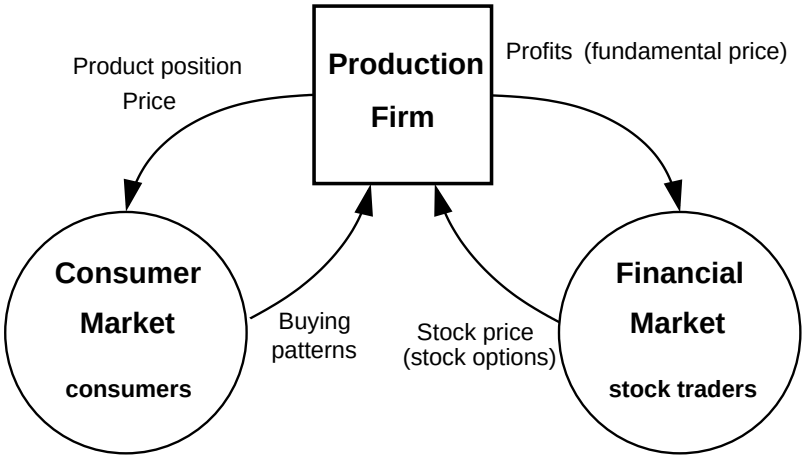

Figure 1: The Integrated Markets Model. Consumers purchase products, and financial traders trade shares. Production firms link the consumer and financial markets, by selling products to consumers and offering their shares in the financial market.

The simulator runs in discrete time steps. Simulation steps consist of the following operations:

1. Consumers make purchase decisions.
2. Firms receive an income based on their sales and their position in product space.
3. Financial traders make buy/hold/sell decisions. Share prices are set and the market is cleared.
4. Every N_p steps, production firms update their products or pricing policies based on performance in previous iterations.

We describe the details of the markets, and how they interact, in the following sections.

3 The Consumer Market

The consumer market consists of firms which manufacture products, and consumers who purchase them. The model is meant to simulate production and purchase of non-durable goods, which the consumers will re-purchase at regular intervals. The product space is represented as a two-dimensional simplex, with product features represented as real numbers in the range [0,1]. Each firm manufactures a single

product, represented by a point in this two-dimensional space. Consumers have fixed preferences about what kind of product they would like to purchase. Consumer preferences are also represented in the two-dimensional product feature space. There is no distinction between product features and consumer perceptions of those features.

3.1 Firms

Every N_p iterations of the simulation the firms must examine market conditions and their own performance in the previous iterations, and then modify their product or pricing. A boundedly rational agent can be subject to several kinds of limitations. We focus here on limits on knowledge, and representational and computational power. How these limitations are implemented is detailed below.

The production firms are adaptive learning agents. They adapt to consumer preferences and changing market conditions via a *reinforcement learning* algorithm [Sutton and Barto 1998]. Reinforcement learning has been used to simulate firm behavior in competitive markets [Tesauro 1999], as well as to estimate optimal controllers in many contexts [Bertsekas and Tsitsiklis 1996]. It can be seen as approximating solving for a Nash equilibrium. It is essentially a sampling-based algorithm to solve a dynamic programming problem. It is used here as an alternative to a genetic algorithm as a model for learning in the firm. Reinforcement learning is more appropriate as a model of learning within a trial and within a single firm. This is in contrast to genetic programming, which models learning within a population across many trials. For completeness, we detail the reinforcement learning representation and algorithm below.

3.1.1 State Description

The firms do not have complete information about the environment in which they operate. In particular, they do not have direct access to consumer preferences. They must infer what the consumers want by observing what they purchase. Purchase information is summarized by performing "k-means" clustering on consumer purchases. The number of cluster centers is fixed at the start of the simulation. The current information about the environment consists of the positions of the cluster centers in feature space, along with some additional information. The information is encoded in a bit-vector of "features". The features are summarized in Table 1.

Table 1: Features Available to Production Firms.

Feature	Description
Assets	1 if assets increased in the previous iteration, 0 otherwise
Share Price	1 if share price increased in the previous iteration, 0 otherwise
Mean Price	1 if product price is greater than mean price of competitors products, 0 otherwise
Cluster Center 1	A bit-vector that encodes the position of cluster center 1
...	...
Cluster Center N	A bit-vector that encodes the position of cluster center N

The cluster centers are encoded as binary vectors. Each cluster center can be described as a pair of numbers in $[0, 1] \times [0, 1]$. Two corresponding binary vectors are generated by "binning". Each axis is divided in to K bins. For all of our experiments, K was set to 10. The bit representing the bin occupied

by each number is set to 1. All other bits are 0. For example, given 10 bins per axis and a cluster center $(0.42, 0.61)$, the resulting bit vector is $(0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0)$.

This information gives a summary of the environment at the current time step. Firms make decisions based on the current "state", which is a finite history of H_s bit vectors. In other words, firms make decisions based on a limited memory of length H_s. This limited history window represents an additional explicit limit on the firm's knowledge.

3.1.2 Actions

In each iteration the firms can take one of several actions. The actions are summarized in Table 2. The "Do Nothing" and "Increase/Decrease price" actions are self-explanatory. The "random" action

Table 2: Actions Available to Production Firms.

Action	Description
Random Action	Take a random action from one of the actions below, drawn from a uniform distribution.
Do Nothing	Take no actions in this iteration.
Increase Price	Increase the product price by 1.
Decrease Price	Decrease the product price by 1.
Move down	Move product in negative Y direction.
Move up	Move product in positive Y direction.
Move left	Move product in negative X direction.
Move right	Move product in positive X direction.
Move Towards Center 1	Move the product features towards cluster center 1.
...	...
Move Towards Center N	Move the product features towards cluster center N.

is designed to allow the firm to explicitly try "risky" behavior. The "Move product" actions move the features of the product produced by the firm a small distance in a direction along the chosen axis or towards or away from the chosen cluster center. For example, if the action selected by firm i is "Move Towards Center j" then the product is modified as follows:

$$b_{i,k,t+1} \leftarrow b_{i,k,t} + \nu \left(c_{j,k} - b_{i,k,t}\right) \tag{1}$$

where $k \in \{1, 2\}$ enumerates product features, and $b_{i,k}$ and $c_{j,k}$ are the k^{th} product feature and feature of cluster center j respectively. The update rate $\nu \in (0, 1]$ is a small fixed constant.

3.1.3 Payment Function

A firm's manager seeks to modify its behavior so as to maximize an external payment signal. The payment signal takes the form of a fixed cash payment , a variable amount based on the firm's profitability, and a variable amount due to change in the value of the firms stock. The payment received by firm i at time t is given by:

$$r_{i,t} = S_f + \alpha_\phi \Phi_{i,t} + \alpha_g \left(p_{i,t} - p_{i,t-1}\right) \tag{2}$$

where S_f is the fixed payment, $\Phi_{i,t}$ denotes the profits of firm i at time t, and $p_{i,t}$ denotes the share price of firm i at time t. In our model we assume no costs, so profits are simply the number of items sold times

the price per item. Assets are the accumulation of profits. For all of our experiments, S_f was set to zero for simplicity. The constants α_ϕ and α_g sum to unity. They are fixed at the beginning of the simulation and held constant throughout. They trade off the relative importance of profits and stock price in a firm's decision-making process. The constant payment signal S_f can be interpreted as a fixed salary paid to the manager of the firm. The profit-based payment can be interpreted as an performance-based bonus given to the manager of the firm, and the stock-based payment as a stock grant or stock option.[1] We will further modify this payment signal in section 5 to explicitly include stock option grants.

