Macroeconomies as Constructively Rational Games

Ekaterina Sinitskaya and Leigh Tesfatsion
Department of Economics, Iowa State University
Ames, Iowa 50011-1070
kate.sinitskaya@gmail.com, tesfatsi@iastate.edu

ISU Econ Department Working Paper No. 14018
Latest Revision: 20 September 2015
Published Version: *J. of Economic Dynamics and Control* 61 (2015), 152-182

Abstract

Real-world decision-makers are forced to be locally constructive; that is, their decisions are necessarily constrained by their interaction networks, information, beliefs, and physical states. This study transforms an otherwise standard dynamic macroeconomic model into an open-ended dynamic game by requiring consumers and firms with intertemporal utility and profit objectives to be locally constructive. Tested locally-constructive decision processes for the consumers and firms range from simple reactive reinforcement learning to adaptive dynamic programming (ADP). Computational experiments are used to explore macroeconomic performance under alternative decision-process combinations relative to a social planner benchmark solution. A key finding is that simpler decision processes can outperform more sophisticated decision processes such as ADP. However, memory length permitting some degree of adaptive foresight is critical for good performance.

JEL Codes: B4, C6, C7, E03, E2

Keywords: Macroeconomics; agent-based modeling; game theory; intertemporal optimization; learning; constructive rationality

1 Introduction

Decision-makers in real-world macroeconomies are necessarily limited to *locally-constructive* actions, that is, to actions constrained by their interaction networks, information, beliefs, and physical states. In contrast, the actions of agents in current macroeconomic models are typically not locally constructive because they are constrained by externally imposed coordination and optimality restrictions. Key examples include the global market clearing conditions and strong-form rational expectations postulates imposed in standard *dynamic stochastic general equilibrium (DSGE)* models based on Smets and Wouters (2003).

These observations raise the following important challenge. Suppose all actions within an otherwise standard macroeconomic model are required to be locally constructive, unsupported by externally imposed coordination and optimality restrictions. What form could these locally-constructive actions take to ensure good outcomes, not only for the individual participants but also for the macroeconomy as a whole?

This study addresses this challenge for a relatively simple macroeconomic model, referred to as the *Dynamic Macroeconomic (DM) Game*. Consumers and firms in the DM Game interact over time in labor and goods markets modeled as double auctions with uniform pricing rules. Each consumer desires to maximize his expected intertemporal (lifetime) utility subject to budget constraints, and each firm desires to maximize its expected intertemporal profit subject to technology constraints.

However, in a departure from standard macroeconomic modeling, consumers and firms in the DM Game are required to be *constructively rational* in the following sense. First, the specification by these agents of their objective functions, decision domains, and decision rules mapping decision domains into decision selections must be locally-constructive actions. Second, the successive determination of DM-Game outcomes must be a purely historical process, unaided by externally imposed coordination and optimality restrictions.

To investigate the implications of constructive rationality for the DM Game, the decision domains for consumers and firms are expressed in stationary form, as vectors of possible parameter selections. In each successive time period an agent's selection of a decision (parameter vector) maps into a sequence of parameterized supply and demand functions for current and future markets, conditional on the agent's current information, beliefs, and physical state.

Computational experiments are then conducted in which consumers and firms make successive selections from their decision domains in accordance with decision processes ranging from simple adaptation to sophisticated anticipatory learning. These decision processes include: (i) a modified version of a reactive reinforcement learning method originally developed by Roth and Erev (1995) and Erev and Roth (1998) on the basis of findings from human-subject experiments; (ii) a forward-looking learning method developed by Watkins (1989), called Q-learning; (iii) a forward-

looking rolling-horizon learning method (Alden and Smith (1992)); and (iv) an adaptive dynamic programming (ADP) learning method based on value-function approximation.

A key issue of interest is which decision-process combinations come closest to achieving the benchmark optimal solution obtainable by a fully informed social planner. In particular, do the decision processes involving relatively more sophisticated use of information tend to result in relatively higher welfare outcomes, either for the individual decision-makers or for the economy at large? Since previous experimental findings have shown that minimally-informed traders using relatively unsophisticated decision processes can match or exceed the performance of better informed traders in some market contexts (Gode and Sunder (1993); Smith (2008)), the answer to this question is not obvious *a priori*. A related issue of interest is which (if any) decision-process combinations constitute Nash equilibria and/or Pareto optimal solutions for the DM Game.

A key finding of this study is that good performance in the DM Game requires decision-makers to engage both in the exploitation of their current information and in searches for new information. Simpler decision processes can outperform more sophisticated decision processes, but only if the simpler processes entail memory lengths permitting some degree of adaptive foresight. Overall, the best performance is achieved when the consumers and firms use rolling-horizon learning methods.

This study is organized as follows. The relationship of our work to previous research is more carefully considered in Section 2, with a particular focus on learning in macroeconomic contexts. Section 3 sets out the basic structure of the DM Game together with its market and payment processes. Section 4 explains the decision processes implemented by the DM-Game consumers and firms, and Section 5 introduces and solves the social planner model used as a benchmark of comparison for our computational experiments. The sensitivity design for our computational experiments is described in Section 6, and key findings from these computational experiments are reported in Section 7. Section 8 concludes. Technical implementation aspects are relegated to the Appendix, and the code is available at https://github.com/wilfeli/DMGameBasic.

2 Relationship to Previous Research

Numerous previous researchers have emphasized the importance and complexity of modeling real-world decision processes. Examples include Simon (1978), Dosi and Egidi (1991), Stiglitz (2002), Smith (2008), Howitt (2008), Kahneman (2011), Kirman (2011), Hommes (2013), and Arthur (2015). Practitioners have also been interested in obtaining an improved understanding of these processes; see, e.g., a recent report issued by Trichet (2010), a former President of the European Central Bank.

Current macroeconomic models are surely complex. For example, standard DSGE models typically include consumers and firms that solve intertemporal utility and profit maximization problems

subject to intertemporal constraints, conditional on announced government policy rules; see, for example, Sbordone et al. (2010) and Tovar (2009). Yet, to avoid aggregation and coordination issues, these models also typically assume the existence of representative consumer and firm agents with strong forms of rational expectations. This reliance on representative agents with rational expectations has been criticized on the grounds it prevents the study of learning and coordination issues critical for understanding the operation of real-world macroeconomies (Howitt (2012)).

Recently, however, a growing number of researchers have become interested in the study of dynamic macroeconomic systems for which agents are forward-looking optimizers with incomplete knowledge about the structure of the economy. As surveyed in Honkapohja et al. (2012) and Evans and Honkapohja (2013), the standard context assumed in this literature is that a representative consumer[1] with learning capabilities resides in a dynamic world consisting of itself, a representative firm, and a government policy-maker. The representative consumer has incomplete information about the structure of its world, and it behaves as an econometrician in its attempts to learn about its world from observed data.

Specifically, the representative consumer is assumed to make consumption and labor decisions in each successive time period conditional on intertemporal budget constraints. These budget constraints depend on current state variables (e.g., financial and physical asset values), on current and forecasted future values for system variables (e.g., goods prices, wages, and interest rates), and on current and forecasted future values for government policy variables (e.g., tax rates). The consumer's system variable forecasts are obtained from a reduced-form econometric model. The consumer estimates and updates the parameters of this econometric model over time, often by means of a least-squares or Bayesian learning method. The consumer's government policy variable forecasts are generated by means of the latest announced government policy rule, assumed to be credible common knowledge.

Functional forms and calibrated maintained parameter values are specified in the initial time period to guarantee the existence of a steady-state solution, assumed to be common knowledge. A temporary equilibrium solution for the macroeconomic model is then approximately determined in differenced form (i.e., differenced from steady-state values) in each successive time period.

The approximate temporary equilibrium solution in any current time period is generally obtained as follows. Consumer and firm first-order necessary conditions for optimality are linearized around steady-state values subject to transversality, no-arbitrage, and no-ponzi-game restrictions. This permits differenced demand and supply decision variables to be expressed as linear affine functions of differenced current state variables, differenced current system variables, differenced current government policy variables, and current exogenous random shocks. For the representative consumer

[1]Some researchers assume a compact continuum of consumers exhibiting some degree of heterogeneity in their preferences for consumption versus leisure; see, e.g., Milani (2005). However, efficient risk-sharing arrangements are then typically assumed so that the consumers in fact face identical intertemporal budget constraints and behave the same in equilibrium, effectively reducing the economy to a representative consumer economy.

with forward-looking learning capabilities, the linearized expressions for its differenced decision variables also include additive terms depending on forecasted differenced values for future system variables, future government policy variables, and future shock realizations.

Period-t market clearing conditions are then imposed to obtain temporary equilibrium solutions for differenced period-t decision and system variables as linear affine functions of differenced period-t state variables and realized period-t shock terms. These solutions are used in turn to derive differenced state variable solutions for period t+1, in preparation for the determination of an approximate temporary equilibrium solution for period t+1.[2]

One key issue addressed in this literature is whether the temporary equilibrium solution path conditional on a particular learning specification exhibits convergence or escape dynamics (Honkapohja et al. (2012); Evans and Honkapohja (2013)). That is, will it converge over time to the steady-state solution (in either a global or local stability sense), or will it persistently deviate from this solution?

A second key issue is how different learning specifications affect the dynamic properties (e.g., persistence and volatility) of the temporary equilibrium solution path, taking the dynamic properties of this solution path under rational expectations as a benchmark of comparison (Milani (2005, 2007)). A third key issue, studied in Mitra et al. (2013) for a real business cycle model, is how the temporary equilibrium solution path is affected by a sudden, permanent, credibly-announced switch in the government's policy rule. A fourth key issue, explored at length in Hommes (2013), is how temporary equilibrium solutions are affected when agents are modeled as adaptive forecasters with heterogeneous beliefs and expectations.

Clearly this literature takes an important step towards more realistic macroeconomic modeling by recognizing the constrained information and computational capabilities of decision-making agents. Nevertheless, external coordination and optimality conditions are still imposed on agents (both intertemporally and cross-sectionally) in order to obtain model solutions. Examples of such conditions include: single representative consumer (or firm) assumptions; the assumed coordination of agents on a single solution; non-constructive transversality conditions; the assumed absence of interest-rate arbitrage opportunities; the assumed absence of ponzi-game opportunities such as persistent debt roll-over; and the assumed absence of excess supplies and demands in markets.

An alternative approach permitting the systematic study of locally-constructive decision processes in macroeconomic contexts without reliance on the external imposition of coordination and optimality conditions is *Agent-based Computational Economics (ACE)*. Under the ACE approach, economic processes (including whole economies) are computationally modeled as open-ended dynamic systems of interacting agents (Tesfatsion and Judd (2006); LeBaron and Tesfatsion (2008);

[2]Researchers in this dynamic macroeconomic learning literature are increasingly resorting to models expressed directly in terms of these reduced-form linear affine relationships. For example, compare the working paper (Milani, 2005) with its later published version (Milani, 2007).

Tesfatsion (2015c); Arthur (2015)). Here "agent" can refer to any physical, biological, social, or institutional entity residing within the system.

An ACE model is an *historical process model* in the following sense: Outcomes are determined in each successive time period based solely on current agent interactions, conditional on current state conditions and current exogenous shock realizations. These successive agent interactions give rise to global regularities characterizing the system as a whole, which in turn affect agent interactions.

ACE macroeconomic research to date has typically postulated decision rules for decision-making agents that are not explicitly derived as solutions for optimization problems, although they are sometimes motivated as heuristic approximations for such solutions. Examples include Oeffner (2008), Dosi et al. (2010), Mandel et al. (2010), Kirman (2011), Salle et al. (2013), Salle and Seppecher (2013), and Dawid et al. (2015).[3] This has led some macroeconomists to dismiss ACE modeling based on the incorrect belief that ACE decision-making agents must necessarily be reactive stimulus-response agents with myopic objectives.

To the contrary, however, the behaviors expressed by decision-making agents in ACE models can range all the way from simple rule-based actions to intertemporal optimization with sophisticated anticipatory learning capabilities.[4] We thus argue that it would be a Pareto improvement to expand the standard macroeconomic toolkit to include ACE as another potentially useful modeling approach.

More precisely, any modeling approach will have both advantages and disadvantages for a particular purpose at hand. For some purposes, imposing external coordination and optimality conditions on decision-making agents could be a perfectly acceptable short-cut. For other purposes it could be important to understand potential outcomes when decision-making agents are constrained to operate within a purely historical process subject to realistically rendered informational and physical limitations. The adoption of ACE modeling for these latter purposes does not require decision-makers to be "irrational".

The primary goal of the current study is to provide concrete support for the above assertions within the context of a relatively simple ACE macroeconomic model, which we refer to as the Dynamic Macroeconomic (DM) Game. As will be demonstrated more carefully in subsequent sections, the DM Game differs from existing macroeconomic models in four key respects:

(D1): Each consumer and firm in the DM Game is a learning agent with an intertemporal objective that it attempts to achieve by successive implementation of a decision process.

[3] See Chen (2012) for a survey of ACE agent modeling, and see Tesfatsion (2015a) for extensive annotated pointers to ACE macroeconomic research.

[4] For an extensive collection of annotated pointers to research on learning algorithms for ACE agents, including approximate dynamic programming and other forward-looking methods for intertemporal optimization, see Tesfatsion (2015b).

(D2): The decision process used by each learning agent in the DM Game is locally constructive.

(D3): The DM Game is an historical process model.

(D4): In the DM Game, heterogeneity in the information, beliefs, and physical states of agents changes endogenously over time through the natural course of market participations.

A final note on terminology is in order. Our conception of *constructive rationality* does not necessarily entail the pursuit of goals solely through the solution of optimization problems. Consequently, it differs from the concept of *procedural rationality* introduced by Simon (1978, p. 9), in which decision-making agents are assumed to pursue the most effective possible processes for the choices of their actions, given their limited information and cognitive powers. Similarly, it differs from the concept of *constructivist rationality* introduced by Smith (2008, p. 2), defined as "the deliberate use of reason to analyze and prescribe actions judged to be better than alternative feasible actions that might be chosen."

Rather, our conception permits *procedural uncertainty* (Dosi and Egidi (1991); Howitt (2008)), in the sense that decision-makers might be uncertain how to use their limited decision-making resources in an attempt to achieve their goals. In this case they might engage in a combined learning and decision process in an attempt to reduce their uncertainty about their world even as they attempt to survive and prosper within that world.

Indeed, the operative question for a reader of this study is as follows: If you were to be suddenly transported into the DM Game as a consumer or firm, forced to implement your decisions in a locally-constructive manner, what decision process would you use in an attempt to achieve your utility or profit goal?

3 The Dynamic Macroeconomic Game

3.1 Overview

This section develops the *Dynamic Macroeconomic (DM) Game*, a relatively simple dynamic macroeconomic model that will permit us to investigate the effects on micro and macro outcomes when consumers and firms with intertemporal utility and profit goals implement various types of decision processes in an attempt to achieve these goals. The basic structure of the DM Game is similar to the structure of standard dynamic macroeconomic models. However, as noted in Section 2, the DM Game differs from these standard models in four key respects: (D1) multiple consumers and firms with learning capabilities; (D2) locally-constructive decision rules; (D3) absence of externally-imposed coordination and optimality conditions; and (D4) endogenous heterogeneity.

Conditions (D1) through (D3) imply that events must proceed through historical time from cause to effect, with no non-causal looping permitted. In particular, the standard determination of market outcomes, in which labor and goods markets are simultaneously cleared at correct equilibrium prices with correct matching of buyers and sellers, with no risk to the traders, must be replaced by market processes permitting risky trades to proceed even if transactions are based on imperfectly informed demands and supplies.

Regarding (D4), heterogeneity among the DM-Game consumers and among the DM-Game firms arises endogenously over time from three sources. One source is that all of the decision processes tested for consumers and firms in this study are adaptive processes involving stochastic aspects in their implementations. A second source is that consumers and firms use "coin flips" to resolve indifference among decision options. A third source is that the rules governing labor and goods market operations include stochastic rationing rules to resolve excess demand and supply situations.

Section 3.2 provides a big-picture understanding of the basic DM-Game structure. Sections 3.3 through 3.5 then explain in greater detail the market and payment processes in the DM Game, as well as the structure of the intertemporal optimization problems for consumers and firms. A detailed description of the particular locally-constructive decision processes to be tested for the consumers and firms by means of computational experiments is given in the following Section 4.

3.2 Basic DM-Game Structure

As depicted in Fig. 1, the DM Game consists of a finite number I of utility-seeking infinitely-lived consumers and a finite number J of profit-seeking infinitely-lived corporate firms that interact in market and payment processes over discrete time periods $t \geq 0$, where period $t = [t, t+1)$.

