PIER Working Paper

21-027

An Economy of Neural Networks: Learning from Heterogeneous Experiences

ARTEM KURIKSHA
University of Pennsylvania

November 19, 2021

An Economy of Neural Networks:
Learning from Heterogeneous Experiences

Artem Kuriksha*

University of Pennsylvania

Abstract

This paper proposes a new way to model behavioral agents in dynamic macro-financial environments. Agents are described as neural networks and learn policies from idiosyncratic past experiences. I investigate the feedback between irrationality and past outcomes in an economy with heterogeneous shocks similar to Aiyagari (1994). In the model, the rational expectations assumption is seriously violated because learning of a decision rule for savings is unstable. Agents who fall into learning traps save either excessively or save nothing, which provides a candidate explanation for several empirical puzzles about wealth distribution. Neural network agents have a higher average MPC and exhibit excess sensitivity of consumption. Learning can negatively affect intergenerational mobility.

Keywords: heterogeneous experiences, rational expectations, bounded rationality, policy learning, artificial intelligence.

1. Introduction

This paper aims to achieve three goals. First, I suggest a new model of bounded rationality in macroeconomics. This model comes from the family of algorithms that successfully learn to perform many important human-like tasks. It also imposes almost no functional restrictions on agents. Second, I attempt to shed light on the problem of learning a decision rule over time, which is different from most of the learning that economics traditionally focuses on. Third, the developed model allows me to investigate the feedback between imperfect decision making and heterogeneous experiences, as policy functions are formed endogenously and depend on agents' past. I focus on the feedback between mistakes in saving and experiences with earnings and wealth. This feedback could have important implications for the economy, but has attracted virtually no attention in the literature.

Bounded rationality and heterogeneous experiences. Figure 1 illustrates the feedback I explore in this study. Consider a model with heterogeneous agents who are subject to idiosyncratic shocks. With rational expectations, only the left arrow is important for the economy. But when perfect rationality is substituted with an experience-based learning process, economic outcomes vary not only because of the shocks but also because of the learned policies. Agents learn to make their choices over time based on what they have observed in the past. Past experiences are heterogeneous across agents and depend both on idiosyncratic shocks and choices made. This creates a feedback between economic outcomes and imperfections in economic choices. In this paper, I suggest a way to model the top right arrow using neural networks, and investigate the feedback that occurs when learning is incorporated into consumption-saving decisions.
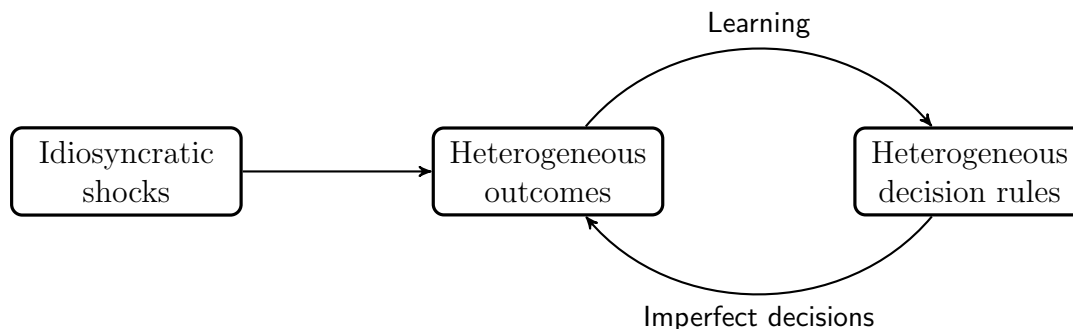
Fig. 1. Feedback between choices and outcomes.