3.1.4 Utility Function

Given the state history of the simulator for the previous H_s time steps, a production firm makes strategic decisions based on its *utility function*. Utility functions (analogous to "cost-to-go" functions in the control theory literature, and value functions in reinforcement learning) are a basic component of economics, reinforcement learning and optimal control theory [Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998]. Given the "payment signal" or payoff r_t at each time step, the learning agent attempts to act so as to maximize the total expected discounted payment, called expected discounted return, received over the course of the task:[2]

$$R_t = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau\right]_\pi \tag{3}$$

Here $E\left[\cdot\right]_\pi$ denotes taking expectations with respect to the distribution π, and π is the *policy* of the firm. The policy is a mapping $\pi : \mathcal{S} \rightarrow \Delta^{|\mathcal{A}|}$ from states to distributions over actions. In our case $\mathcal{S}$ is the set of possible state histories, and $\mathcal{A}$ is the set of possible actions taken by a firm. The range of the policy $\Delta^{|\mathcal{A}|}$ is the set of probability distributions over actions in $\mathcal{A}$. Note that the discount factor γ encodes how "impatient" the firm is to receive payment. It dictates how much future payments are devalued by the agent. If desired, the discount factor can be set to the rate of inflation or the interest rate in economic simulations, such that the loss of interest on deferred earnings is taken into account by the firm's manager. In our simulations, the discount factor was found using the Markov chain Monte Carlo validation technique (see Sallans, Pfister, Karatzoglou, and Dorffner [2003]).

Given the above definitions, the *action-value function* (or Q-function [Watkins 1989; Watkins and Dayan 1992]) is defined as the expected discounted return conditioned on the current state and action:

$$Q^\pi(\mathbf{s}, a) = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau | \mathbf{s}_t = \mathbf{s}, a_t = a\right]_\pi \tag{4}$$

where $\mathbf{s}_t$ and a_t denote the current state information and action respectively (see tables 1 and 2). The action-value function tells the firm how much total discounted payment it should expect to receive, starting now, if it executes action a in the current state $\mathbf{s}$, and then follows policy π. In other words, it is the firm's expected discounted utility (under policy π) conditioned on the current state and the next action. Note that this is not a myopic measure of payment. This utility function takes into account all future (discounted) payments.

The coding scheme used for world states makes the overall state space quite large. However, in practice, the number of world states observed during a typical simulation is not very large. We can therefore represent the action-value function as a table indexed by unique state histories of length H_s and actions.

[1] The stock based payment changes linearly with stock return. This would be consistent with a limited stock grant, or with a call option where the current price of the underlying stock is significantly above the strike price of the option.

[2] We will drop the firm index i in this section for clarity. The same reinforcement learning algorithm is used for each firm, with the same parameter settings. Each firm learns its own value function from experience.

3.1.5 Reinforcement Learning

Reinforcement learning provides a way to estimate the action-value function from experience. Based on observations of states, actions and payments, the learner can build up an estimate of the long term consequences of its actions.

By definition, the action-value function at time $t-1$ can be related to the action-value function at time t:

$$
\begin{aligned}
Q^{\pi}(\mathbf{s}, a) &= E\left[\sum_{\tau=t-1}^{\infty} \gamma^{\tau-t+1} r_{\tau} | \mathbf{s}_{t-1} = \mathbf{s}, a_{t-1} = a\right]_{\pi} && (5)\\
&= E\left[r_{t-1} | \mathbf{s}_{t-1} = \mathbf{s}, a_{t-1} = a\right]_{\pi} + \gamma E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau} | \mathbf{s}_{t-1} = \mathbf{s}, a_{t-1} = a\right]_{\pi} && (6)\\
&= E\left[r_{t-1} | \mathbf{s}_{t-1} = \mathbf{s}, a_{t-1} = a\right]_{\pi} \\
&\quad + \gamma \sum_{\mathbf{s}', a'} P(\mathbf{s}_t = \mathbf{s}' | \mathbf{s}_{t-1} = \mathbf{s}, a_{t-1} = a) \pi(a_t = a' | \mathbf{s}_t = \mathbf{s}') Q^{\pi}(\mathbf{s}', a') && (7)
\end{aligned}
$$

The first line is just the definition of the action-value function. The second line simply unrolls the infinite series one step, and the third line explicitly replaces the second term by the expected action-value function, again by the definition of expectations under policy π. The last line is called the Bellman equations [Bellman 1957]. It is a set of self-consistency equations (one for each state-action pair) relating the utility at time t to the utility at time $t-1$.

One way to compute the action-value function is to solve the Bellman equations. We use a reinforcement learning technique called SARSA [Rummery and Niranjan 1994; Sutton 1996]. SARSA can be viewed as using a Monte Carlo estimate of the expectations in Eq.(7) in order to iteratively solve the Bellman equations. At time t, the estimate of the action-value function $\widehat{Q}_t(s, a)$ is updated by:

$$
\widehat{Q}^{\pi}_t(\mathbf{s}_{t-1}, a_{t-1}) = (1-\lambda)\widehat{Q}^{\pi}_{t-1}(\mathbf{s}_{t-1}, a_{t-1}) + \lambda\left(r_{t-1} + \gamma \widehat{Q}^{\pi}_{t-1}(\mathbf{s}_t, a_t)\right) \tag{8}
$$

where λ is a small learning rate. For all of our experiments, λ was set to 0.1. In words, the utility function is updated as a linear mixture of two parts. The first part is the previous estimate. The second part is a Monte Carlo estimate of future discounted return. It includes a sample of the payment at time $t-1$ (instead of the expected value of the payment), and the utility for a sampled state and action at time t (instead of the expected utility based on the policy and state transition probability). Finally, the current estimate of the action-value function is used, in place of the true action-value function.

Intuitively this learning rule minimizes the squared error between the action-value function and a bootstrap estimate based on the current payment and the future discounted return, as estimated by the action-value function. Theoretically, this technique has been closely linked to stochastic dynamic programming and Monte Carlo approximation techniques [Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998].

After each action-value function update, a new policy π' is constructed from the updated action-value function estimate:

$$
\pi'(a|\mathbf{s}) \leftarrow \frac{\exp\left(\widehat{Q}^{\pi}(\mathbf{s}, a)\right)}{\sum_{a'} \exp\left(\widehat{Q}^{\pi}(a'|\mathbf{s})\right)} \tag{9}
$$

This policy selects actions under a Boltzmann distribution, with better actions selected more frequently.

In theory, using the SARSA algorithm, the action-value function estimate will converge to the optimal action-value function. However, convergence relies on the learner operating in a stationary stochastic Markov environment [Sutton and Barto 1998]. When there is more than one adaptive firm in the environment, the stationarity assumption is violated. Nevertheless, it has been shown that reinforcement learning can be used to approximately solve arbitrary-sum competitive games [Tesauro 1999]. As noted previously, the Markov assumption is also violated, since the state vector of the firm does

not include all information necessary for solving the task. For example, explicit consumer preferences and exact product positions are not known by the firm. Such limited-memory reinforcement learning algorithms have been used previously to approximately solve partially observable problems [Jaakkola, Singh, and Jordan 1995]. This approximation can be seen as another source of "boundedness" in a boundedly rational firm. As well as having a limited knowledge base (represented by the partial state history and state aggregation), the firm has a limited model of the capabilities of other firms (they are assumed to have a stationary policy). Thus our firms implement an algorithm designed to iteratively improve their strategies, given the constraints on their knowledge and computational power.

3.2 Consumers

Consumers are defined by their product preference. Each consumer agent is initialized with a random preference in product feature space. During each iteration of the simulation, a consumer must make a product purchase decision. For each available product, the consumer computes a measure of "dissatisfaction" with the product. Dissatisfaction is a function of product price and the distance between the product and the consumer's preferred product.

Modeling consumer decision making as measuring distance in a product feature space is a common technique. It appears both in computational models of consumer behavior [Buchta and Mazanec 2001] and in theoretical models of customer preferences [Lancaster 1966].