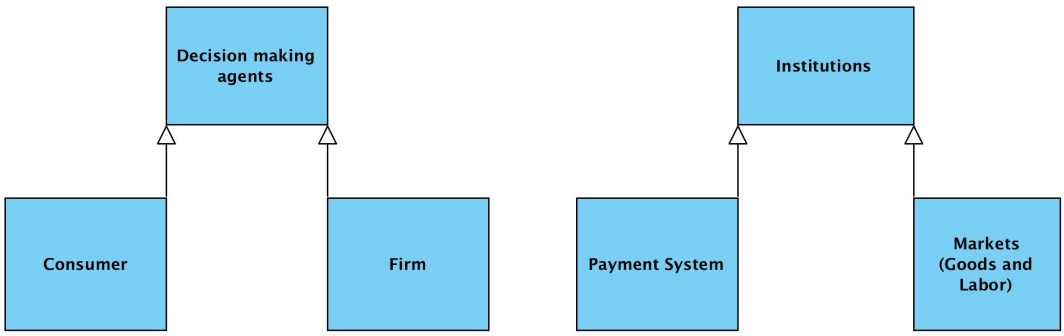

Figure 1: Decision-making agents and institutions for the DM Game

Each consumer and firm has an initial money balance at time 0, measured in book credit; and all subsequent payments and receipts take the form of changes in consumer and firm money balances. The consumers derive utility from leisure and from the consumption of a durable good q purchased from firms. The firms earn profits from the sale of good q to consumers, where q is produced by means of labor services purchased from consumers.

Both the labor market and the goods market are organized as automated double-auction exchanges in which demands and supplies are matched to determine market-clearing prices and quantities. Firm profits are distributed back to consumers in the form of dividend payments. The goal of each consumer is to maximize his expected intertemporal utility subject to budget constraints, where this optimization problem is expressed in locally-constructive terms. The goal of each firm is to maximize its expected intertemporal profits subject to technology constraints, where this optimization problem is expressed in locally-constructive terms.

Each consumer at time 0 is structurally identical to each other consumer; that is, each consumer has the same initial money balance, human capital endowment, and intertemporal utility function. Also, each consumer owns an equal share of each firm, fixed through time, and hence receives the same stream of dividend payments. Similarly, each firm at time 0 is structurally identical to each other firm, meaning that each firm has the same initial money balance, goods stock, production function, intertemporal profit function, and dividend allocation rule.

Market trades in the DM Game are risky in the following sense. In each period the labor market occurs prior to the goods market. Firms engage in forward contracting with consumers for labor services, and carry out goods production using these labor services, prior to the realization of actual goods demands. Firms thus risk bankruptcy if insufficient goods are sold to permit them to meet their wage obligations, and bankrupt firms must exit the DM-Game economy. Since there is no entry mechanism for firms, the bankruptcy of firms can ultimately lead to the collapse of the economy.

When firms are forced to exit the DM Game due to bankruptcy, the remaining firms do not immediately modify their behavior to take into account that they now have a larger share of the market. However, as will be seen in Section 4, all of our tested decision processes involve adaptation to changing state conditions. Consequently, the exit of bankrupt firms will eventually result in changes in the decisions of the remaining firms to the extent that this bankruptcy affects their state conditions.

Consumers risk non-payment for labor services rendered if firms become bankrupt. Since all goods demands must be backed by actual purchasing power, this can reduce the goods demands of the consumers in the next trading period, exacerbating firm cash-flow problems. However, consumers can survive even if their market purchases of consumption goods are zero because they can obtain their basic subsistence needs through non-traded means (e.g., a garden patch).

A key question to be addressed is therefore as follows. Given the potential riskiness of market trading, and the restriction to locally-constructive decision processes, is it worthwhile for the consumers and firms to use relatively sophisticated decision processes derived from intertemporal optimizations? Or should they instead proceed cautiously with simpler forms of decision processes based on incremental adaptations to past trading outcomes?

3.3 Market and Payment Processes in the DM Game

All transactions in the DM Game are accompanied by corresponding payments, hence the payment system is an important underlying institution. For simplicity, this payment system is taken to be a simple clearing house that instantaneously clears each transaction. Although consumers and firms can carry forward savings in the form of money (book-credit), there is no banking system, hence no borrowing/lending opportunities and no interest paid on savings.

A consumer is not permitted to spend more than his current money balance, hence all consumer demands for goods must be backed by actual purchasing power. A firm is declared bankrupt, and removed from the economy, if its current money balance is insufficient to meet its wage-payment obligations to its workers.[5]

The consumers and firms use decision processes in each period t in an attempt to take actions that satisfy their intertemporal utility and profit goals, conditional on current expectations for future wages and goods prices. These actions consist of both labor and goods market decisions, such as whether or not to participate in these markets and what specific quantity and price terms to seek if they do.

The consumers and firms receive feedback from the economy as a result of their period-t actions, and they update their decision processes on the basis of this feedback in preparation for period $t+1$. This feedback includes market-clearing wages and prices for the period-t labor and goods markets, and their own private utility or profit outcomes as a result of their period-t market transactions.

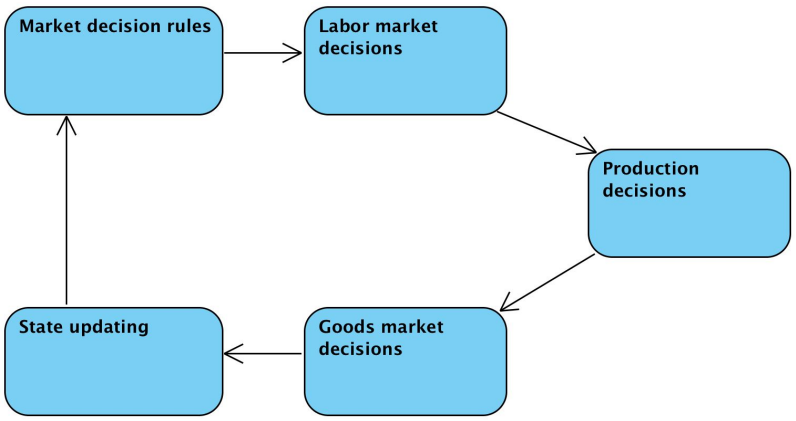

Figure 2: Sequential market decisions during a typical period t.

As depicted in Fig. 2, the labor market occurs before the goods market in each period t. Each consumer participating in the labor market submits a labor supply offer, and each firm participating in the labor market submits a labor demand bid. A labor market clearing solution is then calculated based on these offers and bids. This solution consists of a set of forward labor contracts (supply

[5]Any money held by a bankrupt firm is divided equally among its workers in partial fulfillment of its wage-payment obligations. However, goods stocks of bankrupt firms are assumed to be lost to the economy.

now, get paid later) that determine the amount of labor to be supplied now by each consumer to each firm, and the (common) wage to be paid later by the firms to the consumers for each unit of supplied labor.

After the close of the period-t labor market, the consumers perform labor for the firms in accordance with their forward labor contracts, which results in produced amounts of goods. Next, each consumer participating in the period-t goods market submits a goods demand bid, and each firm participating in the period-t goods market submits a goods supply offer. A goods market clearing solution is then calculated based on these bids and offers. This solution consists of a set of spot contracts that determine the amount of good to be received now by each consumer from each firm, and the (common) goods price to be paid now by the consumers to the firms for each unit of good received.

After the close of the period-t goods market, each firm proceeds to deliver goods to its customers in return for goods payments in accordance with its period-t goods market spot contracts. Each firm then settles its period-t wage-payment obligations to its workers as determined by its period-t forward labor contracts, if it has a sufficient money balance to cover these obligations. Otherwise, the firm is bankrupt and must exit the economy.

At the end of period t, each consumer calculates its period-t utility on the basis of its period-t consumption of market-procured goods and leisure. Also, each (non-bankrupt) firm calculates its period-t profit as its period-t goods-sales revenues minus its period-t wage payments. These period-t utility and profit outcomes are used by the consumers and firms to update their decision processes for period $t+1$.

A portion of any positive profits accrued by a firm during period t is distributed to the firm's consumer-owners as dividend payments at the end of period t. The wage and dividend payments received by a consumer from the firms at the end of period t, together with any other unspent monies held by the consumer at the end of period t, constitute the money balance of the consumer at the start of period $t+1$ to be used for goods purchases in period $t+1$.

This flow of events is illustrated in Fig. 3. Note the use of internal times t:1 through t:6 for events occurring within each period $t = [t, t+1)$. The money balances held by consumers and firms at the end of period t (i.e., at time $t+1$) are determined by the money balances held by consumers and firms at the start of period t together with the additions and subtractions to these money balances arising from period-t market and dividend payments.

Finally, as detailed below in Sections 4.2 and 4.3, reservation wages and prices are used to determine demand and supply functions in the DM Game. Agents thus abruptly enter or drop out of the labor and goods markets as the wage and price increase from 0, which induces vertical and horizontal portions in the aggregate demand and supply functions.

If the aggregate demand and supply functions coincide along a vertical portion, there will be

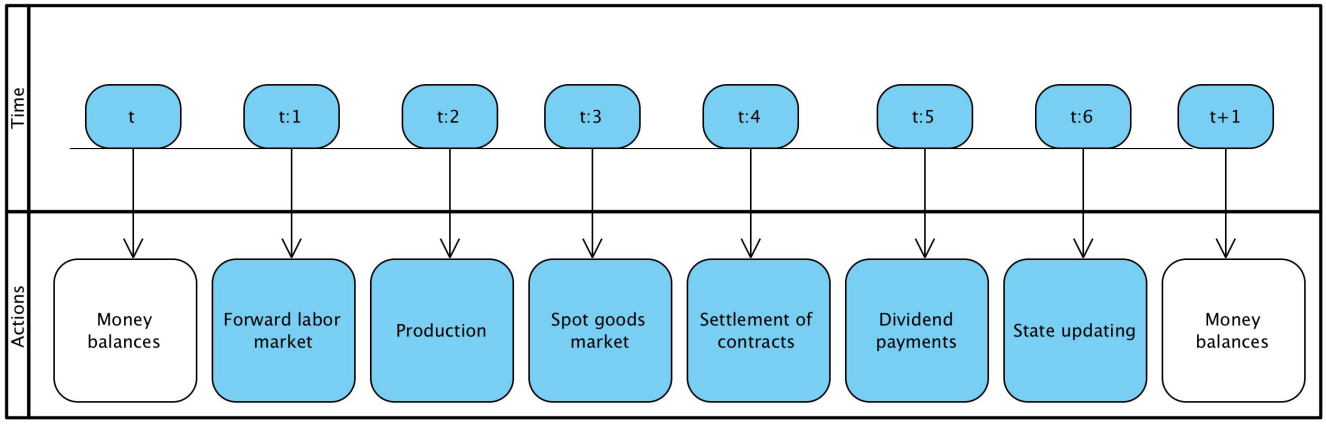

Figure 3: Flow of events during a typical period t.

multiple possible equilibrium prices corresponding to a unique equilibrium quantity. In this case, market rules impose the selection of the maximum possible equilibrium price.

If the aggregate demand and supply functions coincide along a horizontal portion, there will be multiple possible equilibrium quantities corresponding to a unique equilibrium price. In this case, market rules impose a simple stochastic rationing mechanism: namely, the agents that are willing to trade at this unique equilibrium price are allowed to trade in random order. Trading stops when no more trades are possible, at which point the maximum possible equilibrium quantity has been cleared. At such equilibrium points there will typically be traders willing but unable to purchase more goods (excess demand) or traders willing but unable to supply more goods (excess supply).

3.4 Consumer Constraints and Goals in the DM Game

Consumers in the DM Game are structurally identical at the initial time 0. Each consumer i is endowed with the same initial positive money balance M^c_{-1} (in book credit form). Each consumer i also has one unit of time in each period $r \geq 0$ that can be divided between labor $l^c_{i,r:1}$ and leisure $[1 - l^c_{i,r:1}]$. To simplify the analysis, it is assumed that this one unit of time is allocated either all to labor or all to leisure.

Ignoring uncertainties (for the moment), the budget constraints faced by each consumer i in each period $r \geq 0$ take the following form:

$$s_{i,r:3} = M^c_{i,r-1} - p_{r:3} q^c_{i,r:3} \tag{1}$$

$$M^c_{i,r} = s_{i,r:3} + w_{i,r:4} l^c_{i,r:1} + div^c_{r:5} \tag{2}$$

$$s_{i,r:3},\ q^c_{i,r:3} \ \geq \ 0 \tag{3}$$

$$l^c_{i,r:1} \in \{0, 1\} \tag{4}$$

Here $M^c_{i,r-1}$ denotes consumer i's money balance at the start of period r, $p_{r:3}$ denotes the goods price

determined in the goods market at time r:3 (same for all consumers), $q^c_{i,r:3}$ denotes the amount of good purchased by consumer i in the goods market at time r:3, $s_{i,r:3}$ denotes the savings of consumer i immediately subsequent to the goods market at time r:3, $w_{i,r:4}l^c_{i,r:1}$ denotes the actual wage payment received by consumer i at time r:4 arising from its forward labor contract cleared in the labor market at time r:1, and $div^c_{r:5}$ denotes the dividend payment (same for all consumers) received by consumer i at time r:5. The non-negativity constraint $s_{i,r:3} \geq 0$ ensures that consumer i's goods purchase $q^c_{i,r:3}$ is backed by actual purchasing power (money holdings).

The goal of each consumer i at the start of each period $t \geq 0$ is to maximize his expected intertemporal utility over periods $r \geq t$ subject to budget constraints (1)-(4) for periods $r \geq t$. If the labor service and consumption levels of consumer i in periods $r \geq t$ are given by $\left\{l^c_{i,r:1}, q^c_{i,r:3}\right\}_{r=t}^{\infty}$, then the intertemporal utility attained by consumer i over periods $r \geq t$ is given by

$$U_{i,t} = \sum_{r=t}^{\infty} \beta^{r-t} u\left(q^c_{i,r:3}, 1 - l^c_{i,r:1}\right) \ , \tag{5}$$

where $\beta \in (0,1)$ is a time-preference discount parameter.

In summary, as detailed above, the constraints and goals of the consumers in the DM Game depend commonly on the specific settings for $(M^c_{-1}, u(\cdot), \beta)$ at the initial time 0. However, consumers do not know in advance the decision processes in use by firms and other consumers, hence they do not know in advance the market-clearing values for future goods prices and wages nor the extent to which their own future goods demands and labor supplies will be fulfilled. How each consumer i might address this uncertainty through various alternative specifications for its own locally-constructive decision process will be explained in Section 4.

3.5 Firm Constraints and Goals in the DM Game

Firms in the DM Game are structurally identical at the initial time 0. Each firm j is endowed with the same initial positive money balance M^f_{-1} (in book credit form) and the same initial goods stock q^{stock}_{-1}. Also, each firm j has the same stationary production function $q = F(l)$ for the production of good q using labor services l. Ignoring uncertainties (for the moment), the constraints faced by each firm j in each period $r \geq 0$ are derived as follows.

Let $q^{stock}_{j,r-1}$ denote firm j's goods inventory at the start of period $r \geq 0$. Suppose firm j purchases labor services $l^f_{j,r:1}$ in the time-r:1 labor market and uses these labor services to produce a goods amount $q^f_{j,r:2} = F(l^f_{j,r:1})$ at time r:2. The goods amount $q^f_{j,r:3}$ that firm j sells in the time-r:3 goods market cannot exceed its time r:2 goods inventory, $q^{stock}_{j,r:2}$, which is given by its goods inventory at the start of period r plus its time r:2 goods production $q^f_{j,r:2}$:

$$q^{stock}_{j,r:2} \ = \ q^{stock}_{j,r-1} \ + \ q^f_{j,r:2} \ \geq \ q^f_{j,r:3} \tag{6}$$

Firm j's goods inventory $q_{j,r}^{stock}$ at the start of period $r+1$ is then determined from the following inventory accumulation equation:

$$q_{j,r}^{stock} \;=\; q_{j,r:2}^{stock} \;-\; q_{j,r:3}^{f} \tag{7}$$

In addition, firm j must worry about avoiding bankruptcy, since bankrupt firms (i.e., firms unable to meet their wage obligations) must exit the DM-Game economy. Consequently, firm j only distributes dividends in period r if its goods market revenues $p_{r:3}q_{j,r:3}^{f}$ earned at time r:3 exceed its wage obligations $w_{j,r:1}l_{j,r:1}^{f}$ incurred in the forward labor market at time r:1 for settlement at time r:4. Moreover, firm j limits its dividend distributions to its profits (if any). Specifically, firm j's total dividend payments $div_{j,r:5}^{f}$ at time r:5 are determined in accordance with the following allocation rule:

$$div_{j,r:5}^{f} = \begin{cases} \kappa^{div} \cdot \left[p_{r:3}q_{j,r:3}^{f} - w_{r:1}l_{j,r:1}^{f}\right] & \text{if } \; p_{r:3}q_{j,r:3}^{f} - w_{r:1}l_{j,r:1}^{f} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\kappa^{div} \in [0,1]$. Given (8), the no-bankruptcy condition for firm j in period r guaranteeing its period-r wage obligations can be fulfilled takes the form

$$M_{j,r-1}^{f} \;+\; p_{r:3}q_{j,r:3}^{f} \;-\; w_{r:1}l_{j,r:1}^{f} \;\geq\; 0 \tag{9}$$

The money balance $M_{j,r}^{f}$ held by a non-bankrupt firm j at the end of period r (i.e., at the start of period $r+1$) is determined by the money balance $M_{j,r-1}^{f}$ held by firm j at the start of period r adjusted to reflect firm j's market activities and dividend payments during period r, as follows:

$$M_{j,r}^{f} \;=\; M_{j,r-1}^{f} + p_{r:3}q_{j,r:3}^{f} - w_{r:1}l_{j,r:1}^{f} - div_{j,r:5}^{f} \tag{10}$$

Finally, the following non-negativity restrictions on firm j's labor demand $l_{j,r:1}^{f}$ at time r:1 and goods supply $q_{j,r:3}^{f}$ at time r:3 must be satisfied for physical meaningfulness:

$$l_{j,r:1}^{f},\; q_{j,r:3}^{f} \;\geq\; 0 \tag{11}$$

The goal of each firm j at the start of each period $t \geq 0$ is to maximize its expected intertemporal profit over periods $r \geq t$ subject to the technological and feasibility constraints (6)-(11) for periods $r \geq t$. For any given sequence $\left\{w_{r:1}, l_{j,r:1}^{f}, p_{r:3}, q_{j,r:3}^{f}\right\}_{r=t}^{\infty}$ of wage levels, labor service purchases, goods prices, and goods purchases for periods $r \geq t$, the intertemporal profit attained by firm j

over periods $r \geq t$ is given by

$$\Pi_{j,t} = \sum_{r=t}^{\infty} \mu^{r-t} \left[p_{r:3} q^f_{j,r:3} - w_{r:1} l^f_{j,r:1} \right] \tag{12}$$

where $\mu \in (0,1)$ is a time-preference discount parameter.