There are two important reasons why idiosyncratic experiences might affect decision making. The first one is formation of expectations. Even if they are perfectly rational in

solving their optimization problems, agents with different expectations can make different choices in otherwise identical situations. At the same time, there is substantial evidence that agents with different experiences do have different expectations even about the most crucial and widely discussed economic or financial[1] variables. For example, Nagel and Xu (2019) find that lifetime inflation experiences effectively predict inflation expectations for individuals in the US. Goldfayn-Frank and Wohlfart (2020) show that East Germans still have higher inflation expectations than West Germans. Kuchler and Zafar (2019) find that personal experience with employment or the local housing market affects a person's expectations about unemployment or the housing market nationwide. Malmendier and Nagel (2011) find that those who experienced higher stock returns are more willing to take risks in the future, suggesting formation of expectations as the channel. Choi et al. (2009) obtain a similar finding for 401(k) savings. Malmendier and Nagel (2016) use a model of learning with fading memory to explain patterns in expectations about equity market returns.

The second reason for agents to have different decision rules is that these decision rules are products of a learning process that is endogenous and imperfect. Agents who have different pasts have learned how to act in different environmental circumstances. An agent might be unaware of the right mode of behavior in a situation that is unfamiliar to her. Even if learning was successful for some region of the state space, extrapolation to new regions could be problematic. For example, in the consumption-saving problem, agents that are used to low levels of wealth might fail to save optimally when they suddenly become rich. Similarly, agents that are used to a wealthy life might fail to save correctly when their wealth is suddenly gone. This mechanism also works alongside multiple cognition limitations. For example, agents might have limited memory capacity. They also can overreact to personal experiences or be unable to learn the right way to save in rare events such as unusually high values of productivity.

The learning process suggested here incorporates both of these reasons. Policy updating is directed toward solving the dynamic decision problem, and thus expectations[2] are implicitly taken into consideration. The learning process reflects cognitive limitations and is endogenous and based on past experiences. Thus, the learning artifacts described above are also present.

Learning model. Each agent faces the income fluctuation problem and has a saving decision rule that is parameterized. Learning takes the form of updating the parameters, which

---

[1]For a psychological reference, see, for example, Hertwig et al. (2004).

[2]In this paper, the uncertainty is about individual productivity in the future. The model can be generalized to allow agents to choose from a menu of risky assets. In that case, expectations about returns would become important too.

might be different across agents. Specifically, the policy is given by a neural network, and the parameters are its weights. Agents have limited memory of their past. Every period, an agent takes several learning steps. She thinks about what her consumption could have been in the past. The reasoning that she engages in can be informally described as "I would be happier if I would do this instead of that." Subsequently, the agent makes an incremental update of the parameters.

This model of learning has several advantages. First, it exhibits properties that one naturally expects from a human-like agent in a dynamic environment. At the same time, it is indeed a process of learning that is useful for agents as it is "consistent." I show that if an agent has favorable hyperparameters (that is, the parameters of the model that are not changed while learning, e.g., the memory size) and learns from experiences that are sufficiently informative about the state space, then the updating rule converges to very good approximations of the optimal policy.

Second, this model is largely inspired by and related to deep reinforcement learning. Reinforcement learning (RL) studies learning guided by rewards and with the goal of choosing optimal actions in an environment.[3] Incorporating deep neural networks into reinforcement learning (deep RL) has provided a number of breakthroughs in artificial intelligence in recent years. Moreover, neural networks now can achieve a previously unprecedented performance in human-like tasks like image, text, or speech recognition. All that makes neural networks and deep RL a very promising tool for modeling human behavior in economics.

Third, this model relies on few assumptions about the agents. For instance, I do not rely on many of the "restrictive assumptions" that are typical in behavioral macroeconomics. I do not assume anything about cognition and attention costs or the structure of incomplete information. Also, I do not assume that agents perform hard cognitive operations. Here, agents do not operate with expectations or other statistical concepts. Neither do they solve any maximization problem or recursive equation. Further, neural networks in principle can approximate almost any function, and thus no restrictive functional assumptions are imposed. Last but not least, the method is generalizable to other dynamic macroeconomic or financial contexts. See Section 3 for the technical description and a detailed discussion of the learning model.