We use a simplified model of consumer behavior. Consumer i's dissatisfaction with product j is given by:

$$\mathrm{DIS}_{i,j} = \alpha_c \frac{D(\beta_i, b_j)}{\max_{b'} D(\beta_i, b')} + (1 - \alpha_c) \frac{\rho_j}{\max_j \rho_j} \tag{10}$$

where ρ_j denotes the price of product j, and α_c trades off the importance of product features and price. For all of our experiments, $alpha_c$ was set to 0.5. The measure $D(\beta_i, b_j)$ is the Euclidean distance in feature space between the ideal product of customer i and product j:

$$D(\beta_i, b_j) = \frac{(\beta_i - \mathbf{b}_j)^\top \mathbf{W} (\beta_i - \mathbf{b}_j)}{{\beta_i}^\top \mathbf{W} \mathbf{b}_j} \tag{11}$$

Here bold-faced letters denote the feature-vector representations of products and preferences. The diagonal matrix $\mathbf{W}$ is common to all consumers and models the relative importance of features in the feature space. In all of our simulations, the matrix $\mathbf{W}$ was initialized to the identity matrix.

Every consumer is also initialized with a "ceiling" dissatisfaction MAXDIS_i. The ceiling dissatisfaction acts as a limit. If all product dissatisfactions are above its ceiling, a consumer will simply make no purchase in that iteration. For all of our simulations, MAXDIS was set to 0.96.

Given dissatisfaction ratings for all products with dissatisfactions below the ceiling in iteration t, consumer i selects from this set the product j with the lowest dissatisfaction rating:

$$j = \arg\min_k \left\{ \mathrm{DIS}_{i,k} \right\},\ k \in \left\{ l : \mathrm{DIS}_{i,l} < \mathrm{MAXDIS}_i \right\} \tag{12}$$

In order to avoid sharp boundary effects when the dissatisfaction is exceeded, the probability of buying the product specified in equation (12) decreases linearly from 1 to 0 as DIS goes from 0.8 to 0.96. The result is that the probability of purchasing is non-zero for any dissatisfaction below MAXDIS, and drops to zero when dissatisfaction exceeds MAXDIS.

4 The Financial Market

Our financial model is based on a standard capital market model (see e.g. [Arthur, Holland, LeBaron, Palmer, and Tayler 1997; Brock and Hommes 1998; Dangl, Dockner, Gaunersdorfer, Pfister, Soegner,

and Strobl 2001; Sallans, Pfister, Karatzoglou, and Dorffner 2003; Pfister 2003]). Myopic investors maximize their next period's utility subject to a budget restriction. At time t agents invest their wealth in a risky asset with price p_t and in bonds, which are assumed to be risk free. Each agent only trades with stocks of a single firm. Within this section we therefore drop the index i for the firms. There are S stocks paying a dividend d_t. It is assumed that firms pay out all of their profits if positive, therefore each stock gets a proportion of $\frac{1}{S}$ of the profits. In our model dividends should be viewed as an information signal about the fundamental situation of the firm.

The risk free asset is perfectly elastically supplied and earns the risk free and constant interest rate κ. Investors are allowed to change their portfolio in every period. The wealth of investor m at time $t+1$ is given by

$$W_{m,t+1} = (1+\kappa)\, W_{m,t} + (p_{t+1} + d_{t+1} - (1+\kappa)\, p_t)\, q_{m,t} \tag{13}$$

where $W_{m,t+1}$ is the wealth at time t and $q_{m,t}$ the number of stocks of the risky asset hold at time t. As in [Brock and Hommes 1998] , [Levy and Levy 1996] , [Chiarella and He 2001] , and [Chiarella and He 2002] the demand functions of the following models are derived from a Walrasian scenario. This means that each agent is viewed as a price taker (see [Brock and Hommes 1997] and [Grossman 1989])

	Denote
p_t:	Price (ex dividend) per share of the risky asset at time t
d_t:	Dividend at time t
κ:	Risk free rate
S:	Total number of shares of the risky asset
M:	Total Number of investors
$q_{m,t}$:	Number of shares investor m wants to hold at time t
$W_{m,t}$:	Wealth of investor m at time t
ζ_m	Risk aversion of investor m

Let an investor m with wealth W_m maximize his/her utility of the form

$$u\left(W_m\right) = -e^{-\zeta_m W_m} \tag{14}$$

with ζ_m as constant absolute risk aversion. Denote by $F_t = \{p_{t-1}, p_{t-2}, ..., d_{t-1}, d_{t-2}\}$ the information set available at time t[3]. Let $E_{m,t}$ and $V_{m,t}$ be the conditional expectation and conditional variance of investor m at time t based on F_t. Then the demand for the risky asset $q_{m,t}$ solves

$$\max_{q_{m,t}} \left\{ E_{m,t}\left(W_{m,t+1}\right) - \frac{\zeta_m}{2} V_{m,t}\left(W_{m,t+1}\right) \right\} \tag{15}$$

i.e.,

$$q_{m,t} = \frac{E_{m,t}\left(p_{t+1} + d_{t+1}\right) - p_t\left(1+\kappa\right)}{\zeta_m V_{m,t}\left(p_{t+1} + d_{t+1}\right)} \tag{16}$$

Let S be the total number of shares, then the market clearing price p_t is implicitly given by the equilibrium equation

$$S = \sum_{i=1}^{M} q_{m,t} \tag{17}$$

[3]Note that at time t price p_t and dividend d_t are not included in the information set F_t

4.1 Formation of expectations

It is well known that expectations play a key role in modeling dynamic phenomena in economics. Heterogeneous expectations are introduced in the following way:

$$\begin{aligned} E_{m,t}\left(p_{t+1}+d_{t+1}\right) &= F_m\left(p_{t-1},\cdots p_{t-h_m},d_{t-1},\cdots,d_{t-h_m}\right) \\ V_{m,t}\left(p_{t+1}+d_{t+1}\right) &= G_m\left(p_{t-1},\cdots p_{t-h_m},d_{t-1},\cdots,d_{t-h_m}\right) \end{aligned} \tag{18}$$

Heterogeneity arises from different information sets and different prediction functions. Agents have three characteristics:

- Type of prediction function: Fundamentalist or chartist
- Time horizon: Length of history of past prices and dividends used for prediction
- Trade interval: Intraday trader or end-of-day trader

These characteristics are initialized at the beginning of the simulation and are held fixed thereafter. In order to keep the simulation simple, there are only two types of prediction functions and the trade interval can only take one out of two different values. The above three characteristics can be combined in any way, e.g. chartists trading intraday or fundamentalists, who are end-of-day traders.

First let's have a closer look at the type of prediction function. As in many other heterogeneous agents models we assume that two kinds of investors exist: Fundamentalists and chartists. Chartists use past prices to predict the future price whereas fundamentalists calculate a "fair value" based on past dividends. Therefore for a fundamentalist $E_{i,t}\left(p_{t+1}\right)$ is a function of past dividends and for a chartist $E_{i,t}\left(p_{t+1}\right)$ is a function of past prices.

The second characteristic of agents is their time horizon. The time horizon is the length of history h_i of past prices and dividends used for their predictions. Time horizon means how far back agents look into past to predict next period's price and dividend. A long time horizon, i.e. high h_i means that many past observations are used as a forecast. The time horizon is drawn from an uniform distribution. The range of this uniform distribution depends on the agents type of prediction function. It was found that for fundamentalists and chartists different time horizons were appropriate. In particular chartists have to identify trends. The more observations they use, the further they look into the past which also means that longer trends can be identified. For example, if prices followed an upward trend, thereafter a downward trend, a short term chartist, i.e. a chartist with a short time horizon, would extrapolate the downward trend. This trader would predict a further drop in the stock price. For a long term chartist the last two trends might cancel each other and this agent would predict no change in the stock price. However the situation for fundamentalists may be different. Fundamentalists use past dividends for their price prediction. Depending on the type of the dividend process, it might be favorable to base the prediction only on the last observation. The minimum time horizon for a fundamentalist is 1, for a chartist it is 2 (the chartist needs at least 2 prices to calculate a trend, see also equation (26)). We allow two different maximum values for the time horizon for fundamentalists and chartists. For more details please see the results (section 4).