In summary, as detailed above, the constraints and goals of the firms in the DM Game depend commonly on the specific settings for $(M^f_{-1}, q^{stock}_{-1}, F(\cdot), \mu, \kappa^{div})$ at the initial time 0. However, firms do not know in advance the decision processes in use by consumers and other firms, hence they do not know in advance the market-clearing values for wages and goods prices nor the extent to which their own future labor supplies and goods demands will be fulfilled. How each firm j might address this uncertainty through various alternative specifications for its own locally-constructive decision process will be explained in the following Section 4.

4 Locally-Constructive Decision Processes

4.1 Overview of Decision Processes

The locally-constructive decision processes specified for consumers and firms in the DM Game are procedures for the adaptive determination of demand bids and supply offers for the labor and goods markets in each successive period t. These decision processes consist of three distinct components, as follows.

First, *decision domains* are specified for consumers and firms that consist of possible selections of "tuning" parameters for demand and supply functions. To permit more meaningful comparisons among decision processes, the decision domain for each consumer at the start of each period t is specified as a cross-product D^c of finite sets, the same for each consumer. Similarly, the decision domain for each firm at the start of each period t is specified as a cross-product D^f of finite sets, the same for each firm.

Second, *state-conditioned transformation functions* are specified for consumers and firms. The state of a consumer or firm at any time t consists of the time-t information, beliefs, and physical attributes of this agent. The transformation function for each consumer at the start of each period $t \geq 0$ maps each of his possible decisions d^c in D^c into a collection of labor supply and goods demand functions for periods $r \geq t$, parameterized by d^c, and conditional on the consumer's time-t state. Similarly, the transformation function for each firm at the start of each period $t \geq 0$ maps each of its possible decisions d^f in D^f into a collection of labor demand and goods supply functions for periods $r \geq t$, parameterized by d^f, and conditional on the firm's time-t state.

Third, *Reactive Learner (RL)*, *Forward-looking Learner (FL)*, and *Explicit Optimizer (EO) decision rules* are specified for each consumer and firm that determine how this agent selects decisions from its decision domain in each period t. These three types of decision rules cover a range of decision-making behaviors roughly ordered from less to more sophisticated with regard to information utilization, expectation formation, and forward-looking behavior. A summary description of these decision-rule types is given in Table 1.

Agent	**Decision-Rule Type**	**Decision-Rule Description**
Consumer	Reactive Learner (RL)	Adaptively updates decisions in response to realized utility outcomes
	Forward-Looking Learner (FL)	Uses Q-learning in an attempt to maximize expected intertemporal utility
	Explicit Optimizer (EO)	Maximizes expected intertemporal utility using adaptively updated probabilities
Firm	Reactive Learner (RL)	Adaptively updates decisions in response to realized profit outcomes
	Forward-Looking Learner (FL)	Uses Q-learning in an attempt to maximize expected intertemporal profit
	Explicit Optimizer (EO)	Maximizes expected intertemporal profit using adaptively updated probabilities

Table 1: Types of decision rules for consumers and firms in the DM Game.

The construction of the decision domains and the state-conditioned transformation functions for consumers and firms is explained more carefully in Section 4.2 and Section 4.3. Detailed descriptions of the three decision-rule types RL, FL, and EO listed in Table 1 are provided in Section 4.4 through Section 4.6.

4.2 Decision Domain and Transformation Function for Consumers

The decision domain D^c for each consumer i at the start of each period t is given by a cross-product of finite sets having the form

$$D^c \;=\; L^c \otimes \Omega \otimes \Theta \tag{13}$$

where:

- $L^c = \{0, 1\}$
- the elements of $\Omega = \{\omega_1, \ldots, \omega_G\}$ satisfy $0 < \omega_1 < \ldots < \omega_G$
- the elements of $\Theta = \{\theta_1, \ldots, \theta_H\}$ satisfy $0 \leq \theta_1 < \ldots < \theta_H \leq 1$

Consumer i selects a decision $d = (l^c, \omega, \theta)$ from D^c at each time $t \geq 0$ by means of its particular RL, FL, or EO decision rule. The selection of d at time t is then transformed into a sequence $\mathbf{TR}^c_{i,t}(d)$ of labor supply and goods demand functions $(l^c_{i,r:1}(w,d,t), q^c_{i,r:3}(p,d,t))_{r\geq t}$, parameterized by d and conditional on consumer i's time-t state.

Specifically, the labor supply $l^c_{i,r:1}(w,d,t)$ as a function of the time-r:1 labor market wage w is determined as follows. If $l^c = 0$, then $l^c_{i,r:1}(w,d,t) = 0$ for all wages w, meaning that consumer i does not plan to participate in the time-r:1 labor market. On the other hand, if $l^c = 1$, the *reservation wage* of consumer i for the time-r:1 labor market, calculated from the vantage point of the current time t, is given by

$$w^c_{i,r:1}(d,t) \;=\; \omega \cdot E_{i,t}[w_{r:1}] \tag{14}$$

where $E_{i,t}[w_{r:1}]$ denotes the time-r:1 labor market wage expected by consumer i, based on his state at time t.[6] The parameter ω in (14) determines the scale of each consumer's reservation wage relative to his expected wage.

The reservation wage $w^c_{i,r:1}(d,t)$ in (14) has the standard meaning that it is the lowest wage that consumer i expects at time t to be willing to accept for his offered labor at time r:1. If $w < w^c_{i,r:1}(d,t)$, then $l^c_{i,r:1}(w,d,t) = 0$, meaning that consumer i does not plan to participate in the time-r:1 labor market at the labor market wage w. On the other hand, if $w \geq w^c_{i,r:1}(d,t)$, then $l^c_{i,r:1}(w,d,t) = 1$, meaning that consumer i plans to offer his 1 unit of labor into the time-r:1 labor market at the labor market wage w.

Also, the goods demand $q^c_{i,r:3}(p,d,t)$ as a function of the time-r:3 goods market price p takes the form

$$p \cdot q^c_{i,r:3}(p,d,t) \;=\; \theta \cdot M^c_{i,r-1} \tag{15}$$

Thus, consumer i plans in period t to spend a fraction θ of his time-r money balance $M^c_{i,r-1}$ on consumption goods at time r:3, and he specifies his time-r:3 goods demand as a function of the time-r:3 market price p in accordance with this plan. Note that $M^c_{i,r-1}$ will be known to consumer i at time r, prior to the opening of the goods market at time r:3.[7]

The decision domain D^c depends on the grid specifications for Ω and Θ; these grid specifications are explained in Appendix A. The transformation function $\mathbf{TR}^c_{i,t}$ depends on the wage expectation in (14). The method used by consumers to form and update their wage expectations is explained in Appendix B.

[6] Without loss of generality, the reservation wage (14) could be expressed in real terms by dividing each side of (14) by the expected goods price at time r:1, where this price expectation is formed at the current time t.

[7] Recall that consumer i receives no money payments between time r (the start of period r) and the settlement of labor market contracts at time r:4. Thus, consumer i's purchases in the time-r:3 goods market cannot exceed his money balance $M^c_{i,r-1}$ at time r.

4.3 Decision Domain and Transformation Function for Firms

The decision domain D^f for each firm j at the start of each period t is given by a cross-product of finite sets having the form

$$D^f \;=\; L^f \otimes \Gamma \otimes \Lambda \otimes \Psi \tag{16}$$

where:

- the elements of $L^f = \{l_1^f, \ldots, l_L^f\}$ satisfy $0 \leq l_1^f < \ldots < l_L^f$
- the elements of $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$ satisfy $0 < \gamma_1 < \ldots < \gamma_M$
- the elements of $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ satisfy $0 < \lambda_1 < \ldots < \lambda_N$
- the elements of $\Psi = \{\psi_1, \ldots, \psi_R\}$ satisfy $0 \leq \psi_1 < \ldots < \psi_R \leq 1$

Firm j selects a decision $d = (l^f, \gamma, \lambda, \psi)$ from D^f at each time $t \geq 0$ by means of its particular RL, FL, or EO decision rule. The selection of d at time t is then transformed into a sequence $\mathbf{TR}_{j,t}^f(d)$ of labor demand and goods supply functions $(l_{j,r:1}^f(w,d,t), q_{j,r:3}^f(p,d,t))_{r \geq t}$, parameterized by d and conditional on firm j's time-t state.

Specifically, the labor demand $l_{j,r:1}^f(w,d,t)$ as a function of the time-r:1 labor market wage w is determined as follows. If $l^f = 0$, then $l_{j,r:q}^f(w,d,t) = 0$ for all w, meaning that firm j does not plan to participate in the time-r:1 labor market. If $l^f > 0$, the *reservation wage* of firm j for the time-r:1 labor market is given by

$$w_{j,r:1}^f(d,t) \;=\; \gamma \cdot E_{j,t}[w_{r:1}] \tag{17}$$

where $E_{j,t}[w_{r:1}]$ denotes the time-r:1 labor market wage expected by firm j, based on its state at time t.[8] The parameter γ in (17) determines the scale of firm j's reservation wage relative to its expected wage.

The reservation wage (17) has the standard meaning that it is the highest wage that firm j expects at time t to be willing to pay for its demanded labor at time r:1. If $w > w_{j,r:1}^f(d,t)$, then $l_{j,r:1}^f(w,d,t) = 0$, meaning that firm j does not plan to participate in the time-r:1 labor market at the labor market wage w. On the other hand, if $w \leq w_{j,r:1}^f(d,t)$, then $l_{j,r:1}^f(w,d,t) = l^f$, meaning that firm j plans to demand l^f units of labor in the time-r:1 labor market at the labor market wage w.

Also, the goods supply $q_{j,r:3}^f(p,d,t)$ as a function of the time-r:3 goods market price p is determined as follows. The *reservation goods price* of firm j for the time r:3 goods market is given by

$$p_{j,r:3}^f(d,t) = \lambda \cdot E_{j,t}[p_{r:3}] \tag{18}$$

[8] Without loss of generality, the reservation wage (17) could be expressed in real terms by dividing each side of (17) by the expected goods price at time r:1, where this price expectation is formed at the current time t.

where $E_{j,t}[p_{r:3}]$ denotes the time-r:3 goods market price expected by firm j, based on its state at time t. The parameter λ in (18) determines the scale of each firm's reservation goods price relative to its expected goods price.

The reservation price (18) has the standard meaning that it is the lowest price that firm j expects at time t to be willing to accept for its supplied goods at time r:1. If $p < p^f_{j,r:3}(d,t)$, then $q^f_{j,r:3}(p,d,t) = 0$, meaning that firm j does not plan to participate in the time-r:3 goods market at the goods market price p. On the other hand, if $p \geq p^f_{j,r:3}(d,t)$, then

$$q^f_{j,r:3}(p,d,t) \;=\; \psi \cdot q^{stock}_{j,r:2} \tag{19}$$

That is, firm j plans to supply a fraction ψ of its time-r:2 goods stock into the time-r:3 goods market at the goods market price p. The parameter ψ in (19) determines the scale of each firm's goods supply relative to its current stock of goods. Note that $q^{stock}_{j,r:2}$ will be known to firm j at time r:2, prior to the opening of the goods market at time r:3.

The decision domain D^f depends on the grid specifications for L^f, Γ, Λ, and Ψ; these grid specifications are explained in Appendix A. The transformation function $\mathbf{TR}^f_{j,t}$ depends on the wage expectation in (17) and the price expectation in (18). The method used by firms to form and update their wage and price expectations is explained in Appendix B.

4.4 RL Decision Rule for Consumers and Firms

Reinforcement learning embodies the basic common-sense principle that the propensity to select relatively good decisions should be reinforced and the propensity to select relatively poor decisions should be discouraged. Immediate rewards flowing from decisions are typically used to update the propensities for choosing these decisions in an appropriate up or down direction.

The *reinforcement learning (RL) decision rule* used for consumers and firms in the DM Game is an RL algorithm developed by Nicolaisen et al. (2001). This algorithm is referred to as Modified Roth-Erev Reinforcement Learning (MRE-RL) because it introduces modifications to correct for two potentially problematic aspects of an RL algorithm originally developed by Roth and Erev (1995) and Erev and Roth (1998).[9] The RL decision rule is *reactive* in the sense that it asks the following backward-looking question: Given past events, what decision should I make now?

For the DM Game, the immediate reward $R^c_i(d,t)$ received by a consumer i as a result of selecting a decision d in D^c at the start of any period t is taken to be consumer i's realized period-t utility. Similarly, the immediate reward $R^f_j(d,t)$ received by a firm j as a result of selecting a decision d in D^f at the start of any period t is taken to be firm j's realized period-t profit.

[9] As detailed in Nicolaisen et al. (2001), the two problematic aspects of the original Roth-Erev RL algorithm are solution degeneracy for some parameter configurations and no updating of relative choice propensities in response to zero-reward outcomes.

Below we explain the RL decision rule for an arbitrary decision-maker v who selects a decision d from a finite decision domain D in each period t, receiving an immediate reward $R(d,t)$, where v could represent either a consumer or a firm in the DM Game. Let the finite cardinality of D be denoted by $\mathcal{D}$, and let the elements of D be indexed by $d = 1, \ldots, \mathcal{D}$.

Suppose it is the start of the initial period 0, prior to decision selection, and suppose decision-maker v must select a decision from its decision domain D for period 0. Suppose the *initial propensity* of v to select decision d in D at time 0 is exogenously given by $q(d,0)$ for $d = 1, \ldots, \mathcal{D}$. Let the vector of these initial propensities be denoted by $\mathbf{q}(0) = (q(1,0), \ldots, q(\mathcal{D},0))$.

Now suppose it is the start of any period $t \geq 0$, prior to decision selection, and suppose the current propensity of decision-maker v to select decision d in D is given by $q(d,t)$ for $d = 1, \ldots, \mathcal{D}$. The *choice probabilities* that v uses to select a decision for period t are then constructed from these propensities as follows:

$$\text{Prob}(d,t) = \frac{\exp(q(d,t)/C)}{\sum_{k=1}^{\mathcal{D}} \exp(q(k,t)/C)} \ , \ d = 1, \ldots, \mathcal{D} \tag{20}$$

In (20), C is a *cooling parameter* that affects the degree to which v makes use of propensity values in determining his choice probabilities. As $C \to \infty$, then $\text{Prob}(d,t) \to 1/\mathcal{D}$, so that in the limit v pays no attention to propensity values in forming his choice probabilities. On the other hand, as $C \to 0$, the choice probabilities (20) become increasingly peaked over the particular decisions d having the highest propensity values $q(d,t)$, thereby increasing the probability that these decisions will be chosen by v.

At the end of period t, the current propensity $q(d,t)$ that decision-maker v associates with each decision d in D is updated in accordance with the following rule. Let d_t in D denote the decision that v *actually* selected and implemented during period t. Also, let $R(d_t,t)$ denote the reward attained by v at the end of period t as a result of the implementation of d_t. Then, for each decision d in D,

$$q(d,t+1) \ = \ [1-\rho]q(d,t) \ + \ \text{Response}(d,t) \ , \tag{21}$$

where

$$\text{Response}(d,t) = \begin{cases} [1-e] \cdot R(d_t,t) & \text{if } \ d = d_t \\ \\ e \cdot q(d,t)/[\mathcal{D}-1] & \text{if } \ d \neq d_t \end{cases} \tag{22}$$

Note $d \neq d_t$ implies $\mathcal{D} \geq 2$. The *recency parameter* $\rho \in [0,1]$ appearing in (21) controls the relative weighting of past versus current rewards in the updating of the propensities. The *experimentation parameter* $e \in [0,1)$ appearing in (22) permits reinforcement to spill over from a chosen decision to other decisions to encourage experimentation with various decisions in the early stages of the

learning process.[10]

In summary, the RL decision rule is fully characterized once values are specified for the vector of parameter values $(\mathcal{D}, \mathbf{q}(0), C, e, \rho)$. Note that the RL decision rule is well-defined for any decision domain with finite cardinality $\mathcal{D}$; the exact form of the decisions constituting this decision domain is irrelevant. Note, also, that the decision-maker does not need to know his reward function; the RL decision rule only makes use of realized rewards, not potential rewards. The versatility and low-information requirements of the RL decision rule, together with its demonstrated robust performance in diverse situations, have led to its widespread use in learning applications.