Findings. The suggested model allows me to see how agents with potentially different levels of rationality learn to make consumption-saving decisions over time. A major theme of all results I obtain is that learning is a hard task. Learning is prone to instabilities, traps, and

---

[3]See Sutton and Barto (2018) for an introduction to reinforcement learning. See Botvinick et al. (2020) for a review of recent advancements in deep RL and connections of deep RL with neuroscience.

biases, which require additional components and assumptions as a remedy. It is a non-trivial task for agents to find a good policy even when they can dynamically experiment with their saving decisions and when they are guided by their previous consumption experiences, which are informative about the stationary Bewley–Huggett–Aiyagari environment the agents live in. This suggests that the assumption of rational expectations is a strong assumption. The model also allows me to estimate the cost of irrationality, which for the main specification I consider is roughly equivalent to one-tenth of consumption.

The issues that agents face while learning are not specific to this particular task; they also arise in almost all modern applications of RL. This indicates that they can be inherent to the process of learning, and thus it is important to understand their implications for economics. I find that for the consumption-saving problem, they often lead to two particular types of behavior: saving nothing or saving excessively. Agents who are "lost while learning" in these two ways make the outcomes of the model closer to the data along several margins. They comprise a mass of agents who are persistently stuck around the liquidity constraint (so-called hand-to-mouth status) or they comprise a thicker upper tail in the wealth distribution. This and the fact that heterogeneity in decision making is an additional source of heterogeneity in outcomes increase inequality in the economy. The ones who save nothing also make the average marginal propensity to consume in the economy higher. Another learning artifact I find is inattention to productivity shocks: some agents fail to learn to react to them. This might have implications for economic policy, and in the model it contributes to excess sensitivity of consumption.

The learning process itself also has some interesting properties. Neural network agents make better decisions in familiar regions of the state space, but their biases and mistakes often push them into the regions where their decision rules are severely suboptimal. In particular, descendants of parents with little wealth on average know better how to save when they have little savings. However, they often do not know how to save correctly in the "rich" region of the state space. Surprisingly, that makes them more likely to end up being over-savers. I observe a symmetric pattern for descendants of parents with high wealth, who are more likely to end up with no savings.

Finally, the noise from mistakes in consumption can decrease the role of wealth inherited from parents. That and the mechanism described above can equalize agents in a succeeding generation. But I also find that in some cases learning can substantially decrease intergenerational wealth mobility. It happens because learning traps might be transmitted through generations. The bulk of the effect comes from the descendants of over-saving irrational agents, who often keep over-saving and generate even more wealth. This is consistent with the data, where large wealth also tends to have high persistency across generations.

Related literature. Methodologically, this paper is a part of the growing number of applications of machine learning in macroeconomics. Duarte (2018), Azinovic et al. (2019), Fernández-Villaverde et al. (2019), Maliar et al. (2019), Fernández-Villaverde et al. (2020), and Kahou et al. (2021) use neural networks to approximate equilibrium objects and solve computationally-hard problems with high-dimensional state and control spaces. Though bounded rationality and learning are not the objects of interest of any of those papers, the idea to update policy parameters using first-order optimality conditions is related to the learning process considered here.

This paper also utilizes multiple ideas from deep reinforcement learning. The number of applications of reinforcement learning in economics is small but is likely to grow fast. Most of the existing applications are related to allocation of a binary treatment, as in Athey and Wager (2021) or Adusumilli et al. (2019). This paper is among the very first attempts to use reinforcement learning in macroeconomics or for a continuous economic problem. Shi (2021) also does that and solves the stochastic optimal growth model using deep reinforcement learning. Chen et al. (2021) solve a model with monetary and fiscal policy and show that a deep reinforcement learning algorithm can recover steady states that are not learnable with adaptive learning. The latter two papers also argue that neural networks can be a modeling device for bounded rationality. That said, both keep unexplored many of the economic implications of bounded rationality. The main finding there is that RL algorithms can solve control problems from macroeconomics. This paper, in contrast, focuses more on the economic implications of imperfect learning and the interpretability and plausibility[4] of the assumptions behind learning agents.

Another related project is Zhao et al. (2020). It also uses neural networks as a device to model behavioral economic agents. However, the focus there is very different. Neural networks are used to model not decision rules but utility, no learning happens, and the macroeconomic implications are not investigated.