The third characteristic of an agent is the trade interval ϕ_i. Investors update their portfolio every ϕ periods. This corresponds to the parameter N_p for the firms, which update their action every N_p periods. In our model we assume five price fixings per day with the last price setting being the closing price of the day. Therefore every 5^{th}, 10^{th}, 15^{th} etc. price is a closing price. We distinguish between two types of agents concerning their trading interval: Intraday traders with $\phi_i = 1$ and end-of-day traders with $\phi_i = 5$. Intraday traders update their position at every time step, whereas end-of-day traders only trade at the closing price of each day. This means that in the first four price settings per day only intraday traders re-balance their portfolios. In the fifth price fixing (the closing price) all agents trade. End-of-day traders ignore intraday prices, i.e. prices between their updates. The dividends of the five previous

periods are accumulated. In order to keep consistent notation between intraday and end-of-day traders, we will define prices and dividends as used by the intraday traders as they are related to the highest frequency information available. Given the high-frequency price and dividend information $\widehat{p}_t$ and $\widehat{d}_t$, the price p_{t-1} and dividend d_t used by a trader with interval ϕ_i is given by:

$$\begin{aligned} p_{t-1} &= \widehat{p}_{t-\phi_i} \\ d_t &= \sum_{j=0}^{\phi_i-1} \widehat{d}_{t-j} \end{aligned} \tag{19}$$

Therefore the information set used by trader i with interval ϕ_i is given by

$$\begin{aligned} F_{h_i,\phi_i,t} : &\left\{ \widehat{p}_{t-\phi_i}, \widehat{p}_{t-2\phi_i}, ..., \widehat{p}_{t-h_i\phi_i}, \sum_{j=0}^{\phi_i-1} \widehat{d}_{t-j}, \sum_{j=0}^{\phi_i-1} \widehat{d}_{t-\phi_i-j}, ..., \sum_{j=0}^{\phi_i-1} \widehat{d}_{t-h_i\phi_i-j} \right\} \\ &\rightarrow \{p_{t-1}, p_{t-2}, ..., p_{t-h_i}, d_t, d_{t-1}, ..., d_{t-h_i}\} \end{aligned} \tag{20}$$

Now let's have a closer look at the formation of expectation. Let's begin with the expectation for the variance $V_{i,t}$. Agents determine $V_{i,t}$ in the following way

$$\begin{aligned} V_{i,t}\left(p_{t+1} + d_{t+1}\right) &= \frac{1}{h_i} \sum_{j=1}^{h_i} \left(p_{t-j} + d_{t-j} - M\right)^2 \\ \text{with } M &= \frac{1}{h_i} \sum_{k=1}^{h_i} \left(p_{t-k} + d_{t-k}\right) \end{aligned} \tag{21}$$

Fundamentalists and chartists have the same prediction function for the variance. Their information set depend on their history h_i of past prices and dividends and their trade interval ϕ_i.
Now we take a look at the expectation of next period's price and dividend. First we split $E_{i,t}\left(p_{t+1} + d_{t+1}\right)$ into $E_{i,t}\left(p_{t+1}\right)$ and $E_{i,t}\left(d_{t+1}\right)$. Investors form their expectations over the next periods dividend d_{t+1} in the following way:

$$E_{i,t}\left(d_{t+1}\right) = \frac{1}{h_i + 1} \sum_{j=0}^{h_i} d_{t-j} \tag{22}$$

Fundamentalists determine their price expectation according to a model based on fundamental information, which in our model are past dividends. A fundamentalist i assumes that the fair price $p_{i,t}^{\text{Fair price}}$ is a linear function of past dividends, i.e.

$$\begin{aligned} p_{i,t}^{\text{Fair price}} &= F_i\left(d_t, \cdots, d_{t-h_i}\right) \\ &= f \frac{1}{h_i + 1} \sum_{j=0}^{h_i} d_{t-j} + g \end{aligned} \tag{23}$$

This leads to the following price and dividend expectation for the fundamentalist

$$E_{i,t}\left(p_{t+1} + d_{t+1}\right) = \left(f + 1\right) \frac{1}{h_i + 1} \sum_{j=0}^{h_i} d_{t-j} + g \tag{24}$$

Chartists use the past history of the stock prices in order to form their expectations. They assume that the future price change per period equals the average price change during the last h_i periods[4].

$$E_{i,t}\left(p_{t+1}\right) = p_{t-1} + \frac{p_{t-1} - p_{t-h_i}}{h_i - 1} \tag{25}$$

Note that at time t, p_t is not included in the information set F_t. This leads to the following price and dividend expectation of the chartists

$$E_{i,t}\left(p_{t+1} + d_{t+1}\right) = p_{t-1} + \frac{p_{t-1} - p_{t-h_i}}{h_i - 1} + f\frac{1}{h_i + 1}\sum_{j=0}^{h_i} d_{t-j} + g \tag{26}$$

4.2 Sequence of Events

Let us have a look at the timing of the events within the equity market model. First the dividend d_t of the current period is announced and paid. The next step is the formation of expectations. Based on past prices and dividends, including d_t an investor i forms his/her expectation about the distribution of the next period's price and dividend. According to equation (21), (24) and (26) the investor calculates $V_{m,t}\left(p_{t+1} + d_{t+1}\right)$ and $E_{i,t}\left(p_{t+1} + d_{t+1}\right)$. Plugging the expectations into Equation (16) the agent is able to determine the demand function, which is submitted to the stock market via limit buy orders and limit sell orders[5]. After the orders of all agents are collected, the stock market calculates this period's equilibrium price p_t.

The market uses a sealed-bid auction, where the clearance mechanism chooses the price at which trading volume is maximized. The constructed supply and demand curves are based on the transaction requests.

The artificial return series generated by these traders exhibits the most important stylized facts from real markets, such as insignificant autocorrelation of returns, volatility clustering and high kurtosis. The statistics presented in figure 2 and table 3 were computed by averaging the performance for two risk-neutral firms over 20 simulation runs. Figure 2 presents the mean autocorrelation coefficients of returns and absolute returns and the errorbars correspond to the standard error of the coefficients across the 20 simulation runs. The autocorrelation coefficients of the returns are insignificant (except lag 4), and the significants of the autocorrelations of the absolute returns indicates volatility clustering.

Table 3: Summary Statistics for the Artificial Stock Returns.

Statistic	Value
mean	2.9×10^{-4}
standard dev	2.3×10^{-3}
kurtosis	39.2

5 Compensation

Owners of a company delegate authority to a manager. The goal of the owners is to encourage the manager to increase the value of their company. The goal of the manager is to maximize its compensation.

[4]Note that for chartists $h_i >= 2$ for computing an average price change

[5]A limit order is an instruction stating the maximum price the buyer is willing to pay when buying shares (a limit buy order), or the minimum the seller will accept when selling (a limit sell order).