4.5 FL Decision Rule for Consumers and Firms

The *forward-looking (FL) decision rule* used for consumers and firms in the DM Game is a "greedy" variant of the Q-learning algorithm developed by Watkins (1989) that permits decisions to be taken in accordance with dynamic programming policy functions in approximate form. The FL decision rule is *forward looking* in the sense that it asks the following anticipatory question: If I make this decision now, what will happen in the future?

The key conceptual construct underlying Q-learning (and stochastic dynamic programming in general) for a decision-maker v is the *value function* $V_t(x)$, defined to be the optimum expected total reward that can be obtained by v, starting at time t in state x. An important derived conceptual construct is then the *policy function* expressing the optimal decision for v as a function of the time t and state x. Below we provide an intuitive derivation of ϵ-greedy Q-learning as a policy-function approximation method, without consideration of technical details regarding the existence and uniqueness of optimal solutions.

Suppose a decision-maker v is in state x at some current time t. Suppose v implements a decision d, observes a random event realization ω, obtains an immediate reward $R_t(x, d, \omega)$, and transits to a new state $x' = S_t(x, d, \omega)$. Then the best that v can do, starting from time $t+1$, is $V_{t+1}(x')$. Consequently, letting $E[\cdot]$ denote expectation with respect to the random event ω, the best that v

[10] The use of $q(d;t)$ in the response (22) in place of $R(d_t, t)$ is a key modification of the original Roth-Erev RL algorithm that was introduced by Nicolaisen et al. (2001) in order to correct a potentially serious "zero reward-zero updating" feature of the original algorithm. With $R(d_t, t)$ in place of $q(d;t)$ in (22), if a selected action d_t results in a reward $R(d_t, t) = 0$, then the Response(d, t) in (22) is zero. This implies that the choice propensities $q(d, t)$ in (21) are uniformly changed by a factor of $[1-\rho]$, hence their *relative* sizes are unchanged. As reported in Nicolaisen et al. (2001), this failure to update the relative sizes of the choice propensities in response to zero-reward outcomes can result in a substantial loss of market efficiency in auction markets because participants whose bids and offers fail to clear do not learn from their mistakes. The use of $q(d;t)$ in (22) ensures that any selected action d_t with a positive propensity that results in a reward $R(d_t;t) = 0$ will have its propensity (hence choice probability) reduced relative to other actions with positive propensities, thus encouraging the decision-maker to move away from zero-reward actions towards better actions.

can do, starting in state x at time t, is

$$V_t(x) = \max_d E\left[R_t(x,d,\omega) + \beta V_{t+1}(S_t(x,d,\omega))\right] \tag{23}$$

Finally, let $d^*(t,x)$ denote the *optimal policy function* expressing the optimal decision d in (23) as a function of the current time t and state x. Then (23) can equivalently be written as

$$V_t(x) = E\left[R_t(x,d^*(t,x),\omega) + \beta V_{t+1}(S_t(x,d^*(t,x),\omega)\right] \tag{24}$$

The recursive relationships (23) and (24) provide simple illustrations of Richard Bellman's celebrated *principle of optimality* for stochastic dynamic programming problems. As detailed in Powell (2011, 2014), one practical difficulty is how to compute the value function $V_t(x)$ and the optimal policy function $d^*(t,x)$. Another practical difficulty is that the reward function $R_t(x,d,\omega)$ and/or the state transition function $S_t(x,d,\omega)$ might not be known.

The Q-learning method provides a way to implement decisions in approximate accordance with optimal policy functions for certain classes of decision problems. Below we provide a simple exposition of Q-learning that is applicable for the DM Game.

Suppose a decision problem has an infinite planning horizon, random events ω are governed by a stationary probability distribution, and the reward, state transition, and value functions have time-invariant forms $R(w,d,\omega)$, $S(x,d,\omega)$, and $V(x)$. For each state x and decision d, define

$$Q(x,d) = E\left[R(x,d,\omega) + \beta V(S(x,d,\omega))\right] \tag{25}$$

where the expectation in (25) is taken with respect to the stationary probability distribution governing ω. If the Q-values in (25) can be learned, then the optimal policy function $d^*(x)$ is determined as follows: For any state x,

$$d^*(x) = \arg\max_d Q(x,d) \tag{26}$$

Hence, the learning of the Q-values in (25) avoids the need for separate learning or knowledge of the reward, state transition, and value functions.

In its simplest form, Q-learning uses the following iterative procedure to determine estimates $\widehat{Q(x,d)}$ for the Q-values $Q(x,d)$ in (25), conditional on a user-specified recency parameter α and a user-specified discount factor β:

Step 1: Initialize $\widehat{Q(x,d)}$ to a random value for each possible state x and decision d.

Step 2: Observe an actual state x'.

Step 3: Pick a decision d' and implement it.

Step 4: Observe the next state x'' and the next reward R''.

Step 5: Update the estimate $\widehat{Q(x', d')}$ as follows:

$$\widehat{Q(x', d')} \quad \leftarrow \quad [1 - \alpha]\widehat{Q(x', d')} + \alpha \left[R'' + \beta \max_{d} \widehat{Q(x'', d)}\right] \tag{27}$$

Step 6: Loop back to Step 2 and repeat.

The above procedure does not specify how the decision in Step 3 is to be picked. Let ϵ be any number in $(0, 1)$. The ϵ-greedy variant of Q-learning replaces the above Step 3 with an alternative Step $3'$ incorporating a specific decision selection process that accommodates two goals: (i) Exploit current information for maximum possible current gain; and (ii) seek new information to improve opportunities for future gains. This decision selection process is as follows: With probability ϵ the decision-maker v in Step $3'$ experiments by selecting a random decision d'. However, with probability $[1 - \epsilon]$ the decision-maker v instead "greedily" chooses a decision $\hat{d}$ that maximizes the current estimator $\widehat{Q(x', d)}$ for $Q(x', d)$.

In summary, the ϵ-greedy Q-learning method for a decision-maker v is fully characterized once values are specified for the initial Q-value estimates $\widehat{Q(x, d)}$ and the three parameters $(\alpha, \beta, \epsilon)$.

4.6 EO Decision Rules for Consumers and Firms

Each EO agent (consumer or firm) at the start of each period $t \geq 0$ attempts to maximize an explicit expression for their expected reward (utility or profit) over current and future periods $r \geq t$, subject to constraints. The EO agents use a combined open-loop/closed-loop optimization approach in the following sense: They undertake their maximization problems in each period t conditional on updated state information, yet in these maximizations they ignore the fact that they will re-optimize their decision selections at the start of each future period $r > t$. They also ignore that rationing can occur on the margin in the market clearing processes.

Specifically, at the start of each period $t \geq 0$ an EO consumer i selects a decision d in D^c that maximizes his expected intertemporal utility over current and future periods $r \geq t$. In this maximization, consumer i makes use of the transformation function $\mathbf{TR}^c_{i,t}(d)$ detailed in Section 4.2 to map each possible decision d in D^c at time t into a collection of current and future labor supply and goods demand functions $(l^c_{i,r:1}(w, d, t), q^c_{i,r:3}(p, d, t))_{r \geq t}$.

Formally stated, an EO consumer i's maximization problem at the start of each period $t \geq 0$ takes the following form:

$$\max_{d \in D^c} E_{i,t} U_t(\mathbf{TR}^c_{i,t}(d), \mathbf{w}_{t:1}, \mathbf{p}_{t:3}) \tag{28}$$

subject to the budget and feasibility constraints (1)-(4) for $r \geq t$, where

$$\mathbf{w}_{t:1} = (w_{r:1})_{r=t}^{\infty} \tag{29}$$

$$\mathbf{p}_{t:3} = (p_{r:3})_{r=t}^{\infty} \tag{30}$$

$$\mathbf{div}_{t:5} = (div_{r:5})_{r=t}^{\infty} \tag{31}$$

and

$$U_t(\mathbf{TR}_{i,t}^c(d), \mathbf{w}_{t:1}, \mathbf{p}_{t:3}) = \sum_{r=t}^{\infty} \beta^{r-t} \left[u\left(q_{i,r:3}^c(p_{r:3}, d, t), 1 - l_{i,r:1}^c(w_{r:1}, d, t) \right) \right] \tag{32}$$

Similarly, an EO firm j's maximization problem at the start of each period $t \geq 0$ takes the following form:

$$\max_{d \in D^f} E_{j,t} \Pi_t(\mathbf{TR}_{j,t}^f(d), \mathbf{w}_{t:1}, \mathbf{p}_{t:3}) \tag{33}$$

subject to the technological and feasibility constraints (6)-(11) for $r \geq t$, where $\mathbf{w}_{t:1}$ and $\mathbf{p}_{t:3}$ are defined as in (29) and (30), and

$$\Pi_t(\mathbf{TR}_{j,t}^f(d), \mathbf{w}_{t:1}, \mathbf{p}_{t:3}) = \sum_{r=t}^{\infty} \mu^{r-t} \left[p_{r:3} q_{j,r:3}^f(p_{r:3}, d, t) - w_{r:1} l_{j,r:1}^f(w_{r:1}, d, t) \right] \tag{34}$$

As explained in Appendix B, the expectations in the maximization problems (28) and (33) for each period t are based on estimated probability distributions for future labor market wages, future goods market prices, and future dividend payments (for consumers), conditional on the states of consumer i and firm j at time t.

As explained in Appendix C, approximate solutions for the maximization problems (28) and (33) are derived using two different decision rules. Briefly summarized, the first decision rule, referred to as the *EO Adaptive Dynamic Programming (EO-ADP) decision rule*, derives an approximate solution in each period t by solving a stochastic dynamic programming recurrence relation, assuming a basis-function approximation for the value function. The second decision rule, referred to as the *EO Finite Horizon (EO-FH) decision rule*, replaces the infinite planning horizon in each period t with a finite planning horizon of length T, called the *forecasting horizon*, and then derives an approximate solution by means of direct search across the decision domain.

5 Social Planner Benchmark Model

In order to report comparative performance outcomes for our tested decision-rule combinations, it is desirable to have a benchmark model with a provably unique optimal solution against which the performance of each combination can be compared. This section explains our construction of a social planner model for this purpose.

As detailed in Section 3, in the DM Game all consumers are structurally identical at the initial time 0 and all firms are structurally identical at the initial time 0. Moreover, there are no external shocks. In consequence, heterogeneity among consumers and among firms only arises endogenously, over time, as a result of their market participations.

All sources of uncertainty for the DM Game thus disappear if market decision-making by consumers and firms is replaced by a social planner who maximizes the intertemporal utility of a representative consumer subject only to technological feasibility constraints, conditional on the restriction that the structurally-identical consumers must all be treated alike and the structurally-identical firms must all be treated alike. The resulting model, hereafter referred to as the *Social Planner (SP) Benchmark Model*, is introduced here in order to have a benchmark of comparison for the DM-Game simulation findings reported in Section 7.

Specifically, suppose the number I of DM-Game consumers and the number J of DM-Game firms are arbitrary positive integers, and let $q_{-1}^{stock} \geq 0$ denote the exogenously given goods stock of each firm at the start of period 0. We consider a social planner who solves the following social welfare optimization problem at time 0 on behalf of the representative DM-Game consumer:[11]

$$\max \sum_{t=0}^{\infty} \beta^t u(q_{t:3}^c, 1 - l_{t:1}^c) \tag{35}$$

with respect to $\{l_{t:1}^c, q_{t:3}^c\}_{t=0}^{\infty}$, subject to the following constraints for each $t \geq 0$:

$$J \cdot q_t^{stock} \;=\; J \cdot q_{t-1}^{stock} \;+\; J \cdot F(l_{t:1}^f) \;-\; I \cdot q_{t:3}^c \tag{36}$$

$$l_{t:1}^f \;=\; \frac{I \cdot l_{t:1}^c}{J}$$

$$0 \;\leq\; q_t^{stock},\; q_{t:3}^c$$

$$l_{t:1}^c \in \{0, 1\}$$

To obtain a concrete SP Benchmark Model solution, we assume that the utility function $u(\cdot)$ in (35) takes the form

$$u(q, 1 - l) \;=\; \delta_0^c \cdot \ln\left(b(q) \;+\; q\right) + \delta_1^c \cdot [1 - l] \tag{37}$$

where[12]

[11] Given the exponential form of the discount factor in (35), the social planner would exhibit time consistency, meaning that re-optimization in successive periods would not result in any deviation from the optimal solution determined at time 0.

[12] In order to permit consumers to constructively compare consequences for failure to participate in the goods market, the valuation they place on failure to participate needs to be finite. As will be seen in Section 7, the advantage of introducing the discontinuous valuation function $b(q)$ in (38) is that a consumer's utility takes on a negative value only if he fails to participate in the goods market, thus providing an easily detected signal of this non-participation.

$$b(q) = \begin{cases} 1.0 & \text{if } \; q > 0 \\ b \in (0,1) & \text{if } \; q = 0 \end{cases} \tag{38}$$

Also, the production function $F(\cdot)$ in (36) is assumed to take the form

$$F\,(l) \;=\; \delta_0^f l^{\delta_1^f} \tag{39}$$

We further assume that the values specified for the parameters appearing in this SP Benchmark Model are as listed in Table 2. Finally, for each $t \geq -1$ we let

$$s_t^{stock} \;\equiv\; \frac{J \cdot q_t^{stock}}{I} \tag{40}$$

denote the per-consumer amount of goods stock carried forward from period t to period $t+1$.

Parameter	Value
q_{-1}^{stock}	0.0
β	0.95
δ_0^c	3.0
δ_1^c	0.5
b	0.5
δ_0^f	1.0
δ_1^f	1.0

Table 2: Maintained parameter values for the SP Benchmark Model and the DM Game

Given these concrete specifications, the SP Benchmark Model (35) can be expressed in the following reduced representative-consumer form:

$$\max \; \sum_{t=0}^{\infty} \beta^t \Big[3.0 \cdot \ln(b(q_{t:3}^c) + q_{t:3}^c) + 0.5 \cdot (1 - l_{t:1}^c) \Big] \tag{41}$$

with respect to $\{l_{t:1}^c, q_{t:3}^c\}_{t=0}^{\infty}$, subject to the following constraints for each $t \geq 0$:

$$\begin{gathered} s_t^{stock} \;=\; s_{t-1}^{stock} \;+\; l_{t:1}^c \;-\; q_{t:3}^c \\ 0 \;\leq\; s_t^{stock},\; q_{t:3}^c \\ l_{t:1}^c \in \{0,1\} \\ s_{-1}^{stock} = 0 \end{gathered} \tag{42}$$

The optimal solution of the reduced SP Benchmark Model (41) is a full-employment solution with $l_{t:1}^c = q_{t:3}^c = 1$ and $s_t^{stock} = 0$ for all $t \geq 0$. The proof, by induction, is provided in Appendix D.

Given this optimal solution, the representative consumer attains the stationary per-period utility level

$$u(1,0) = [3.0 \cdot \ln(2)] \approx 2.08 \tag{43}$$

and the intertemporal utility level

$$\sum_{t=0}^{\infty} \beta^t u(1,0) \;=\; \sum_{t=0}^{\infty} [0.95]^t 3.0 \cdot \ln(2) \;=\; 3.0 \cdot \ln(2) \frac{1}{1-0.95} \;\approx\; 41.59 \tag{44}$$

Note that the smallest single-period utility outcome that a representative consumer can feasibly attain under the SP Benchmark Model assumptions is $u(0,0) = 3.0 \cdot \ln(0.5) \approx -2.08$.

6 Sensitivity Design

6.1 Design Overview

The main focus of this study is the degree to which consumers in the DM Game are able to attain optimal utility outcomes when the DM-Game consumers and firms use different combinations of locally-constructive decision rules. The tested decision-rule combinations for consumers (C) and firms (F), identified by assigned case numbers Nk, are displayed in Table 3.

	C:RL	C:FL	C:EO-FH	C:EO-ADP
F:RL	N1−N10	N21	N31	N39
F:FL	N22	N11−N20	N32	N40
F:EO-FH	N33	N34	N23−N30	N41
F:EO-ADP	N42	N43	N44	N35−N38

Table 3: Tested combinations of locally-constructive decision rules

For each of the forty-four cases in Table 3, the utility functions, production functions, initial goods stocks, initial money balances, and initial demand/supply decisions of the consumers and firms were set the same for the Social Planner (SP) Benchmark Model developed in Section 5 and for the DM Game. In particular, the parameter value settings listed in Table 2 for the SP Benchmark Model were also adopted as fixed parameter settings for the DM Game.

Given these common specifications, the unique optimal solution for the SP Benchmark Model is also the unique optimal solution for consumers in the DM Game.[13] Any DM Game-departures from optimality can be attributed solely to the fact that DM-Game outcomes are determined by

[13]Recall from Section 5 that the optimal solution for the SP Benchmark Model is expressed in stationary per-capita form for arbitrary positive numbers of consumers and firms. Consequently, it remains the optimal solution for the DM Game even if some firms become bankrupt and are forced to exit the DM-Game economy.

the locally-constructive decisions of imperfectly informed consumers and firms over time rather than by the intertemporal decisions of a perfectly-informed social planner at the initial time 0.

For each decision-rule case in Table 3, certain key decision-rule parameters were selected as treatment factors while all other parameters were maintained at fixed values. Combinations of treatment-factor values were then selected for testing. For each combination of interest, the number of runs was set at NRuns=10 (corresponding to 10 seed values for pseudo-random number generation) to control for stochastic effects.[14] The length of each run was set to LRun= 1000 periods. To reduce dependence on transient effects, outcomes from the first LOmit=50 periods in each run were omitted from all calculated performance metrics.