Conceptually, this project is another attempt to analyze the effects of bounded rationality on the macroeconomy. The vast majority of the literature, however, focuses on the frictions in agents' ability to observe or forecast aggregate variables like the interest rate or inflation. One

---

[4]In particular, the approach implemented in this paper allows for learning in a reasonable amount of time. In contrast, in Shi (2021) and Chen et al. (2021) the agents need hundreds of thousands or even millions of periods to converge to the rational behavior (see Section 3.2 for a discussion of this issue). In Chen et al. (2021), for example, the first phase over which the learning error *increases* takes several hundreds of thousands of periods. This is problematic if one wants to interpret agents as humans. Further, it might be hard to interpret why real-world agents would have multiple learning episodes. Finally, both papers are concerned with learning by a representative agent. At the same time, heterogeneity is a desirable feature for a macroeconomic model of learning and bounded rationality. Learning mistakes are likely to be heterogeneous across agents.

approach is to postulate the slow spreading of information, as happens in Mankiw and Reis (2002). A particularly large class of models follows Sims (2003). In these models, learning the true state of the world is costly, and having limited information becomes rational. Deviations from rational expectations can arise then both on the side of firms, as in Mackowiak and Wiederholt (2009) or Stevens (2020), and on the side of consumers, as in Gabaix (2016). Another direction is represented by Farhi and Werning (2019) and García-Schmidt and Woodford (2019), who allow for a limited ability of agents to analyze general equilibrium effects while forecasting.

This literature mainly revolves around various questions about monetary policy. The center of attention is agents' ability to process unexpected shocks to inflation and interest rates. After these frictions are considered and the expectations are formed, agents always behave rationally.[5] Thus, the possibility of imperfect decision making per se is set aside. Because of that and because the information frictions in these models typically are not experience-dependent, there is no place for a feedback between heterogeneous experiences and bounded rationality. This paper, in contrast, abstracts from the potential inability of agents to observe or forecast aggregate variables. Those are constant in the model and agents know them perfectly. The focus is entirely on frictions in decision making and learning decision rules. This learning is experience-dependent, and thus the aforementioned feedback is possible.

An exception is Ilut and Valchev (2020).[6] Similar to this paper, it is also a model where agents observe the state perfectly but make imperfect decisions, trying to learn a good decision rule over time. In Ilut and Valchev (2020), learning happens in a Bayesian way.[7] The framework allows the authors to explain facts that the baseline model fails to fit well: high inequality, high marginal propensity to consume, and a large share of agents with zero wealth. Learning that will be discussed in this paper also allows an explanation of these facts. That said, learning with neural networks has some advantages as a model of bounded rationality. First, the learning model of Ilut and Valchev (2020) is heuristic[8] and

---

[5]Lian (2020) allows for future mistakes in consumption, but those again are rationally processed in the agents' decisions today.

[6]See also the older version Valchev and Ilut (2017).

[7]Uncertainty over the policy function is modeled as a Gaussian process, and this generates local spillovers from learning the true policy in a particular point. Reasoning is costly (agents pay for mutual information between the signal and the posterior), and thus an agent might decide not to update her beliefs if her uncertainty at some state is small.

[8]The assumption is that agents can receive signals about the value of the optimal policy function at points of the state space. However, the optimal policy in a dynamic model should solve the recursive problem, in which maximization happens conditional on behaving optimally in the future. This means that it is hard to separate learning in a particular point from learning the policy for other points of the state space. All of that makes Bayesian learning of a policy function less conceptually justifiable than, for example, Bayesian learning of a demand function in Ilut et al. (2020).

omits many details about the important process of "converting information into a policy function." In that sense, the learning model presented here is more micro-founded. Second, the results about the feedback between learning and the past outcomes in Ilut and Valchev (2020) come from a specific mechanism – agents stick to wrong guesses in states that occur often. This mechanism might depend on the assumptions about the distributions and the cognition costs. Besides, the role of the choice of the prior distribution is also a concern.[9] Finally, the Bayesian approach might not be robust to increases in dimensionality, while real-world agents have high-dimensional state and action spaces.[10] I discuss the advantages of neural networks as a modeling device in Section 3.