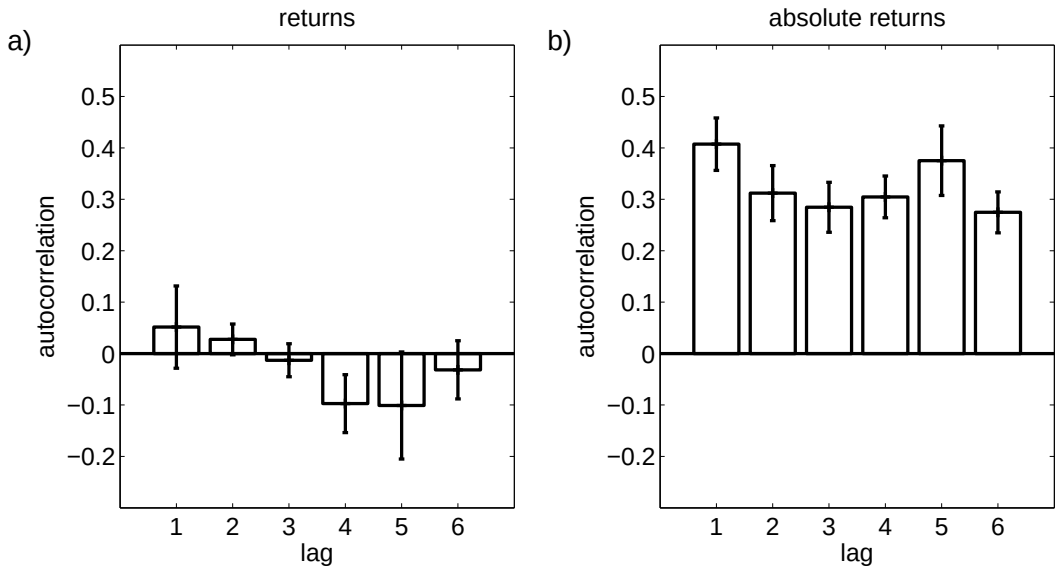

Figure 2: Autocorrelations for (a) the stock returns and (b) the absolute stock returns. The correlations of the returns are not significant (except for lag four), and for the absolute returns are significantly positive.

In the IMM, each manager seeks to modify its behavior so as to maximize an external payment signal. In the basic model, this payment takes the form of a fixed cash salary, a variable amount based on the firm's profitability, and a variable amount due to change in the value of the firm's stock. We modify Eq.(2) to add compensation from stock options. The payment received by manager i at time t is given by:

$$r_{i,t} = S_f + \alpha_\phi \Phi_{i,t} + \alpha_g \mathrm{g_i} + \alpha_o \mathrm{o_i} \tag{27}$$

where S_f is the fixed salary, $\Phi_{i,t}$ denotes the profit-based bonus at time t, and $\mathrm{g_i}$ and $\mathrm{o_i}$ are bonuses from stock grants and stock-options respectively.

For all of our experiments, S_f was set to zero for simplicity.[6] The constants α_ϕ, α_g and α_o are fixed at the beginning of the simulation and held constant throughout. They trade off the relative importance of profits and stock-based pay in a manager's decision-making process.

5.1 Cash and Stock

A manager in the IMM is rewarded once in every time period, using a combination of cash and stock-based bonuses. Specifically, the manager's compensation is based on a profit-based cash bonus, a stock grant, and a stock option.

The profit-based bonus is proportional to the profits of the firm:

$$\Phi_{i,t} = c_{i,t} - c_{i,t-1} \tag{28}$$

where $c_{i,t}$ are the current assets of firm i at time t.[7]

[6]The absolute level of salary can influence the result. For example, if the salary is much higher than the bonus, then the manager is not motivated to do anything to achieve the extra bonus. By setting the basic salary to zero, we focus on the influence of the bonus on manager behavior

[7]This is slightly unrealistic, in that all of the bonuses can be either positive or negative. In psychology, it is well known that positive and negative payments do not have symmetric impacts. However, in our reinforcement learning model of firm behavior, the sign of the payment is not significant. By centering bonuses around zero, we improve the numerical stability of the algorithm.

The stock grant bonus is proportional to the change in stock price:

$$g_{i,t} = p_{i,t} - p_{i,t-1} \tag{29}$$

where $p_{i,t}$ is the stock price of firm i at time t.

The value of the stock option bonus is the change in value of the stock option held by the manager:

$$o_{i,t} = b_{i,t} - b_{i,t-1} \tag{30}$$

where $b_{i,t}$ is the value of the stock option held by manager i at time t. The value of the stock option is computed using the Black-Scholes formula. It is dependent on the current stock price; the strike price; the risk-free interest rate; the volatility of the underlying stock; and the time periods until the option vests. In the following, X is the strike price; κ is the risk free interest rate; T is the number of time periods until the option vests; v is the per-time-period variance of the stock price; and p is the current stock price. The function $N(x)$ returns the cumulative probability of x being drawn from a Normal distribution with zero mean and unit variance (it is the Normal cumulative distribution function).

$$\begin{aligned} C &= p\exp\left(-\kappa * T\right) && (31)\\ s &= \sqrt{v * T} && (32)\\ x_1 &= \log\left(p/C\right)/s + s/2 && (33)\\ x_2 &= \log\left(p/C\right)/s - s/2 && (34)\\ b &= SN(x_1) - CN(x_2) && (35) \end{aligned}$$

Black-Scholes is used because it is a commonly-used technique for option valuation of management options.

5.2 Risk Aversion

In order to assess the worth and riskiness of an action, managers in the IMM estimate two quantities. First, as discussed in the model description, they estimate the expected discounted payment they will receive after taking an action given the current world state:

$$Q^{\pi}(\mathbf{s}, a) = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau} | \mathbf{s}_t = \mathbf{s}, a_t = a\right]_{\pi} \tag{36}$$

We have dropped the firm index i for simplicity. All firms use the same algorithms.

Second, they estimate the variance of this discounted payment:

$$V^{\pi}(\mathbf{s}, a) = V\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau} | \mathbf{s}_t = \mathbf{s}, a_t = a\right]_{\pi} \tag{37}$$

Variance is estimated in the same way as the expected return, using stochastic dynamic programming.

Given this estimate of expected value and variance of value, the manager selects which action to perform based on its risk-adjusted expected discounted value:

$$\begin{aligned} R_p^a &= \sqrt{|\widehat{V}^{\pi}(\mathbf{s}, a)|} - E\left[\sqrt{|\widehat{V}^{\pi}(\mathbf{s}, a')|}\right]_{\pi(a')}\\ P(a|\mathbf{s}) &= \frac{\exp\{-\widehat{Q}^{\pi}(s, a) - \rho R_p^a\}}{\sum_{a'} \exp\{-\widehat{Q}^{\pi}(s, a') - \rho R_p^{a'}\}} \end{aligned} \tag{38}$$

The quantities $\widehat{Q}^{\pi}$ and $\widehat{V}^{\pi}$ denote the firm's estimate of Q^{π} and V^{π} respectively. The firm's utility function (Eq.(36)) is modified by subtracting the risk aversion penalty. This, in turn, alters the policy used

to select actions (Eq.(9)). The risk penalty associated with action a, R_p^a, is the above-average variance of the payoff associated with that action (i.e. the average variance across all actions is subtracted).

The risk aversion factor ρ indicates how risk-averse the manager is. At $\rho = 0$, the manager is risk-neutral. Negative ρ would make the manager risk-prone.

The reader should note that both the value and risk used by the manager are estimates, based on past experience. Unlike many analytic models, we do not assume that the manager has a priori perfect knowledge of value or risk. In fact, the quality of the estimates will be influenced by the actions taken by the manager, which in turn are influenced by the estimates.

6 Simulation Results

In this section we present simulation results from the IMM. First we test the robustness of the predictions of theoretical compensation models. We then propose new hypotheses based on new simulation experiments. All simulations were run with two competing firms and 100 traders, and lasted for 5000 periods. Each simulation scenario was repeated 20 times.