6.2 Performance Metrics

DM-Game firms are corporate entities for the facilitation of production. Hence, for the most part, our performance metrics focus on utility outcomes for the DM-Game consumers.

Since different cases involve different planning-horizon lengths, the main ex post performance metric used for each case Nk in Table 3 is *average realized single-period utility* $\bar{u}^k$, bounded above and below by two standard deviations $\sigma_{\bar{u}^k}$. Other ex post performance metrics used to report results include the *average realized single-period utility for period* t, denoted by $\bar{u}^k_t$, the *average realized cumulative utility through period* t, denoted by $\bar{u}^{cumul,k}_t$, the *average realized real market-clearing wage*, denoted by $\bar{w}^{real,k}$, the *average realized real market-clearing wage for period* t, denoted by $\bar{w}^{real,k}_t$, and *average realized single-period profits*, denoted by $\bar{\pi}^k$.

The precise calculation for each of these performance metrics is given in Appendix E.

6.3 Structural Parameter Values Maintained for All Cases

As detailed in Section 3.4, the constraints and goals of the I consumers in the DM Game depend commonly on the specific settings for $(M^c_{-1}, u(\cdot), \beta)$ at the initial time 0. Also, as detailed in Section 3.5, the constraints and goals of the J firms in the DM Game depend commonly on the specific settings for $(M^f_{-1}, q^{stock}_{-1}, F(\cdot), \mu, \kappa^{div})$ at the initial time 0. All of these functions and parameters have fixed specifications for all cases reported in this study. The utility and production function specifications $u(\cdot)$ and $F(\cdot)$, plus the values of β and q^{stock}_{-1}, are set the same as in Section 5 for the SP Benchmark Model, and the values for the remaining parameters are set as in Table 4.

[14] Specifically, these ten seed values were as follows: $\{2012, 2013, 2014, 1, 2, 3, 100, 101, 102, 345\}$.

Parameter	Value
I	10
J	3
M^c_{-1}	1.00
M^f_{-1}	10.00
μ	0.95
κ^{div}	0.50

Table 4: Maintained parameter values for the constraints and goals of consumers and firms

The transformation function $\mathbf{TR}^c_{it}$ for consumer i in period t postulates that consumer i calculates at time t a reservation wage (14) for each current and future period $r \geq t$, which in turn depends on consumer i's expectation for the wage in periods $r \geq t$. Similarly, the transformation function $\mathbf{TR}^f_{j,t}$ for firm j in period t postulates that firm j at time t calculates a reservation wage (17) and a reservation goods price (18) for each current and future period $r \geq t$, which in turn depend on firm j's expectations for the wage and goods price in periods $r \geq t$.

As detailed in Appendix B, the methods used by the consumers and firms to form and update these wage and goods price expectations in each period t depend on these agents' prior beliefs regarding wages and goods prices, and also on their memory length, i.e., the number of past periods they take into account when forming these expectations. The prior-belief parameters are set at maintained values, given in Table 16. However, as will be clarified below in Section 6.5, two different settings are tested for the memory length.

6.4 Parameter Values Maintained for Each Decision Rule

The decision domain D^c in (13) for each consumer i depends on the grid specifications for Ω and Θ. Also, the decision domain D^f in (16) for each firm j depends on the grid specifications for L^f, Γ, Λ, and Ψ. As detailed in Tables 12 through 15 in Appendix A, two different forms are considered for these grid specifications: namely, a *small* form and a *big* form.

The RL decision rule described in Section 4.4, based on the MRE-RL algorithm developed in Nicolaisen et al. (2001), is characterized by the vector of parameter values $(\mathcal{D}, \mathbf{q}(0), C, e, \rho)$. The recency parameter ρ plays a key role in the determination of performance in many previous RL applications, e.g., the work of Roth and Erev cited in Section 4.4. Consequently, we focus attention on ρ as a treatment factor for the RL decision rule.

The maintained values for the remaining RL parameters are set as follows. The parameter $\mathcal{D}$ is the cardinality of the decision domain D^c for an RL consumer or D^f for an RL firm. This cardinality is determined by the grid-type specification for D^c or D^f, which is always set to *small* for an RL consumer or RL firm. The vector $\mathbf{q}(0)$ of initial propensities has dimension $\mathcal{D}$. This vector is set

equal to a fixed vector $\mathbf{q}^{c,*}$ for an RL consumer and to a fixed vector $\mathbf{q}^{f,*}$ for an RL firm, where these fixed vectors are defined as follows. For an RL consumer, the initial propensity assigned by $\mathbf{q}^{c,*}$ to a decision $d^c = (l^c, \omega, \theta) \in D^c$ is 1.1 if $l^c = 1$ and 1.0 otherwise. For an RL firm, the initial propensity assigned by $\mathbf{q}^{f,*}$ to a decision $d^f = (l^f, \gamma, \lambda, \psi) \in D^f$ is 1.1 if $l^f = l_L^f$ and 1.0 otherwise. The cooling parameter C is set to 1.0. Finally, based on the results reported in Nicolaisen et al. (2001) and subsequent MRE-RL studies, the experimentation parameter e is set to 0.95. These maintained values are summarized in Table 5.

Parameter	Value
grid-type	small
$\mathbf{q}(0)$	$\mathbf{q}^{c,*}, \mathbf{q}^{f,*}$
C	1.00
e	0.95

Table 5: Maintained parameter values for RL decision rules

The FL decision rule described in Section 4.5, a "greedy" variant of Q-learning, is characterized by the vector $\mathbf{Q}_0$ of initial Q-value estimates $\widehat{Q(x,d)}$ and the parameter vector $(\alpha, \beta, \epsilon)$. To facilitate comparisons with the RL decision rule, we select the recency parameter α to be a treatment factor for the FL decision rule.

The state-space for each FL agent is discretized in order to keep computational solution-times manageable. The state $x_{i,t}$ of an FL consumer i at each time $t \geq 0$ is given by his time-t money balance $M_{i,t-1}^c$, discretized into the following three bins: $[0.0, 5.0)$,$[5.0, 10.0)$, $[10.0, \infty)$. The state $x_{j,t}$ of an FL-firm j at each time $t \geq 0$ consists of its time-t money balance M_{t-1}^f and its time-t goods stock q_t^{stock}, each also discretized into three bins, as follows: for the money balance, $[0.0, 50.0)$, $[50.0, 100.0)$, $[100.0, \infty)$; and for the goods stock, $[0.0, 5.0)$,$[5.0, 10.0)$, $[10.0, \infty)$.

The maintained parameter values for the FL decision rule are then set as follows. The vector $\mathbf{Q}_0$ is set equal to a fixed vector $\mathbf{Q}^{c,*}$ for an FL consumer and to a fixed vector $\mathbf{Q}^{f,*}$ for an FL firm. For an FL consumer, the initial Q-value estimate assigned by $\mathbf{Q}^{c,*}$ to a state-decision pair (x, d^c), where $d^c = (l^c, \omega, \theta) \in D^c$, is 0.5 if $l^c = 1$ and 0.0 otherwise. For an FL firm, the initial Q-value estimate assigned by $\mathbf{Q}^{f,*}$ to a state-decision pair (x, d^f), where $d^f = (l^f, \gamma, \lambda, \psi) \in D^f$, is 0.5 if $l^f = l_L^f$ and 0.0 otherwise. Finally, the Q-learning discount parameter β is set to 0.95 and the greedy parameter ϵ is set to 0.10. These maintained values are summarized in Table 6.

Parameter	Value
grid-type	small
$\mathbf{Q}_0$	$\mathbf{Q}^{c,*}, \mathbf{Q}^{f,*}$
β	0.95
ϵ	0.10

Table 6: Maintained parameter values for FL decision rules

Maintained parameter values for the EO-ADP and EO-FH decision rules are provided in Appendix C, together with detailed discussions of their formulations and implementations.

6.5 Tested Specifications for Case Treatment Factors

As detailed in Appendix A, two different settings are tested for the decision-domain grid specifications: namely, a *small* setting and a *big* setting. Although a small grid-type is maintained for both the RL and FL decision rules, both small and big grid-types are tested for EO agents.

As detailed in Appendix B, two different settings are tested for the memory parameter wm used by consumers and firms to adaptively update their expectations. The first setting, wm=*1*, indicates that consumers and firms in each period $t > 0$ only make use of realizations from the previous period $t-1$ to form their expectations for periods $r \geq t$. The second setting, wm=*all*, indicates that consumers and firms in each period $t > 0$ make use of realizations from all previous periods $\{0, \ldots, t-1\}$ to form their expectations for periods $r \geq t$.

Note that all tested cases depend on the setting for wm. This dependence arises because, as detailed in Sections 4.2 and 4.3, the transformation functions $\mathbf{TR}^c_{i,t}$ and $\mathbf{TR}^c_{j,t}$ mapping consumer and firm period-t decisions into collections of demand and supply functions for periods $r \geq t$ depend on the wage, price, and dividend payment expectations of the consumers and firms, which in turn depend on wm.

For the cases listed along the diagonal in Table 3, the tested combinations of values for the treatment-factor parameters are as shown in Tables 7 through 10. All cross-products of the listed parameter values are tested.

Parameter	Range of Values
ρ	$\{0.05, 0.10, 0.5, 0.90, 0.95\}$
wm	1, all

Table 7: Treatment-factor values for the RL decision rules in Cases N1-N10

Parameter	Range of values
α	$\{0.05, 0.10, 0.50, 0.90, 0.95\}$
wm	1, all

Table 8: Treatment-factor values for the FL decision rules in Cases N11-N20

Parameter	Range of values
T	$\{5, 20\}$
wm	1, all
grid-type	small, big

Table 9: Treatment-factor values for the EO-FH decision rules in Cases N23-N30

Parameter	Range of values
wm	1, all
grid-type	small, big

Table 10: Treatment-factor values for the EO-ADP decision rules in Cases N35-N38

For the remaining cases in Table 3, the treatment-factor values are as shown in Table 11. Superscripts are used to indicate for which decision rule each treatment-factor value applies.

Parameter	Value
ρ^{RL}	0.05
wm^{RL}	*all*
α^{FL}	0.05
wm^{FL}	*all*
T^{EO-FH}	20
wm^{EO-FH}	*all*
grid-type^{EO-FH}	*small*
wm^{EO-ADP}	*all*
grid-type^{EO-ADP}	*small*

Table 11: Treatment-factor values for Cases N21, N22, N31-N34, and N39-N44

7 Key Findings for the DM Game

7.1 Findings Overview

This section reports key findings for the forty-four DM-Game cases N1-N44 listed in Table 3. Each case corresponds to a distinct setting of the treatment-factor values for this case's indicated decision-rule combination.

An important point to keep in mind in interpreting these findings is that bankrupt firms must exit the DM-Game economy, and there is no mechanism for firm entry. Consequently, the number of firms in each simulation run either stays the same or declines. It is therefore a very challenging problem for the DM-Game economy to sustain good performance over long periods of time.

In particular, since consumer and firm agents in the DM Game have no *a priori* information regarding the form of the optimal SP Benchmark Model solution, they do not know that their initially-set conditions are in fact optimal conditions. Consequently, departures from optimality immediately begin to arise as the consumer and firm agents start exploring their decision domains in search of better utility and profit outcomes. These exploratory efforts result in a highly non-stationary environment that makes learning difficult.

As will be seen below, given some decision-rule combinations, the DM Game consumer and firm agents are able to learn their way back towards the optimal solution for the SP Benchmark Model with full employment. This tends to occur more frequently when the agents have long memories (wm=all), because long memories permit the agents to recall their initial utility and profit outcomes when in fact their selected decisions were close to optimal. In such cases, good average performance results.

Given other decision-rule combinations, however, the agents' early exploratory activities result in "mistakes" that propagate throughout the DM-Game economy, causing a downward spiraling of performance from which the economy does not recover. For example, some consumers and firms might make decisions that, given current market conditions, result in disastrous consequences for them. Firms might become bankrupt, consumers might lose money, firms might be unable to secure workers, and consumers might be unable to find work. These bad outcomes then result in further bad outcomes. The majority of cases for which performance is poor exhibit growing unemployment and increasing divergence from the optimal SP Benchmark Model solution.

Overall, cases in which each consumer and firm agent uses a rolling fixed-horizon EO-FH decision rule tend to achieve better performance than cases in which these agents use RL, FL, and/or EO-ADP decision rules. Unlike the RL decision rule, the EO-FH and EO-ADP decision rules entail adaptive *foresight*. Unlike the RL and FL decision rules, the EO-FH and EO-ADP decision rules exploit the *structural form* of the intertemporal optimization problems for consumers and firms. However, unlike the EO-ADP decision rule, the EO-FH decision rule relies on only one structural approximation: a truncation of the planning horizon. In contrast, the EO-ADP decision rule relies on the basis-function approximation of dynamic programming value functions.

Nevertheless, comparative performance is also seen to depend strongly on the specific settings for the treatment-factor parameters. For example, all else equal, a long memory covering all previous periods (wm=all) tends to result in better performance than a short memory covering only the latest period (wm=1).

Sections 7.2 through 7.5 report findings obtained for the diagonal cases in Table 3, for which the DM consumers and firms all use the same type of decision rule. Section 7.6 reports findings for the off-diagonal cases in which mixed combinations of decision rules are used.

7.2 Findings for the Pure RL Cases N1-N10

Consider cases N1-N10 in Table 3, for which all DM-Game consumers and firms use an RL decision rule entailing reactive reinforcement learning. Each of these cases corresponds to a distinct setting of the RL treatment factors (ρ, wm) in Table 7, taking as given the maintained parameter values in Table 5.

As seen in Section 4.4, the RL recency parameter $\rho \in [0, 1]$ determines the weight $[1 - \rho]$ that is placed on accumulated past single-period utility outcomes relative to the weight $[1 - e]$ placed on the most recent single-period utility outcome. Since e is set at the maintained value $e = 0.95$, a reduction in ρ implies an increase in the weight placed on past utility outcomes relative to the weight placed on the most recent utility outcome.

Figure 4 reports performance outcomes for cases N1-N10 in Table 3. The performance of each case Nk is measured by average realized single-period utility $\bar{u}^k$, and cases are reported from left to right in ascending performance order.

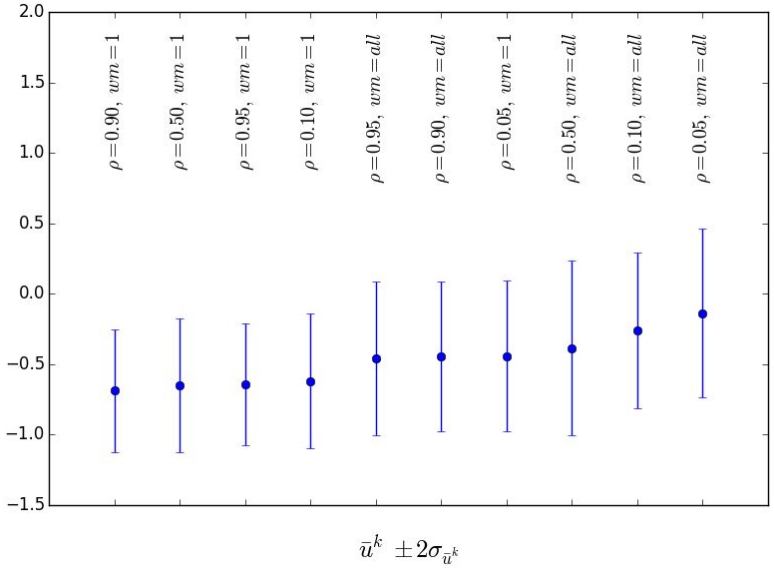

Figure 4: Pure RL Cases N1-N10: Average realized single-period utility $\bar{u}^k$ with bounds of $\pm$ two standard deviations $\sigma_{\bar{u}^k}$

Given the longer memory length wm=all, it is seen that smaller RL recency parameter values ρ (i.e., larger weights on past utility outcomes) tend to result in better performance than larger ρ values. Given a one-period memory length wm=1, however, a relatively low performance level results for all ρ values. Moreover, even in the best-performing cases, performance is significantly below 2.08, the optimal stationary per-period utility level (43) obtained by the representative consumer in the SP Benchmark Model

7.3 Findings for the Pure FL Cases N11-N20

Consider cases N11-N20 in Table 3, for which all DM-Game consumers and firms use an FL decision rule based on Q-learning. Each of these cases corresponds to a distinct setting of the FL treatment factors (α, wm) in Table 8, taking as given the maintained parameter values in Table 6.

As seen in Section 4.5, the FL recency parameter $\alpha \in [0, 1]$ determines the weight $[1 - \alpha]$ that is placed on past Q-value estimates relative to the weight α placed on current and anticipated future utility outcomes. Since these two weights sum to 1.0, a reduction in α implies an increase in the weight placed on past utility outcomes relative to current and anticipated future utility outcomes.

Figure 5 reports performance outcomes for cases N11-N20 in Table 3. The performance of each case Nk is measured by average realized single-period utility $\bar{u}^k$, and cases are reported from left to right in ascending performance order.