Finally, this paper is related to the study of systems consisting of multiple agents who adapt and learn according to some algorithm. Holland and Miller (1991) suggested studying artificial adaptive agents in economics. Back then, the two main tools were classifier systems and genetic algorithms. Arifovic (1996) and Bullard and Duffy (1998) applied this idea in macroeconomics, using genetic algorithm learning in models with overlapping generations. Lettau and Uhlig (1999) used a learning model with classifiers to suggest a behavioral explanation for excess sensitivity of consumption. In Marimon et al. (1990), artificial agents learn to use a medium of exchange. The modern research on multi-agent systems is usually interested in agents modeled as neural networks, which have almost completely superseded other RL techniques. See Zhang et al. (2019) for a review of the literature on multi-agent reinforcement learning. This paper is among the first to analyze the macroeconomy as such a system. Another example is Zheng et al. (2020).

The paper proceeds as follows. Section 2 describes an environment in the spirit of Aiyagari (1994). Section 3 describes in detail the implemented model of behavioral agents. Section 4 describes the behavior of these agents. Section 5 describes the aggregate economic implications. Section 6 concludes.

2.   Economic Environment

The description of the environment is almost identical to the one in Aiyagari (1994). The two main ingredients of the problem that the agents in the economy have are idiosyncratic income fluctuations and a liquidity constraint. They introduce the precautionary saving motive: agents should save or dissave in order to smooth their consumption, and they do not want to be close to the liquidity constraint because smoothing is impossible there. I will

---

[9]In particular, it is suggested by the difference in intuition and interpretations between Ilut and Valchev (2020) (where the prior coincides with the optimal policy) and Valchev and Ilut (2017) (where the prior is constant).

[10]For example, because agents are likely to simultaneously solve other economic problems.

consider two types of agents. Agents of the first type (*rational expectations agents*, or just *RE agents*) will solve this optimization problem in the standard rational way. Agents of the other type (*neural network agents*, or just *NN agents*), which will be introduced in the next section, will learn over time how to save reasonably.

2.1. Agents

There is a unit mass of agents, indexed by i, who derive utility from consumption $c_{i,t}$ of a numeraire good. Agents discount the future at a rate β. Productivity shocks introduce fluctuations in an agent's labor income $wz_{i,t}$, where w is the wage in the economy and $z_{i,t}$ is the productivity of the agent. Levels of productivity are independent across agents but are correlated over time for an individual agent. Each agent can smooth her consumption by saving at the interest rate r, but she is subject to a liquidity constraint $\underline{a}$. Thus, there are two state variables ($a_{i,t}$ and $z_{i,t}$, the wealth and productivity at period t), one control variable ($a_{i,t+1}$, the saving for period $t+1$), and each agent maximizes

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_{i,t}) \tag{1}$$

subject to

$$c_{i,t} = (1+r)a_{i,t} + wz_{i,t} - a_{i,t+1}, \tag{2}$$

$$a_{i,t+1} \geqslant \underline{a}. \tag{3}$$

For the liquidity constraint, I choose the standard conservative option $\underline{a} = 0$. The choices for the utility function and the discount parameter are also standard:

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}, \ \gamma = 2, \ \beta = 0.96. \tag{4}$$

An agent's productivity evolves according to a Markov process, so that

$$\mathbb{P}(z_{i,t+1} = z' | z_{i,t} = z) = \Gamma_{zz'}. \tag{5}$$

The Markov process has 20 states and approximates an AR(1) process given by

$$\ln z_{i,t} = \rho \ln z_{i,t-1} + \sigma \epsilon_{i,t}, \tag{6}$$

where $\epsilon_{i,t} \sim \mathcal{N}(0,1)$ are i.i.d. The persistence parameter $\rho = 0.6$ and the variance parameter $\sigma = 0.3$ are similar to the values considered in Aiyagari (1994). To obtain the states and the

transition matrix Γ for the approximating Markov chain, I use the discretization procedure of Tauchen (1986).