6.1 Model Parameters

The firm's learning algorithm and trader's decision rules have tuning parameters. Parameter values must be selected before a simulation can be run. These parameters have been introduced in earlier sections describing each of the agents in the model. Parameter values were found using a novel Markov chain Monte Carlo simulation technique described in [Sallans, Pfister, Karatzoglou, and Dorffner 2003]. The parameters are "tuned" so that the model reproduces empirical behaviors from real consumer and financial markets. The parameters are summarized for convenience in table 4. The "reference" column indicates where in the text the parameter was introduced, and where more details can be found. The "value" column indicates the value used for simulations, unless noted otherwise.

Table 4: Parameters for Integrated Markets Simulator

Parameter	Description	Value	Reference
α_ϕ	strength of profitability reinforcement	0.47	Eq.(2)
α_g	strength of stock price reinforcement	0.53	Eq.(2)
α_o	strength of stock option reinforcement	0.05	Eq.(27)
ν	product update rate	0.003	Eq.(1)
γ	reinforcement learning discount factor	0.85	Eq.(3)
N_f	Proportion of fundamentalists	0.57	section 4
N_f	Proportion of intra-day traders	0.48	section 4
N_p	product update frequency	5	section 2
λ	reinforcement learning rate	0.1	Eq.(8)
α_c	Consumer feature/price tradeoff	0.5	Eq.(10)
MAXDIS_i	Maximum dissatisfaction for consumer i	0.8	section 3.2
κ	Risk-free Interest rate	0.4% per period	section 4
ζ	trader risk aversion	0.001	section 4
h_i	time horizon for fundamentalist/chartist trader i	[1,13] / [2,74]	section 4.1
ϕ_i	trading interval for trader i	$\{1,5\}$	section 4.1
T	duration of stock options	250	section 5.1
ρ	firm risk aversion	0.5	section 5.2

6.2 Risk Aversion and Options

Because they limit down-side risk, stock options become more valuable in volatile conditions. This is reflected in option pricing models such as Black-Scholes [Black and Scholes 1973]. Because of this, it has been theorized that executive stock options should reduce risk aversion. That is, a manager's risky actions will result in stock price volatility, increasing the value of the options. Empirical studies suggest that in some industries, stock options can lead to more risk-taking behavior [Rajgopal and Shevlin 2002]. However, there are other conflicting and inconclusive results. To add to the confusion, it is not entirely clear how executive options should be valued [Hall and Murphy 2002], or if this mechanism is even considered when awarding options. See Murphy [1999] for a review of executive compensation.

In order to clarify whether stock options enhance risk-taking behavior, we simulated the use of options with risk-neutral and risk-averse managers. We also used two different market conditions: one where the firms can quickly modify their products, and one where the products can only be modified slowly.

Our purpose for doing this simulation is twofold. First, it is unclear from empirical studies of stock option usage if and when awarding options increases risk-taking. Theoretical models indicate this should be the case. We will use the IMM as a test-bed to confirm this theory. There is no explicit link between option value and risk-taking in our model, and the firm decision-making was not designed with this test in mind. The IMM therefore offers a good test of the robustness of the influence of stock options under conditions that are more realistic than the theoretical models: the managers are uncertain, can suffer from misconceptions, and do not know with certainty the outcome or riskiness of their actions. To our knowledge, this is the first attempt to validate the predictions of principle-agency theory in a boundedly-rational, learning system.

Second, we would like to demonstrate that reinforcement learning is a good tool for modeling learning and decision-making in the firm. If the behavior of the model matches theoretical predictions, then we can conclude that the managers are behaving rationally, within the bounds of their knowledge. This would indicate that reinforcement learning is a good alternative to evolutionary algorithms for modeling learning during the lifetime of a firm.

For each market condition we did three types of simulations: One with risk-neutral managers, one with risk-averse managers, and one with risk-averse managers with stock options. For the risk-averse managers, the risk aversion $\rho = 0.5$, and the strength of the stock option compensation was $\alpha_o = 0.05$. The results, averaged over the 20 simulations, are shown in figure 3. For the options, the option duration was 250 periods; the option is granted slightly out of the money at 1.05 times the current stock price; and the interest rate κ was 0.4% per period. For valuation purposes, the current variance of the stock was estimated over the last 10 periods. This short interval was chosen to decrease the time lag between a variance-enhancing action and its subsequent effect on option value.

In both market conditions, the risk-averse managers have much worse performance, and lower variance in their achieved profits. The effect of risk aversion is clearly reduced by adding stock options to the manager's compensation. In the case of the slow-market condition, the average profits achieved by the firm are equivalent to the risk-neutral case. The performance with stock options in the fast-market condition is slightly inferior to the performance in the slow-market condition (the difference is significant at the 5% level, according to a t-test).

The simulated stock options have the effect predicted by principal-agent theory. The behavior of the risk-averse manager leads to lower but more stable profits on average. The use of stock options boosts both expected profits and profit volatility. The effect of options is smaller in the fast market condition. This may simply be because the effect of product update actions is easier to estimate in the slow market condition. When all actions have similar volatility, it requires fewer iterations to get an accurate estimate. This is because the dynamic programming algorithm uses information from future actions to estimate the volatility of current actions. An inaccurate estimate of action volatility results in inaccurate action selection, and lower profits during learning. The performance with no stock options (in the risk-averse case) is also slightly inferior in the fast market condition to the slow market condition

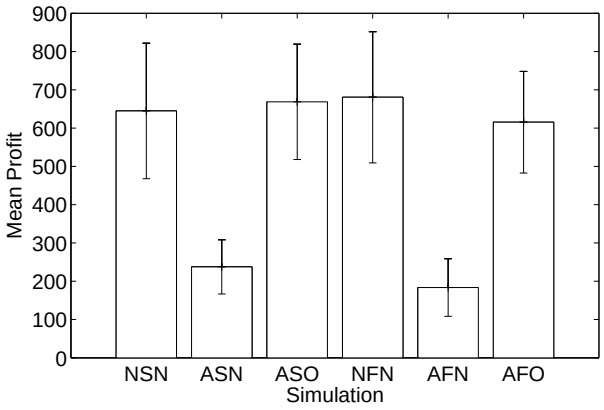

Figure 3: The effect of stock options on risk aversion. Mean and standard error of per-time-step profits are shown for six conditions: N??==risk neutral, A??==risk averse; ?S?==slow product movement, ?F?==fast product movement; ??N==no stock options, ??O==stock options. The risk-averse managers cause mean profits to drop, and the variance of the profits to be reduced. In the slow market case, options boost expected profits and profit volatility. In the fast market conditions, the options also boost profits, but not as much as in the slow market case.

(the difference is significant at the 5% level, according to a t-test). We show the difference in action volatility in the two conditions in figure 4).