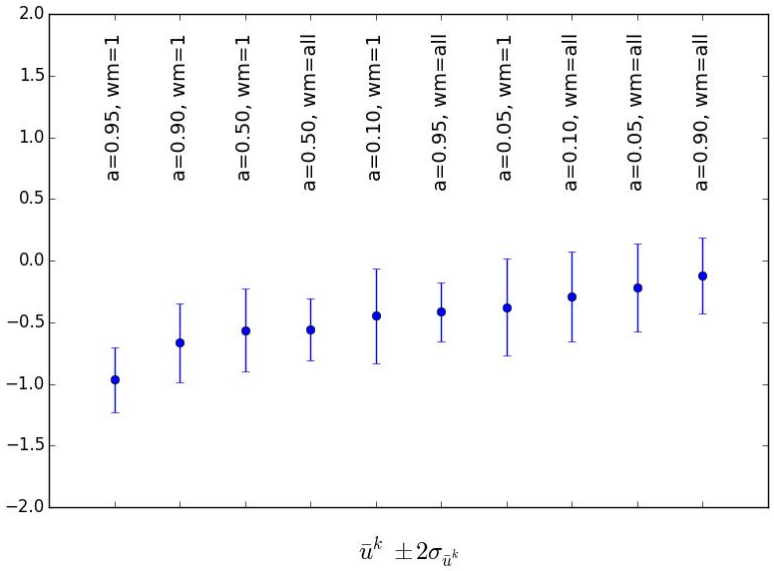

Figure 5: Pure FL Cases N11-N20: Average realized single-period utility $\bar{u}^k$ with bounds of $\pm$ two standard deviations $\sigma_{\bar{u}^k}$

The best pure FL performance is achieved for the case in which the memory length wm is long (wm=all) and the FL-recency parameter α is set at 0.90. Surprisingly, however, this best performance is also nearly achieved for wm=all with $\alpha = 0.05$ or $\alpha = 0.10$. However, for each tested α setting, performance improves as wm is increased from wm=1 to wm=all.

The implication we draw from these findings is that our performance metrics are not very sensitive to the setting of the FL recency parameter α in the pure FL experiments conducted to date.

7.4 Findings for the Pure EO-FH Cases N23-N30

Consider cases N23-N30 in Table 3, for which all DM-Game consumers and firms use an EO-FH decision rule entailing explicit intertemporal optimization by means of a rolling fixed horizon. Each of these cases corresponds to a distinct setting of the EO-FH treatment factors T, wm, and grid-type in Table 9, taking as given the maintained parameter value NDrawsFH=10 discussed in Appendix C.2.

The length T of the forecasting horizon controls the extent to which an EO-FH agent is forward looking. This anticipation could be beneficial if the agent's anticipations are an accurate reflection of future uncertainties, but harmful if not. Restricting the number of potential decision selections by specifying grid-type=*small* rather than grid-type=*big* increases the sampling density, i.e., the frequency with which each potential decision is tried. On the other hand, grid-type=*small* results in a cruder approximation of the decision domain, which could prevent an EO-FH agent from determining a truly best decision.

Figure 6 reports performance outcomes for cases N23-N30 in Table 3. The performance of each case Nk is measured by average realized single-period utility $\bar{u}^k$, and cases are reported from left to right in ascending performance order.

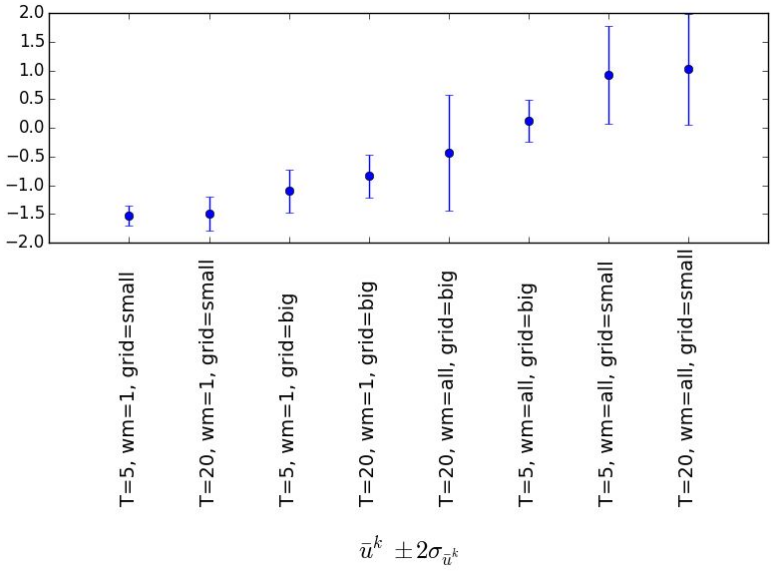

Figure 6: Pure EO-FH Cases N23-N30: Average realized single-period utility $\bar{u}^k$ with bounds of $\pm$ two standard deviations $\sigma_{\bar{u}^k}$

Given a one-period memory length wm=1, performance is relatively low regardless of the grid-type or the length T of the forecasting horizon. However, given a longer memory length wm=*all*, it is seen that having a small grid-type results in better performance than a large grid-type.

Moreover, for wm=*all* and grid-type=*small*, the longer forecasting horizon T=20 yields slightly better performance than the short forecasting horizon T=5. Indeed, as indicated by the standard deviation bounds in Fig. 6, for this combination of treatment factors the average realized single-period utility level $\bar{u}_t^k$ attained in some periods t comes close to matching the optimal stationary single-period utility level 2.08 achieved by the representative consumer in the SP Benchmark Model. This occurs despite the rather simplistic Monte Carlo method used by EO-FH agents to handle their uncertainty regarding future wages, prices and dividends.

Given the relatively good performance of the EO-FH decision rule under some treatments, it is interesting to delve deeper into the underlying dynamics. Time-series for utility and real wage outcomes are depicted below for two illustrative cases: (i) a "good" case N26 with T=20, wm=*all*, and grid-style=*small*; and (ii) a "bad" case N29 with $T = 20$, wm=1, and grid-style=*big*.

For the "good" case N26, depicted in Fig. 7, the average realized single-period utility $\bar{u}_t^{26}$ eventually stabilizes at a level of about 0.5. For the "bad" case N29, depicted in Fig. 8, the average realized single-period utility $\bar{u}_t^{29}$ quickly stabilizes at a much lower level of about -1.0.

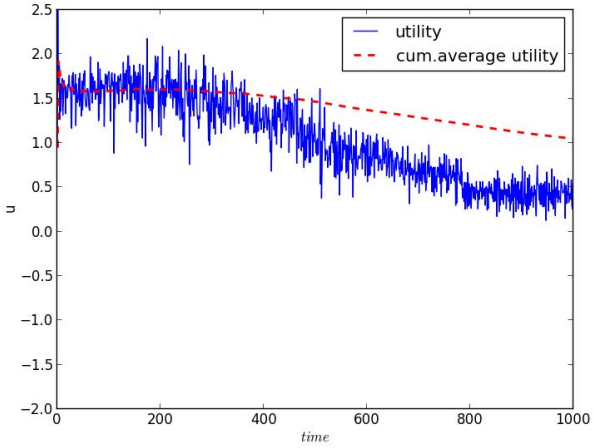

Figure 7: Pure EO-FH Case N26: Average realized single-period utility $\bar{u}_t^{26}$ for period t and average realized cumulative utility $\bar{u}_t^{cumul,26}$ through period t, over successive periods t

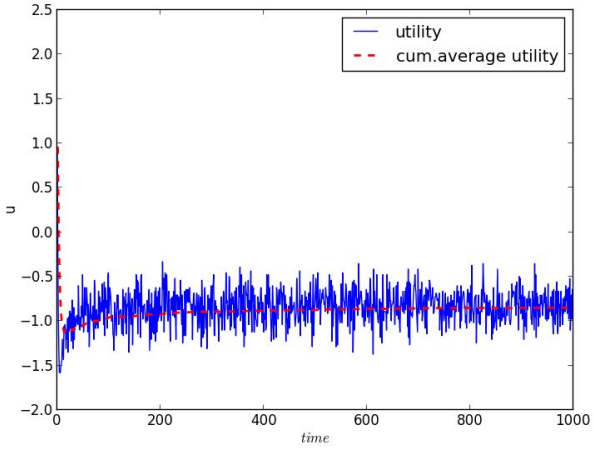

Figure 8: Pure EO-FH Case N29: Average realized single-period utility $\bar{u}_t^{29}$ for period t and average realized cumulative utility $\bar{u}_t^{cumul,29}$ through period t, over successive periods t

The behavior of average real market-clearing wages reflects overall macroeconomic performance. For the "good" case N26, it is seen in Fig. 9 that $\bar{w}_t^{real,26}$ appears to be stabilizing at a level of

about 0.30. In contrast, for the "bad" case N29, it is seen in Fig. 10 that $\bar{w}_t^{real,29}$ rapidly drops towards zero.

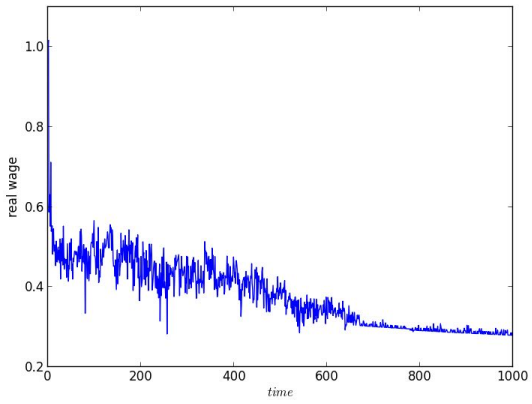

Figure 9: Pure EO-FH Case N26: Average realized real market-clearing wage $\bar{w}_t^{real,26}$ for period t, over successive periods t

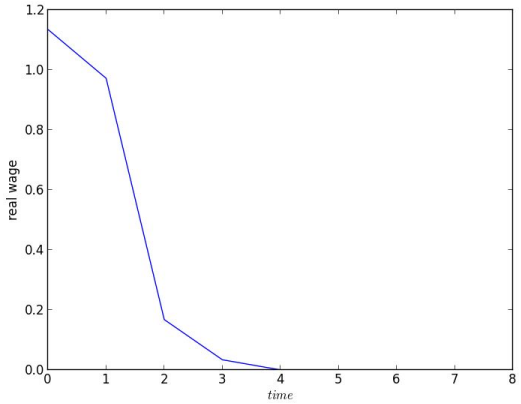

Figure 10: Pure EO-FH Case N29: Average realized real market-clearing wage $\bar{w}_t^{real,29}$ for period t, over successive periods t

7.5 Findings for the Pure EO-ADP Cases N35-N38

Consider cases N35-N38 in Table 3, for which all DM-Game consumers and firms use an EO-ADP decision rule entailing explicit optimization via the approximation of dynamic programming value functions. Each of these cases corresponds to a distinct setting of the EO-ADP treatment factors wm and grid-type in Table 10, taking as given the maintained parameter values listed in Table 17.

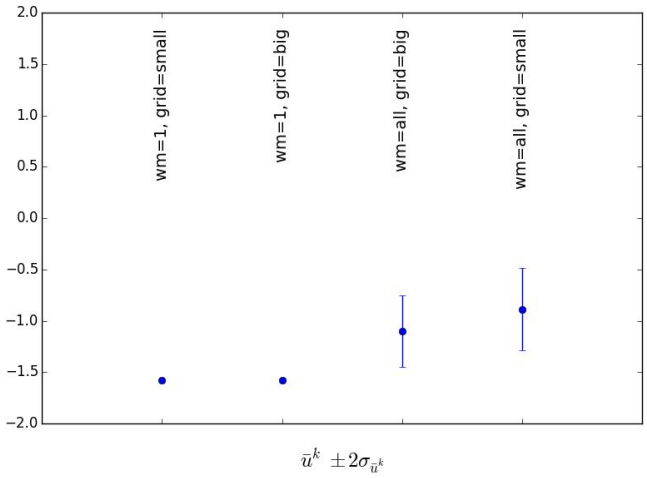

Figure 11: Pure EO-ADP Cases N35-N38: Average realized single-period utility $\bar{u}^k$ with bounds of $\pm$ two standard deviations $\sigma_{\bar{u}^k}$

Figure 11 reports performance outcomes for these pure EO-ADP cases. The performance of each case Nk is measured by average realized single-period utility $\bar{u}^k$, and cases are reported from left to right in ascending performance order.

EO-ADP performance is clearly better with a longer memory length (wm=*all*) than with a one-period memory length (wm=1). Moreover, given a longer memory, performance is slightly better with grid-style=*big* in comparison with grid-style=*small*. Overall, however, a low performance level is attained for all tested settings of the EO-ADP treatment factors in comparison with the overall performance attained using the RL, FL, and EO-FH decision rules.

7.6 Findings for Mixed Combinations of Decision Rules

From a social welfare point of view, it is only consumer utility outcomes that matter in the DM Game. However, the players in the DM Game are utility-seeking consumers and profit-seeking firms, where the latter act on behalf of their shareholders (who receive their profits as dividend payments) but not consciously on behalf of consumer welfare per se. Consequently, it is of interest to construct consumer and firm payoff matrices for the DM Game, interpreting the decision rules RL, FL, EO-FH, and EO-ADP as possible pure strategy choices for these players.

We therefore tested the off-diagonal cases in Table 3 representing mixed combinations of decision rules. We then used the performance outcomes obtained for these off-diagonal cases together with the performance outcomes obtained for the diagonal cases to construct DM-Game payoff matrices, one for consumers and one for firms, under the restriction that all consumers use the same decision rule and all firms use the same decision rule.

The consumer payoff matrix, depicted in Fig. 12, reports the average realized single-period utility $\bar{u}^k$ attained by consumers for each indicated case Nk, with darker shades of color corresponding to higher values of $\bar{u}^k$. The firm payoff matrix, depicted in Fig. 13, reports the average realized single-period profits $\bar{\pi}^k$ attained by firms for each indicated case Nk, with darker shades of color corresponding to higher values of $\bar{\pi}^k$.

It is important to note the following non-standard aspect of these payoff matrices. For each pairing of consumer and firm decision rules along the diagonals, the treatment-factor parameters are selected in an attempt to permit each agent type to do as well as possible in this pairing. This is reflected in the fact that, in contrast to Table 3, only single cases are considered along the diagonals.

As seen from the firm payoff matrix in Fig. 13, EO-FH is a dominant strategy for firms, given the particular case selections and treatment-factor specifications used to form this payoff matrix. Interestingly, as seen from the consumer payoff matrix in Fig. 12, this is not true for consumers. For example, the best response of consumers to a firm choice of FL is to choose FL, not EO-FH. Nevertheless, it is also seen from these two payoff matrices that (F:EO-FH, C:EO-FH) is a Pareto optimal Nash equilibrium

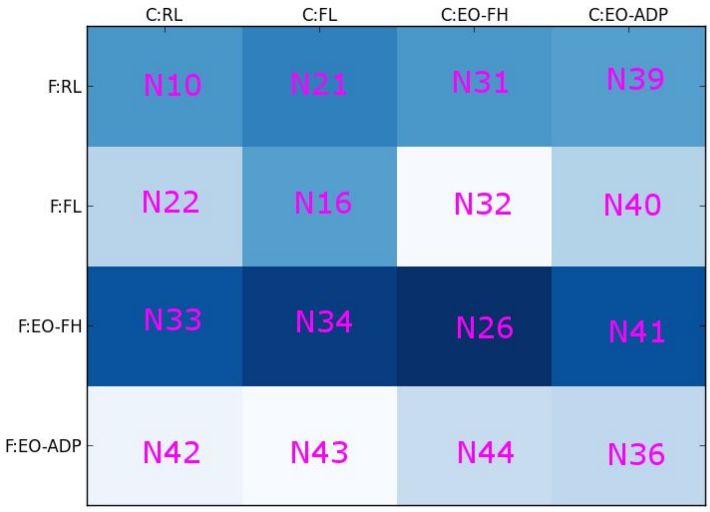

Figure 12: Consumer payoff matrix for the DM Game reporting average realized single-period utility $\bar{u}^k$ for the indicated cases Nk. A darker shade of color indicates a higher value for $\bar{u}^k$.

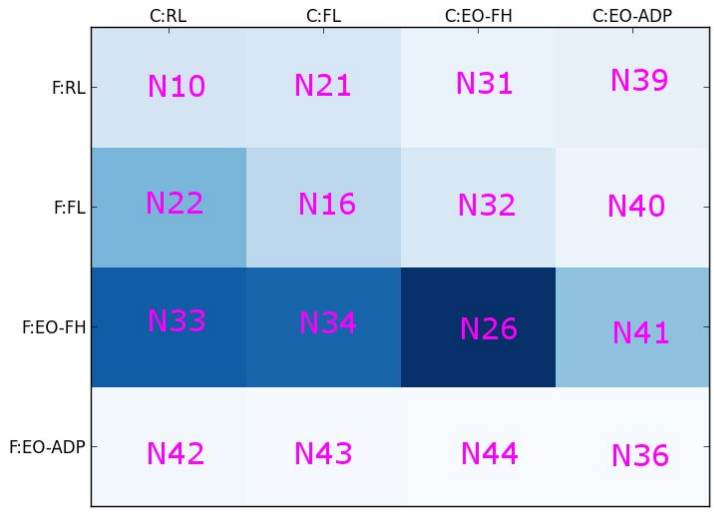

Figure 13: Firm payoff matrix for the DM Game reporting average realized single-period profits $\bar{\pi}^k$ for the indicated cases Nk. A darker shade of color indicates a higher value for $\bar{\pi}^k$.