2.2. Production

The production side of the economy is also standard. Output is produced according to

$$Y_t = K_t^\alpha L_t^{1-\alpha}. \tag{7}$$

The capital endowment K_t in the economy is equal to aggregate savings, and the labor endowment L_t comes from the effective labor supply of the agents:

$$K_t = \int_0^1 a_{i,t} di, \tag{8}$$

$$L_t = \int_0^1 z_{i,t} di. \tag{9}$$

Because there is an infinite amount of agents, L_t is time-invariant. Capital depreciates at rate δ. The markets for capital and labor are competitive, and thus the interest rate and the wage are respectively

$$r_t = \frac{\partial}{\partial K_t} K_t^\alpha L_t^{1-\alpha} - \delta, \tag{10}$$

$$w_t = \frac{\partial}{\partial L_t} K_t^\alpha L_t^{1-\alpha}. \tag{11}$$

The values of the productivity parameters are also taken to be standard and equal to $\alpha = 0.33$ and $\delta = 0.1$. There are no aggregate shocks in the economy.

2.3. Rational expectations equilibrium

The benchmark for the analysis is the steady-state equilibrium under rational expectations. I use a standard procedure to find it:

1. Start with a guess for the equilibrium interest rate r;
2. Find the capital K implied by equation 10 and use equation 11 to find the wage w;
3. Using value function iteration, find the policy $\pi^{RE}(a, z)$ that solves the problem of agents for r and w;
4. Find the stationary distribution of savings $a_{i,t}$ for agents following the policy $\pi^{RE}(a, z)$;
5. Find the stationary capital supply $\tilde{K}$ using equation 8;

6. Update r to solve $K - \tilde{K} = 0$ (e.g., using the standard binary search algorithm).

The interest rate in the rational expectations equilibrium is 3.29%.

3. Neural Network Agents

We need a plausible model of human behavior that would allow for bounded rationality in the context of the consumption-saving problem. In the economy described, every period agents should make choices in different points of a continuous state space. Thus, the object they learn is not a finite-dimensional state but a policy function. As humans, agents have a limited memory about their past. A learning model should render memories into a decision rule. This can be done with neural networks, which are systems inspired by the structure of human or animal brains. See Appendix A for a brief introduction to neural networks.

3.1. Learning process

I assume the following model of behavior for neural network agents. At every point, an agent has in mind a stationary policy

$$\pi(\cdot|\theta) : (a_t, z_t) \mapsto a_{t+1}, \tag{12}$$

which maps the current state (the wealth and productivity at period t) into an element of the action space (the saving for period $t+1$). This policy function is given by a neural network with weights θ.[11] Values of θ are heterogeneous across agents and determine which decision rules agents follow. Neural networks have high expressivity, and almost any function can be approximated if an appropriate θ is chosen. See Appendix A for a brief description of neural networks. See Section 3.2 for a discussion of why I choose neural networks and not some other function approximator.

The key assumption that defines the learning process is the rule to update θ. As always happens with neural networks and as is likely to happen in human learning, updates are incremental. At every step, the agent sets new weights θ' to be

$$\theta' = \theta + \alpha \nabla_\theta \sum_{\tau \in E} \left[u(\tilde{c}_\tau) + \beta u(\tilde{c}_{\tau+1}) \right] = \tag{13}$$

---

[11]I clip $\pi(a, z|\theta)$ if the output of the network is below the liquidity constraint. To ensure non-negative consumption, I also clip the output of the network from above if it is too large. I force consumption to be not smaller than the minimal possible labor income. There are two motivations for this. The first one is that no rational agent would ever choose to consume less than wz_{min} at any state. The second is that the ratio of the minimal consumption to the median wage in the model becomes roughly equal to the ratio of the poverty threshold in the US to the median wage in the US.

$$\theta + \alpha \sum_{\tau \in E} \Big(\beta(1+r)u'(\tilde{c}_{\tau+1}) - u'(\tilde{c}_\tau) \Big) \nabla_\theta \pi(a_\tau, z_\tau | \theta), \tag{14}$$

where

$$\tilde{c}_\tau = (1+r)a_\tau + wz_\tau - \pi(a_\tau, z_\tau | \theta), \tag{15}$$

$$\tilde{c}_{\tau+1} = (1+r)\pi(a_\tau, z_\tau | \theta) + wz_{t+1} - \tilde{a}_{\tau+2}. \tag{16}$$