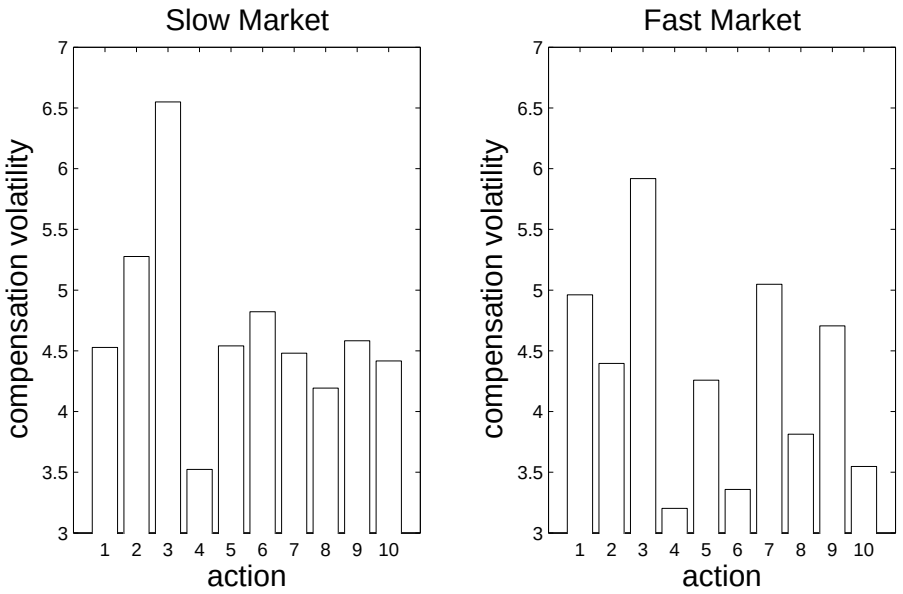

Figure 4: The manager's average estimate of standard error for each potential action in the a) slow-moving consumer market and b) fast-moving consumer market. The actions are: 1. Take a random action from the remaining actions. 2. Do nothing. 3. Increase product price by 1. 4. Decrease product price by 1. 5-10. Modify product attributes.

What does this mean to real markets? The benefit of stock options depends on being able to correctly identify high-risk actions. If the manager is not able to accurately estimate how risky an action is, the benefit of stock options will be muted.

Notice a byproduct of this simulation: In a competitive market with price-sensitive consumers, raising prices is very risky. In the fast-moving market, doing nothing actually becomes a safer bet (has lower variance) than in the slow-moving market. Lowering prices is always safe.

These simulations also validate the choice of reinforcement learning to model learning and decision-

making in the firm. Using this technique managers are able to estimate the value and the volatility of their actions. The managers act rationally, within the bounds of their knowledge. Reinforcement learning offers a good technique for modeling learning in the firm.

6.3 Market Competition and Options

In this section we model a scenario which is difficult to address with an analytic model: How stock options influence the performance of a new competitor entering a market dominated by an incumbent firm. The incumbent has the benefit of prior experience in the market. This could also be seen as a disadvantage: The competitor does not have to overcome old habits that are no longer valid in the new competitive environment.

Hypothesis 1: The incumbent will have an inherent advantage because of its prior knowledge of the market.

Hypothesis 2: Options will help the incumbent, because they will promote exploration.

Hypothesis 3: Options will help the entrant for the same reason.

Here "exploration" means trying new products or pricing strategies, in order to learn whether or not they work. Exploration is inherently a risky action, because it sacrifices current profits in an attempt to boost future profits. Knowledge acquisition by the manager is a crucial consideration in any realistic model of firm behavior.

We modeled four scenarios: The incumbent and entrant have no stock options; the entrant receives stock options with the incumbent having none; and two scenarios where both receive options. In these last two scenarios, the incumbent receives its options either from the start of the simulation, or when the new entrant arrives. All managers are risk averse, and all simulations run for 5000 iterations. The entrant enters the market at iteration 2000 (see figure 5).

Given no options, the incumbent on average does better than the new entrant (significant at the 1% level according to a t-test). These results support hypothesis 1. The results also support hypothesis 2: When the entrant is granted options, it does as well as the incumbent. Note that the incumbent now does less well because of competition from the entrant. The results for hypothesis 3 are mixed. When options are granted to the incumbent at the beginning of the simulation, it does no better than when it had no options (i.e. the result is not significant according to a t-test). However, when options are granted only after the new entrant appears, it does better on average than without options (significant at the 5% level according to a t-test).

This suggests that encouraging risk-taking is not enough, but rather the incumbent manager needs to be encouraged to take risks specifically after the new threat appears. This is followed by a period of new exploration and learning, which results in the higher profits. Figure 6 shows average profits versus time for the incumbent firm for the last two stock options conditions. Initially, there is no difference in performance between the two. It is only after some learning and exploration time that the performance difference is seen.

We can therefore make a new prediction to be tested using empirical data: While options will help a new entrant be competitive in a new market, they will be most effective in helping the incumbent when they are granted after the new entrant appears. This suggests that re-examination of compensation is particularly important when a firm is facing new competition. Encouraging risk-taking at this time can be particularly helpful.

7 Conclusions

In this article we have presented a computational economics model of managerial compensation. We have simulated risk-averse managers with and without stock-option compensation, and shown that the

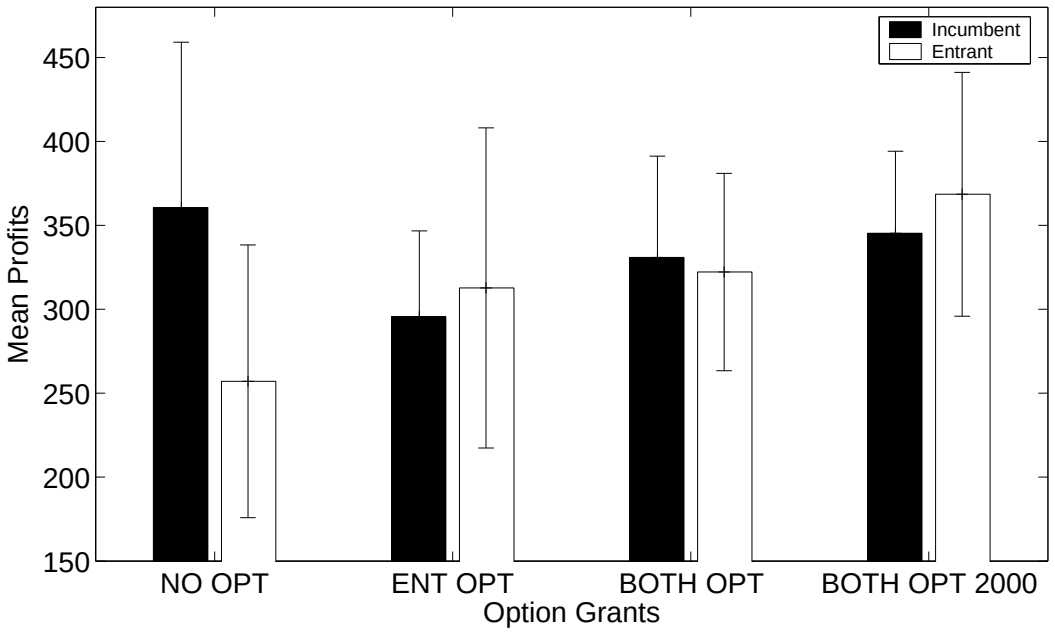

Figure 5: Results of a new entrant in the market. The bar graphs show the average per-time-step profits and standard errors for four scenarios: no stock options (NO OPT); an entrant firm with options (ENT OPT); both with options (BOTH OPT); and both with options, where the incumbent firm has options granted at the time the new entrant arrives (time t=2000) (BOTH OPT 2000). The incumbent is black, and the new entrant is white. The new entrant was always granted stock options from time t=1. The averages were taken over the time from the arrival of the new entrant (time t=2000) to the end of the simulation (time t=5000). Options granted to the entrant help it to compete with the incumbent. Options helped the incumbent when they were granted at the time that the new entrant arrived.

computational model confirms the predictions of principal agency theory. In particular, stock options encourage risk taking in otherwise risk-averse managers, and can boost overall profits by encouraging exploration. In addition, we show that these effects are quite robust, occurring in our model in the presence of learning and incomplete knowledge. As a byproduct, we show that reinforcement learning offers a valuable alternative for modeling learning and decision-making in economic agents. We believe that it is a better model of learning within the lifetime of a process than alternatives such as genetic algorithms.