8 Conclusion

In studies involving a single learning agent operating in a stochastic environment, the form of learning that is best for this agent will depend strongly on whether the stochastic properties of the environment are stationary or non-stationary. Long memories are typically found to be desirable in stationary environments but not necessarily in non-stationary environments.

However, in multi-agent learning situations such as the DM Game, additional considerations complicate the determination of optimal learning rules. Even if the physical and institutional environment is stationary in terms of its stochastic properties, the presence of other learning agents can induce non-stationarity in the learning environment.

More precisely, if multiple learning agents in a stationary physical and institutional environment are relatively more responsive to the recent actions of other learning agents, in the sense that they put relatively higher weight on more recent observations, this can induce persistent suboptimal fluctuations and cycling as agents continually try to adapt to each other's recent actions. Conversely, longer memories with a more even weighting of recent and past observations will tend to induce inertia in the system, which can result in the system settling down to a particular outcome; but this outcome is not guaranteed to be optimal.

Although our DM-Game study has a stationary institutional environment, the learning environment is non-stationary. First, the physical and financial environment is time-varying due to production and the accumulation of goods stocks and money balances. Second, all of our tested

decision rules are adaptive rules in the sense that they are conditioned on the current state of the world. More precisely, the multiple learning agents use these adaptive rules to make successively determined decisions conditional on time-varying information, beliefs, and physical states. In such a world, it is not clear *a priori* whether it would be better for agents to have long memories that take into account a long history of past observations or shorter memories that only take into account more recent observations.

A key finding of our DM-Game study is, indeed, the importance of memory length in determining DM-Game performance. Simpler decision rules such as RL, FL, and EO-FH can outperform more sophisticated decision rules such as EO-ADP, but only if coupled with a relatively long memory length. The benefit of a long memory arises because we initialize the DM Game to optimal SP Benchmark Model settings. It is then beneficial for consumers and firms to be able to recall the utility and profit outcomes they attained during initial periods because their decisions were in fact close to optimal during these initial periods.

However, additional studies are surely needed to better understand the implications of locally-constructive decision making in macroeconomic contexts such as the DM Game. To date we have only explored a small number of initial conditions and parameter settings. In particular, there is no guarantee that the parameter values we have set for our decision rules are well-tailored for the DM-Game environment in which these decision rules are being used. For example, the cooling parameter C appearing in equation (20) for the RL decision rule plays an important role in many well-known learning methods and solution algorithms, such as simulated annealing; yet the current study treats C as a maintained parameter value fixed at C=1.

Moreover, our current DM-Game study imposes various structural restrictions that would be interesting to relax. For example, the labor offers of the DM consumers are restricted to be in 0-1 binary form in order to simplify the analysis and reporting of comparative learning outcomes. It would be of interest to explore more general labor market formulations in future studies of macroeconomic systems with constructively rational agents.

It would also be of great interest to consider the effects on macroeconomic performance when consumers have positive subsistence consumption needs and have to trade for consumption goods in order to survive. This would put stronger pressure on consumers to participate in both labor and goods markets, which in turn would affect their reservation wages and prices. In particular, it would be intriguing to investigate what the best-performing decision rules for consumers and firms would be in response to a systematic increase in the subsistence consumption need level. Would conservative decision-making based on reactive reinforcement learning eventually dominate forward-looking decision-making, such as the EO-FH decision rule that performs best in our current DM Game with zero subsistence consumption needs?

Moreover, in our current DM-Game modeling, remaining firms do not immediately modify their behavior to take into account their larger market shares when some firms are forced to exit due to

bankruptcy. Permitting firms to understand the strategic implications of having other firms exiting due to bankruptcy would be an interesting extension to consider in future work. For example, it could lead to temporary deep price-cutting by firms that have relatively deep pockets (i.e., that are able to withstand temporary dips in profits) in an attempt to drive other firms out of business. However, to achieve a more compelling modeling of this type of strategic firm behavior, it would presumably also be necessary to consider the contestability of markets and to permit possible firm (re)-entry if price-cutting firms subsequently attempt to exploit market power opportunities by raising prices higher than competitive levels.

Clearly, then, much further study is needed to understand the ramifications of modeling macroeconomies as constructively-rational games, in the sense that agent decision-making is based solely on own interaction networks, beliefs, information, and physical states without external support from modeler-imposed coordination and optimality conditions. In particular, a large unexplored gap exists between constructive rationality and *constructive optimality*, i.e., the assurance that the combination of locally-constructive decision processes in use by agents satisfies some stated optimality property, such as Pareto optimality.

Nevertheless, the primary goal of this study has been accomplished. It has been demonstrated that decision-makers in computational macroeconomic models can implement locally-constructive decision processes ranging all the way from reactive reinforcement learning to adaptive intertemporal optimization within the context of a purely historical sequence of events, without the imposition of external coordination and optimality restrictions.

Another important goal has been the development of the DM Game as a computational laboratory. Coded in C++, the DM Game is a modular, extensible, and scalable macroeconomic framework that permits the comparative analysis of different physical and institutional environments populated by a mix of decision-making agents with diverse decision processes. In future work the current structure of the DM Game will be extended to include additional critical features, such as a central government, a central bank, and a commercial banking system.

Acknowledgement

The authors are grateful to J. Bhattacharya, S. Kautz, P. Orazem, S. M. Ryan, I. Salle, the JEDC Editor, and three anonymous JEDC referees for helpful comments on this work.

References

Alden, J. M. and R. L. Smith (1992). Rolling horizon procedures in nonhomogeneous markov decision processes. *Operations Research 40*(3), S183–S194.

Arthur, W. B. (2015). *Complexity and the Economy*. Oxford University Press, Oxford, UK.

Chen, S.-H. (2012). Varieties of agents in agent-based computational economics: A historical and an interdisciplinary perspective. *Journal of Economic Dynamics and Control 36*(1), 1–25.

Dawid, H., S. Gemkow, P. Harting, S. van der Hoog, and M. Neugart (2015). Agent-based macroeconomic modeling and policy analysis: The Eurace@Unibi Model. In S.-H. Chen (Ed.), *Handbook on Computational Economics and Finance*. Oxford University Press.

Dosi, G. and M. Egidi (1991). Substantive and procedural rationality. *Journal of Evolutionary Economics 1*, 145–168.

Dosi, G., G. Fagiolo, and A. Roventini (2010). Schumpeter meeting keynes: A policy-friendly model of endogenous growth and business cycles. *Journal of Economic Dynamics and Control 34*(9), 1748–1767.

Erev, I. and A. E. Roth (1998). Predicting how people play games with unique, mixed-strategy equilibria. *American Economic Review 88*, 848–881.

Evans, G. W. and S. Honkapohja (2013). Learning as a rational foundation for macroeconomics and finance. In R. Frydman and E. Phelps (Eds.), *Rethinking Expectations: The Way Forward for Macroeconomics*. Princeton University Press.

Gode, D. K. and S. Sunder (1993). Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy 101*, 119–137.

Hommes, C. (2013). *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems*. Cambridge University Press, UK.

Honkapohja, S., K. Mitra, and G. Evans (2012). Notes on agents' behavioral rules under adaptive learning and studies in monetary policy. Technical report, Centre for Dynamic Macroeconomic Analysis, University of St. Andrews, UK. Working Paper CDMA 11/02.

Howitt, P. (2008). Macroeconomics with intelligent autonomous agents. In R. Farmer (Ed.), *Macroeconomics in the Small and the Large: Essays on Microfoundations, Macroeconomic Applications and Economic History in Honor of Axel Leijonhufvud*. Edward Elgar.

Howitt, P. (2012, March). What have central bankers learned from modern macroeconononomic theory? *Journal of Macroeconomics 34*(1), 11–22.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kirman, A. (2011). *Complex Economics: Individual and Collective Rationality*. Routledge, London.

LeBaron, B. and L. Tesfatsion (2008). Modeling macroeconomies a open-ended dynamic systems of interacting agents. *American Economic Review (Papers and Proceedings) 98*(2), 246–250.

Mandel, A., C. Jaeger, S. Fürst, W. Lass, D. Lincke, F. Meissner, F. Pablo-Marti, S. Wolf, et al. (2010). Agent-based dynamics in disaggregated growth models. *Université Paris1 Panthéon-Sorbonne (Post-Print and Working Papers)*.

Milani, F. (2005). Expectations, learning, and macroeconomic persistence. Technical report, EABCN Working Paper. http://www.socsci.uci.edu/ fmilani/Milani_ELMP.pdf.

Milani, F. (2007). Expectations, learning, and macroeconomic persistence. *Journal of Monetary Economics 54*(7), 2065–2082.

Mitra, K., G. W. Evans, and S. Honkapohja (2013). Policy change and learning in the rbc model. *Journal of Economic Dynamics and Control 37*(10), 1947–1971.

Nicolaisen, J., V. Petrov, and L. Tesfatsion (2001). Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. *IEEE Transactions on Evolutionary Computation 5*(5), 504–523.

Oeffner, M. (2008). Agent-based keynesian macroeconomics - an evolutionary model embedded in an agent-based computer simulation. *Dissertation, Julius-Maximilians-Universitaet Wuerzburg, Germany*.

Powell, W. (2014). Clearing the Jungle of Stochastic Optimization: INFORMS Tutorial. http://www2.econ.iastate.edu/tesfatsi/ClearingJungle.WPowell.June2014.pdf.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Volume 842. Wiley.

Roth, A. E. and I. Erev (1995). Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior 8*, 164–212.

Salle, I. and P. Seppecher (2013). Technical report, Groupe de REcherche en Droit, Economie, Gestion (GREDEG CNRS) Working Paper 2013-18, University of Nice, Sophia Antipolis.

Salle, I., M. Yildizoglu, and M. Senegas (2013). Inflation targeting in a learning economy. *Economic Modelling 34*, 114–128.

Sbordone, A. M., A. Tambalotti, K. Rao, and K. Walsh (2010). Policy analysis using DSGE models: An introduction. *FRBNY Economic Policy Review*.

Simon, H. A. (1978). Rationality as process and as product. *American Economic Reviewr 68*(2), 1–12.

Smets, F. and R. Wouters (2003). An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association 1*(5), 1123–1175.

Smith, V. L. (2008). *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge University Press.

Stiglitz, J. E. (2002). Information and the change in the paradigm in economics. *The American Economic Review 92*(3), 460–501.

Tesfatsion, L. (2015a). ACE Research Area: Agent-Based Macroeconomics. http://www2.econ.iastate.edu/tesfatsi/amulmark.htm.

Tesfatsion, L. (2015b). ACE Research Area: Learning and the Embodied Mind. http://www2.econ.iastate.edu/tesfatsi/aemind.htm.

Tesfatsion, L. (2015c). Agent-Based Computational Economics: Homepage. http://www2.econ.iastate.edu/tesfatsi/ace.htm.

Tesfatsion, L. and K. L. Judd (Eds.) (2006). *Handbook of Computational Economics: Volume 2, Agent-Based Computational Economics*. Handbooks in Economics Series, North-Holland, Elsevier.

Tovar, C. E. (2009). DSGE models and central banks. *The Open-Access, Open Assessment E-Journal 3*, 1–31.

Trichet, J.-C. (2010). Reflections on the nature of monetary policy non-standard measures and finance theory. *available online: http://www.ecb.int/press/key/date/2010/html/sp101118.en.html*.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. *Cambridge University, Cambridge, England, Doctoral thesis*.

Appendix

A Tested Grid Specifications for Decision Domains

Decision Component	Set of Possible Values
l^c	$L^c = \{0, 1\}$
ω	$\Omega = \{0.8, 1.0, 1.2\}$
θ	$\Theta = \{0.0, 0.5, 1.0\}$

Table 12: Small-grid discretization of the consumer decision domain D^c

Decision Component	Set of Possible Values
l^f	$L^f = \{0, 2.5, 5.0, 7.5, 10\}$
γ	$\Gamma = \{0.8, 1.0, 1.2\}$
λ	$\Lambda = \{0.8, 1.0, 1.2\}$
ψ	$\Psi = \{0.0, 0.5, 1.0\}$

Table 13: Small-grid discretization of the firm decision domain D^f.

Decision Component	Set of Possible Values
l^c	$L^c = \{0, 1\}$
ω	$\Omega = \{0.10, 0.55, 1.00, 1.45, 1.90\}$
θ	$\Theta = \{0.0, 0.5, 1.0\}$

Table 14: Big-grid discretization of the consumer decision domain D^c

Decision Component	Set of Possible Values
l^f	$L^f = \{0, 2.5, 5.0, 7.5, 10\}$
γ	$\Gamma = \{0.10, 0.55, 1.00, 1.45, 1.90\}$
λ	$\Lambda = \{0.10, 0.55, 1.00, 1.45, 1.90\}$
ψ	$\Psi = \{0.0, 0.5, 1.0\}$

Table 15: Big-grid discretization of the firm decision domain D^f

B Wage, Price, and Dividend Expectation Updating

Consumers and firms in the DM Game are assumed to follow the same methods in forming and updating their expectations regarding the distributions governing future market-clearing wages,

market-clearing goods prices, and dividend payments (for consumers). These methods are characterized by prior-belief parameters and a memory length parameter. The prior-belief parameters are maintained parameters set at fixed values for all test cases reported in this study. The memory length parameter is a treatment factor set to reflect either a fixed one-period memory or an expanding memory that takes into account all previous observations at each time t.

Let v denote any consumer or (non-bankrupt) firm agent in the DM Game. At each time $t \geq 0$, agent v forms normal probability distributions for the market-clearing wage w, the market-clearing goods price p, and the dividend payment div in current and future periods. These distributions are characterized by state-conditioned estimates for their means and variances, as follows:

$$w \sim \mathcal{N}\left(\bar{w}_{v,t-1}, \sigma^{2\ L}_{v,t-1}\right) \tag{45}$$

$$p \sim \mathcal{N}\left(\bar{p}_{v,t-1}, \sigma^{2\ G}_{v,t-1}\right) \tag{46}$$

$$div \sim \mathcal{N}\left(\bar{d}_{v,t-1}, \sigma^{2\ D}_{v,t-1}\right) \tag{47}$$

After the determination of a market-clearing wage $w_{t:1}$ in the forward labor market at time t:1, a market-clearing price $p_{t:3}$ in the goods market at time t:3, and a dividend payment $div_{t:5}$ at time t:5, agent v updates the means and variances for these distributions in order to obtain updated estimates for these distributions for use in period $t+1$.[15]

The method used to obtain updated mean and variance estimates for the wage distribution (45) is characterized by the following three parameters: a prior wage $w_{v,0}$; a prior weight $n^L_{v,0}$, and a memory length wm. If $wm = all$, then agent v calculates these estimates as follows:

$$\bar{w}_{v,t} = \frac{\sum^{t}_{r\ =\ 0} w_{r:1} \ +\ n^L_{v,0} \cdot w_{v,0}}{t\ +1\ +\ n^L_{v,0}} \tag{48}$$

$$\sigma^{2,L}_{v,t} = \frac{\sum^{t}_{r\ =\ 0} \left(w_{r:1}\ -\ \bar{w}_{v,t}\right)^2\ +\ n^L_{v,o} \cdot \left(w_{v,0}\ -\ \bar{w}_{v,t}\right)^2}{t\ +1\ +\ n^L_{v,0}} \tag{49}$$

In other words, the mean of the distribution for the expected market-clearing wage is determined by averaging all market-clearing wages observed to date, together with the prior wage, while the dispersion of the expected market-clearing wage is determined by averaging the squares of the deviations of the observed market-clearing wages and the prior wage from the currently estimated mean market-clearing wage.

If $wm = 1$, then agent v sets the expected market-clearing wage equal to the most recently observed

[15] In this updating procedure it is assumed for simplicity that consumers and firms ignore the fact that the actual wages determined at the contract settlement time t:4 could differ from the market-clearing wage $w_{t:1}$ determined at time t:1 if some firms become bankrupt at time t:3 due to poor performance in the time-t:3 goods market.

market-clearing wage:

$$\bar{w}_{v,t} = w_{t:1} \tag{50}$$

Also, agent v sets the expected variance equal to 1% of this expected market-clearing wage:

$$\sigma^{2,L}_{v,t} = 0.01 \cdot \bar{w}_{v,t} \tag{51}$$

Similar equations are used to obtain updated estimates $\bar{p}_{v,t}$, $\sigma^{2,G}_{v,t}$, $\overline{div}_{v,t}$, and $\sigma^{2,D}_{v,t}$ for the means and variances for the goods price distribution (46) and the dividend distribution (47) for $wm = all$ and $wm = 1$, with $p_{r:3}$ or $div_{r:5}$ replacing $w_{r:1}$, $p_{v,0}$ or $div_{v,0}$ replacing $w_{v,0}$, and $n^G_{v,0}$ or $n^D_{v,0}$ replacing $n^L_{v,0}$.