We simulate the scenario of a new entrant challenging an incumbent firm. We show that stock options can boost the competitiveness of the entrant, and also help the incumbent to fight off competition. In the latter case, the options are most effective when they are introduced as a response to the new competition, boosting exploration and learning in the new competitive environment. This suggests that it is particularly important to revisit compensation contracts, introducing risk-taking incentives, when the state of competition in the market changes. It also suggests that empirical research into the effectiveness of stock options should take the experience of the manager and age of the firm into account. In our simulations, options are most effective in the situation where new knowledge acquisition is crucial to success.

Acknowledgements

This work was funded by the Austrian Science Fund (FWF) under grant SFB#010: “Adaptive Information Systems and Modeling in Economics and Management Science”. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.

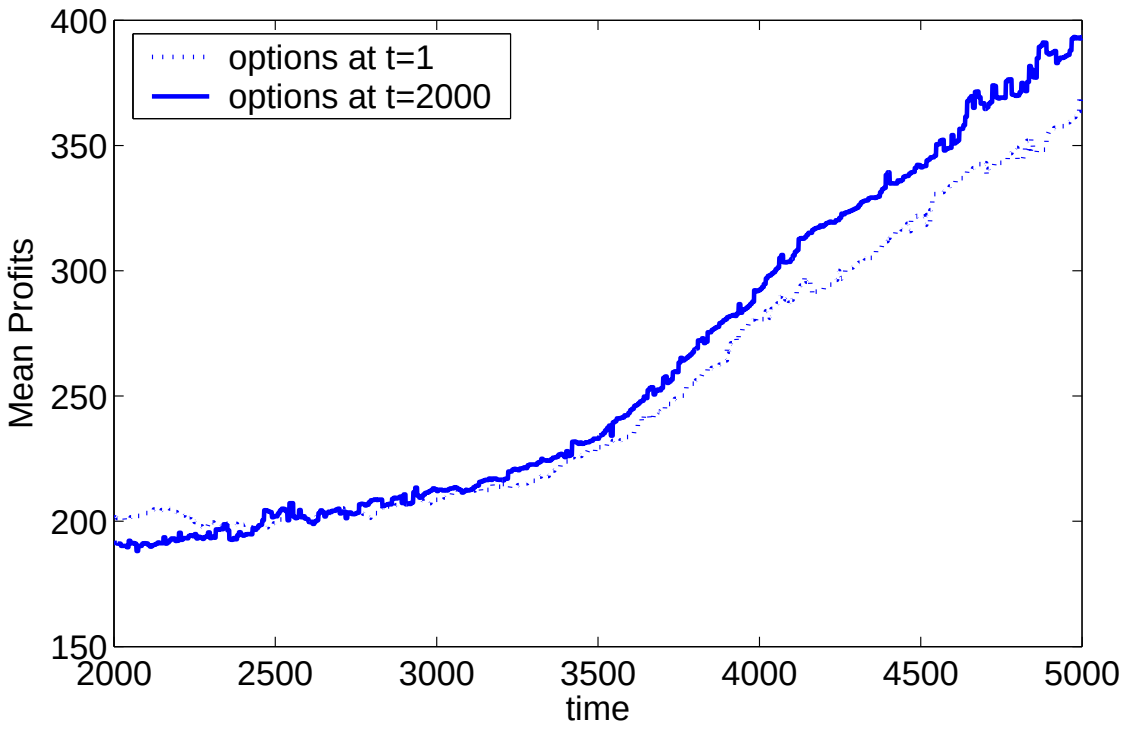

Figure 6: Effect of learning on the granting of stock options. The dotted line shows the average profits of the incumbent when options are granted from the start of the simulation. The solid line shows the average profits of the incumbent when options are granted only starting at time t=2000, when the new entrant appears. After a period of learning and exploration in the new environment, the second scenario produces greater profits.

References

Arthur, W. B., J. Holland, B. LeBaron, R. Palmer, and P. Tayler (1997). *The Economy as an Evolving Complex System II*, Chapter Asset pricing under endogenous expectationsin an artificial stock market, pp. 15–44. Reading, MA: Addison-Wesley.

Baiman, S. and R. Verrecchia (1995). Earnings and price-based compensation contracts in the presence of discretionary trading and incomplete contracting. *Journal of Accounting and Economics 20*, 93–121.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*, 637–654.

Brock, W. and C. Hommes (1997). A rational route to randomness. *Econometrica 65*, 1059–1095.

Brock, W. and C. Hommes (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control 22*, 1235–1274.

Buchta, C. and J. Mazanec (2001). SIMSEG/ACM: A simulation environment for artificial consumer markets. Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science.

Bushman, R. and R. Indjejikian (1993). Accounting income, stock price and managerial compensation. *Journal of Accounting and Economics 16*, 3–24.

Chiarella, C. and X. He (2001). Asset pricing and wealth dynamics under heterogeneous expectations. *Quantitative Finance 1*, 509–526.

Chiarella, C. and X. He (2002). Heterogeneous beliefs, risk and learning in a simple asset pricing model. *Computational Economics 19*, 95–132.

Choe, C. (1998). A mechanism design approach to an optimal contract under ex ante and ex post private information. *Review of Economic Design 3*(3), 237–255.

Dangl, T., E. Dockner, A. Gaunersdorfer, A. Pfister, A. Soegner, and G. Strobl (2001). Adaptive erwartungsbildung und finanzmarktdynamik. *Zeitschrift für betriebswirtschaftliche Forschung 53*, 339–365.

Erev, I. and A. E. Roth (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review 88*(4), 848–881.

Flache, A. and M. W. Macy (2002). The power law of learning. *Journal of Conflict Resolution 46*(5), 629–653.

Grossman, S. (1989). *The Informational Role of Prices*. Cambridge, MA: MIT Press.

Hall, B. J. and K. J. Murphy (2002). Stock options for undiversified executives. *Journal of Accounting and Economics 33*, 3–42.

Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics 10*, 74–91.

Jaakkola, T. S., S. P. Singh, and M. I. Jordan (1995). Reinforcement learning algorithm for partially observable Markov decision problems. In G. Tesauro, D. S. Touretzky, and T. K. Leen (Eds.), *Advances in Neural Information Processing Systems*, Volume 7, pp. 345–352. The MIT Press, Cambridge.

Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy 74*, 132–157.

Levy, M. and H. Levy (1996). The danger of assuming homogeneous expectations. *Financial Analysists Journal 52*(3), 65–70.

Murphy, K. J. (1999). *Handbook of Labor Economics*, Volume 3, Chapter Executive Compensation. Amsterdam: North Holland.

Pfister, A. (2003). Heterogeneous trade intervals in an agent based financial market. Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science.

Rajgopal, S. and T. J. Shevlin (2002). Empirical evidence on the relation between stock option compensation and risk taking. *Journal of Accounting and Economics 33*(2).

Rummery, G. A. and M. Niranjan (1994). On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University.

Sallans, B., G. Dorffner, and A. Karatzoglou (2002). Feedback effects in interacting markets. In C. Urban (Ed.), *Proceedings of the Third Workshop on Agent-Based Simulation*, pp. 126–131. SCS-European Publishing House, Ghent, Belgium.

Sallans, B., A. Pfister, A. Karatzoglou, and G. Dorffner (2003). Simulation and validation of an integrated markets model. *Journal of Artificial Societies and Social Simulation 6*(4).

Simon, H. A. (1982). *Models of Bounded Rationality, Vol 2: Behavioral Economics and Business Organization*. Cambridge, MA: The MIT Press.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, pp. 1038–1044. The MIT Press, Cambridge.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.

Tesauro, G. (1999). Pricing in agent economies using neural networks and multi-agent Q-learning. In *IJCAI-99*.

Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up. *Artificial Life 8*(1), 55–82.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. Cambridge, UK: Cambridge University. Ph.D. thesis.

Watkins, C. J. C. H. and P. Dayan (1992). Q-learning. *Machine Learning 8*, 279–292.