The estimated means $\bar{w}_{v,t}$ and $\bar{p}_{v,t}$ for the market-clearing wage and goods price are used to determine the reservation wage and reservation price for agent v's transformation function mapping described in Sections 4.2 and 4.3. Specifically, $E_{v,t}[w_{r:1}] = \bar{w}_{v,t-1}$ and $E_{v,t}[p_{r:3}] = \bar{p}_{v,t-1}$ for all $r \geq t$. Thus equations (14), (17), and (18) take the form

$$w^c_{i,r:1}(d,t) = \omega \cdot \bar{w}_{i,t-1} \tag{52}$$

$$w^f_{j,r:1}(d,t) = \gamma \cdot \bar{w}_{j,t-1} \tag{53}$$

$$p^f_{j,r:3}(d,t) = \lambda \cdot \bar{p}_{j,t-1} \tag{54}$$

As clarified below in Section C, the EO-FH and EO-ADP agents make use of the full probability distributions (45) through (47) in their decision rules. The updating of these distributions requires specifications for prior variance values as well as prior mean values.

A complete listing of the maintained values for all of the prior-belief parameters is given in Table 16.

Parameter	Value
$w_{v,0}$	1.00
$p_{v,0}$	1.00
$div_{v,0}$	0.00
$n^L_{v,0}$	10.00
$n^G_{v,0}$	10.00
$n^D_{v,0}$	0.00
$\sigma^{2\ L}_{v,0}$	0.50
$\sigma^{2\ G}_{v,0}$	0.50
$\sigma^{2\ D}_{v,0}$	0.01

Table 16: Maintained values for prior-belief parameters

C Implementation of EO Decision Rules

Various computational approximation methods could be used to implement the EO-FH and EO-ADP decision rules. The methods used in this study are outlined below. Detailed explanations of these methods can be found in Powell (2011).

C.1 Implementation of the EO-ADP Decision Rule

Consumers in the DM Game have the same general form of budget and feasibility constraints (1)-(4) for periods $r \geq 0$, the same general form of intertemporal utility objective function (5) for periods $t \geq 0$, and the same single-period utility function $u(\cdot)$ given by (37). The state $\mathbf{x}_{i,t}$ of any consumer i at any time $t \geq 0$ is given by:

$$\mathbf{x}_{i,t} \;=\; \left[t, M^c_{i,t-1}, \bar{w}_{i,t-1}, \sigma^{2\ L}_{i,t-1}, \bar{p}_{i,t-1}, \sigma^{2\ G}_{i,t-1}, \bar{div}_{i,t-1}, \sigma^{2\ div}_{i,t-1}\right] \tag{55}$$

The dimension of the state (55) is fixed at eight, independently of i and t. The normality assumptions imposed on the wage, price, and dividend distributions (45) through (47) imply that each of these distributions is fully characterized in each period t by its estimated mean and variance appearing in (55).

The value function for consumer i at time t in state $\mathbf{x}_{i,t}$ takes the form:

$$V^c\left(\mathbf{x}_{i,t}\right) \;=\; \max_{d \in D^c} E_{i,t} \sum_{r=t}^{\infty} \beta^{r-t} \left[u\left(q^c_{i,r:3}(p_{r:3}, d, t), 1 - l^c_{i,r:1}(w_{r:1}, d, t)\right)\right] \tag{56}$$

The right-side maximization in (56) is constrained by the budget and feasibility constraints (1)-(4) for periods $r \geq t$, conditional on $\mathbf{x}_{i,t}$, and implicitly depends on the $\mathbf{TR}^c_{i,t}$ function that maps each potential period-t decision $d \in D^c$ into a sequence of labor supply and goods demand functions for periods $r \geq t$. The expectation in (56) is taken with respect to the wage, price, and dividend probability distributions (45) through (47), conditional on $\mathbf{x}_{i,t}$.

The structure of the state transition function $\mathbf{S}^c$ mapping each possible state $\mathbf{x}_{i,t}$, decision $d \in D^c$, and realization $(w_{t:1}, p_{t:3}, w_{i,t:4}, div_{t:5})$ into an updated state $\mathbf{x}_{i,t+1}$ for period $t+1$ is time invariant and the same for all consumers i. Also, the left-side summation in (56) is time separable. Consequently, the value function $V^c(x_{i,t})$ can equivalently be expressed in recursive form, as follows:

$$\begin{aligned} V^c\left(\mathbf{x}_{i,t}\right) = \max\nolimits_{d \in D^c} E_{i,t} \big[& u\left(q^c_{i,t:3}(p_{t:3}, d, t), 1 - l^c_{i,t:1}(w_{t:1}, d, t)\right) \\ & + \;\; \beta V^c\left(\mathbf{S}^c\left(\mathbf{x}_{i,t}, d, w_{t:1}, p_{t:3}, w_{i,t:4}, div_{t:5}\right)\right)\big] \end{aligned} \tag{57}$$

We assume that each EO-ADP consumer i at each time t derives an estimate for the value function

(56) that solves the recursive relationship (57) by means of a type of *adaptive dynamic programming (ADP)* algorithm surveyed in (Powell, 2011, p. 407). The latter algorithm, designed for infinite-horizon dynamic programming problems, is an approximate policy iteration method implemented by means of least-squares temporal differencing. During this value function estimation at time t, the mean and variance estimates $\bar{w}_{i,t-1}$, $\sigma^{2\ L}_{i,t-1}$, $\bar{p}_{i,t-1}$, $\sigma^{2\ G}_{i,t-1}$, $\bar{d}_{i,t-1}$, and $\sigma^{2\ div}_{i,t-1}$ in consumer i's state $\mathbf{x}_{i,t}$ are held fixed. No new information is obtained by consumer i during his value function estimation, so he does not update his state information during this estimation.

A critical step in the EO-ADP algorithm at each time t is the selection of basis functions for approximating the general form of the value function prior to conducting the value function estimation. We assume each EO-ADP consumer i at each time t uses a single linear basis function, as follows:

$$V^c(\mathbf{x}_{i,t}) \ = \ \sum_k \theta^\pi_k \phi_k \left(\mathbf{x}_{i,t}\right) \ = \ \theta^\pi \cdot M^c_{i,t-1} \tag{58}$$

where $M^c_{i,t-1}$ denotes the time-t money balance of consumer i. The value function estimation problem at time t thus reduces to the estimation of the scalar parameter θ^π over some specified domain, which in this study is taken to be the interval $[0.01, 1000]$.

It is assumed that EO-ADP firms use a similar EO-ADP decision rule to estimate their time-t value functions. The state $\mathbf{x}_{i,t}$ of a non-bankrupt EO-ADP firm j at time t is given by

$$\mathbf{x}_{j,t} \ = \ \left(t, M^f_{i,t-1}, \bar{w}_{j,t-1}, \sigma^{2\ L}_{j,t-1}, \bar{p}_{j,t-1}, \sigma^{2\ G}_{j,t-1}\right) \tag{59}$$

and its value function is given by

$$V^f_t \left(\mathbf{x}_{j,t}\right) \ = \ \max_{d \in D^f} E_{j,t} \sum_{r=t}^{\infty} \mu^{r-t} \left[p_{r:3} q^f_{j,r:3}(p_{r:3}, d, t) - w_{r:1} l^f_{j,r:1}(w_{r:1}, d, t)\right] \tag{60}$$

The right-side maximization in (60) is constrained by the technological and feasibility constraints (6)-(11) for periods $r \geq t$, conditional on $\mathbf{x}_{j,t}$, and implicitly depends on the $\mathbf{TR}^f_{j,t}$ function that maps each potential period-t decision $d \in D^f$ into a sequence of labor demand and goods supply functions for periods $r \geq t$. The expectation in (60) is taken with respect to the wage and price probability distributions (45) and (46), conditional on $\mathbf{x}_{j,t}$.

For reasons analogous to arguments given above for EO-ADP consumers, the value function (60) for firm j can be expressed in the following recursive form:

$$\begin{aligned} V^f \left(\mathbf{x}_{j,t}\right) \ = \ \max_{d \in D^f} E_{j,t} \Big[& p_{t:3} q^f_{j,t:3}(p_{t:3}, d, t) - w_{t:1} l^f_{j,t:1}(w_{t:1}, d, t) \\ & + \ \beta V^f \left(\mathbf{S}^f \left(\mathbf{x}_{j,t}, d, w_{t:1}, p_{t:3}\right)\right)\Big] \end{aligned} \tag{61}$$

where the structure of the state transition function $\mathbf{S}^f$ does not depend on j or t. Firm j at time t

is assumed to use a simple linear basis function to estimate the value function $V^f(\mathbf{x}_{j,t})$ that solves (61), as follows:

$$V^f(\mathbf{x}_{j,t}) \;=\; \sum_z \theta_z^\pi \phi_z\left(\mathbf{x}_{j,t}\right) \;=\; \theta^\pi \cdot M_{j,t-1}^f \tag{62}$$

where $M_{j,t-1}^f$ denotes the money balance of firm j at time t.

The following parameters need to be specified in order to implement the EO-ADP algorithm for EO-ADP consumers and EO-ADP firms: the number of runs for the inside and outside estimation loops; the number of random number draws in an internal maximization algorithm ; the number of basis functions; the initial parameter value B^0 for recursive least squares estimation (dependent on I = number of consumers); and the initial parameter value $\theta^{\pi,0}$ for the coefficient in the basis-function representation of the value function. These parameters are maintained at the fixed values listed in Table 17 for all EO-ADP agents. The tested values for the two EO-ADP treatment factors, wm and grid-type, are given in Table 10.

Parameter	Value
EstRunIn	5
EstRunOut	5
BasisNum	1
NDrawsADP	5
B^0	$0.005 \cdot I$
$\theta^{\pi,0}$	1.0
β	0.95

Table 17: Maintained parameter values for EO-ADP agents

C.2 Implementation of the EO-FH Decision Rule

The EO-FH algorithm is a brute-force method for the direct estimation of an optimal solution in each period t over a finite rolling forecasting-horizon T. It is performed by EO-FH consumers and firms by undertaking a complete search of their finite decision domains, with a corresponding evaluation of expected outcomes over the next T periods, in order to determine a decision achieving the maximum possible expected intertemporal utility or profit outcome over these next T periods. Thus, in contrast to the EO-ADP algorithm, the EO-FH algorithm does not involve estimation over an infinite horizon, and it does not involve the use of value functions. Consequently, it is conceptually simpler and faster to implement than the EO-ADP algorithm.

Specifically, each EO-FH consumer i at each time t in some state $\mathbf{x}_{i,t}$ uses direct search to solve an optimization problem identical in form to (56) except that the infinite horizon is replaced by a finite horizon $t+T$. Similarly, each non-bankrupt EO-FH firm j at each time t in some state $\mathbf{x}_{j,t}$ uses direct search to solve an optimization problem identical in form to (60) except that the infinite horizon is replaced by a finite horizon $t+T$.

The EO-FH consumers and firms at each time t use Monte Carlo simulation to calculate the expectations in their finite-horizon maximization problems by taking NDrawsFH draws from each of their estimated probability distributions (45), (46), and (47). The value of the parameter NDrawsFH is maintained at 10 for all EO-FH agents. The tested values for the three EO-FH treatment factors T, wm, and grid-type are given in Table 9.

D Social Planner Benchmark Model Solution

This section provides a proof by induction that the Social Planner (SP) Benchmark Model in reduced representative-consumer form (41) has the following solution: $l^c_{t:1} = q^c_{t:3} = 1$ and $s^{stock}_t = 0$ for all periods $t \geq 0$.

By assumption, $s^{stock}_{-1} = 0$. Given this assumption, the social planner's optimal choices for labor, consumption, and goods stock for period 0 are given by $l^c_{0:1} = q^c_{0:3} = 1$ and $s^{stock}_0 = 0$. To establish this, first note that leisure $le^c_{0:1} = [1 - l^c_{0:1}]$ has a constant marginal utility equal to 0.5 whereas goods consumption $q^c_{0:3}$ over the range $(0, 1]$ has a marginal utility that is bounded below by 1.5. Consequently, the social planner will set $le^c_{0:1} = 0$ (hence $l^c_{0:1} = 1$). Given the production function assumptions for the SP Benchmark Model, the maximum amount of good that can be produced in period 0 is thus 1 unit.

Now suppose the social planner contemplates setting aside a portion $s^{stock}_0 \in [0, 1]$ of this period-0 production as goods stock for period 1. Given s^{stock}_0, the maximum utility achievable in period 0 by the representative consumer is $3.0 \ln(2 - s^{stock}_0)$ if $s^{stock}_0 < 1$ and $3.0 \ln(0.5)$ if $s^{stock}_0 = 1$. Also, given s^{stock}_0, the maximum utility achievable by the representative consumer in period 1 is then attained by setting $l^c_{1:1} = 1$, allocating all of the resulting period-1 production of 1 unit of good towards time-1:3 consumption, and allocating all of the goods stock s^{stock}_0 towards time-1:3 consumption,. From the standpoint of period 0, the resulting maximum utility achievable by the representative consumer in period 1 is thus given by $\beta[3.0 \ln(2 + s^{stock}_0)]$. However, since β is less than 1, the sum of these two maximum achievable utility levels,

$$3.0 \ln(2 - s^{stock}_0) + \beta \cdot \left[3.0 \ln(2 + s^{stock}_0)\right] \quad , \tag{63}$$

is a strictly decreasing function of s^{stock}_0 over $s^{stock}_0 \in [0, 1)$ (with a discontinuous further jump down at $s^{stock}_0 = 1$). Consequently, the maximum achievable intertemporal utility for the representative consumer over periods 0 and 1, considered together, is obtained by setting $s^{stock}_0 = 0$. Similar arguments can be used to argue that no future use of a positive s^{stock}_0 can result in a (discounted) utility gain for the representative consumer that outweighs his resulting loss of period-0 utility. Consequently, the social planner should set $s^{stock}_0 = 0$.

Now consider any arbitrary period $t \geq 0$ for which the goods stock s^{stock}_{t-1} is zero. Then the same

argument used above can be applied to period t to show that the social planner's optimal choices for period t are to set $l^c_{t:1} = q^c_{t:3} = 1$ and $s^{stock}_t = 0$. It follows by induction that the optimal solution to the SP Benchmark Model (41) is $l^c_{t:1} = q^c_{t:3} = 1$ and $s^{stock}_t = 0$ for all periods $t \geq 0$.

E Performance Metrics for Case Comparisons

Let Nk denote any of the tested cases in Table 3. This section describes the various performance metrics used to evaluate the performance of the DM-Game economy under case Nk.

The primary indicator used to measure ex post performance is $\bar{u}^k$, the *average realized single-period utility* attained by the I DM-Game consumers. Using notation introduced in Section 6.1, and recalling that the initial period is numbered 0, $\bar{u}^k$ is calculated as follows:

$$\bar{u}^k = \frac{\sum_{i=1}^{I} \sum_{\tau=LOmit}^{LRun} \sum_{r=1}^{NRuns} u^k_{i,\tau,r}}{I \cdot (LRun - LOmit + 1) \cdot NRuns} \tag{64}$$

where $u^k_{i,\tau,r}$ is the utility attained by consumer i in period τ of run r.

Some use is also made of additional performance metrics. For each period $\tau \in \{\text{LOmit}, \ldots, \text{LRun}\}$, the *average realized single-period utility for period* τ is calculated as follows:

$$\bar{u}^k_\tau = \frac{\sum_{i=1}^{I} \sum_{r=1}^{NRuns} u^k_{i,\tau,r}}{I \cdot NRuns} \tag{65}$$

The average value of $\bar{u}^k_\tau$ across the time periods $\tau \in \{\text{LOmit}, \ldots, \text{LRun}\}$ is then given by (64), and the standard deviation of $\bar{u}^k_\tau$ across these same time periods is given by

$$\sigma_{\bar{u}^k} = \left(\frac{\sum_{\tau=LOmit}^{LRun} \left(\bar{u}^k_\tau - \bar{u}^k\right)^2}{LRun - LOmit + 1} \right)^{1/2} \tag{66}$$

The *average realized cumulative utility through period* t is calculated as follows for periods $t \geq$ LOmit:

$$\bar{u}^{cumul,k}_t = \frac{\sum_{\tau=LOmit}^{t} \bar{u}^k_\tau}{t - \text{LOmit} + 1} \tag{67}$$

Suppose that a market-clearing wage $w^k_{t:1,r}$ and a market-clearing goods price $p^k_{t:3,r}$ are both well-defined[16] for some period t for all runs $r \in R^*$, where the subset R^* has cardinality $NRuns^*$. Then

[16]Since the demands and supplies of the DM-Game consumers and firms depend on reservation wages and prices, there can exist periods for which all of these agents decide not to participate in the labor market and/or the goods market.

the *average realized real market-clearing wage for period t* is calculated as follows:

$$\bar{w}_t^{real,k} \;=\; \frac{\sum_{r=1}^{NRuns^*}\left[\frac{w_{t:1,r}^k}{p_{t:3,r}^k}\right]}{NRuns^*} \tag{68}$$

The *average realized real market-clearing wage* $\bar{w}^{real,k}$ is then calculated as the average of $\bar{w}_t^{real,k}$ over all periods t for which $\bar{w}_t^{real,k}$ is well defined.

Finally, in analogy to (64), the *average realized single-period profits* attained by the J DM-Game firms is calculated as follows:

$$\bar{\pi}^k \;=\; \frac{\sum_{j=1}^{J}\sum_{\tau=LOmit}^{LRun}\sum_{r=1}^{NRuns}\pi_{j,\tau,r}^k}{J\cdot(LRun-LOmit+1)\cdot NRuns} \tag{69}$$

where $\pi_{j,\tau,r}^k$ denotes the profit attained by firm j in period τ of run r.