In the above, $\tilde{c}_\tau$ and $\tilde{c}_{\tau+1}$ are counterfactual consumptions that the agent would have if she were to follow the policy $\pi(\cdot|\theta)$. The implied next-period saving $\tilde{a}_{\tau+2}$ is treated as a constant. It is a result of applying a saving policy to the saving choice at τ that would be made before the learning step:

$$\tilde{a}_{\tau+2} = \pi\big(\pi(a_\tau, z_\tau | \theta), z_{\tau+1} | \tilde{\theta}\big). \tag{17}$$

$E \subset \{t - M, \ldots, t - 1\}$ stands for the set of episodes that the agent uses[12] for the learning step.[13] Agents are assumed to remember the information about the previous M experiences $\{(a_\tau, z_\tau)\}_{\tau=t-M}^{t-1}$ they had. Finally, α is the learning rate, which is the parameter that determines the size of the learning step.

Interpretations. Before making a new consumption-saving decision, an agent thinks about some episodes that she remembers. As memories are fading and cognitive limitations are present, only a small number of the recent episodes can be processed at once. Thus, she only thinks about a subset E of her previous experiences. She analyzes *ex-post* what saving choices would be better than the ones that the current policy prescribes. She attempts to make an improvement to her utility flow from the episodes she thinks about, which is given by the summation term in equation 13. After that, she makes an adjustment to her policy and uses the updated policy in the future.

Equation 14 has the following economic interpretation. The parameters are updated either in the direction of increasing savings or decreasing savings at the states from the memory. The direction of the change is given by a scalar, which is the Euler equation error. For example, suppose that an agent remembers some period τ such that the marginal utility of consumption at τ under the current policy is much lower than at $\tau + 1$. The agent understands that this is a case when she could save more wisely. She is starving right after consuming excessively, and this happens because she is not saving enough. The Euler equation error is positive. Equation 14 implies that θ is updated in a way that makes her savings $\pi(a_\tau, z_\tau | \theta)$ larger.

---

[12]The episodes in E are selected randomly. See Section 3.4 for details.

[13]I abuse notation here. Counterfactual consumptions $\tilde{c}_\tau, \tilde{c}_{\tau+1}$ are calculated independently for each index τ of the sum. They are functions of a_τ, z_τ, and $z_{\tau+1}$. Equivalently, the learning rule can be defined in terms of triples $(a_\tau, z_\tau, z_{\tau+1})$ stored in E.

Another interpretation for the updating rule is given by the reinforcement learning literature. Silver et al. (2014) derive an expression for the gradient of the performance objective function with respect to policy parameters. Ideally, agents would use this gradient for all policy updates as it would yield the fastest learning. However, for that they should learn and remember the state-action value function. This seems to be unrealistically hard. Equation 14 can be considered a feasible behavioral approximation to the optimal policy gradient step.[14]

3.2. Discussion

Role of neural networks. In principle, function approximators of other types (e.g., Chebyshev polynomials) could be used to implement the same logic of learning by updating parameters. Neural networks, however, have advantages that make them a much better choice both for agents and a researcher looking for a credible model of agents' behavior. First, in the real world people usually face high-dimensional problems. Agents not only solve for the consumption-saving trade-off but also divide their portfolios among different assets and choose location, job, working hours, etc.[15] Human minds should be able to learn to solve these problems. Neural networks are known to work well in high-dimensional settings for which polynomial approximations seem to be completely infeasible (see Bach, 2017). Not only are these settings high-dimensional, they are also often related to the problems human minds are solving every day: image, speech, or text recognition, physical movements, etc. It seems reasonable to choose the class of algorithms that works universally for human-like tasks as the device to model human agents. Second, neural networks are more stable. For polynomials, there is likely to be a region for which the approximation values are extreme, and it seems unrealistic that a human agent would make decisions like that. Third, neural networks are able to approximate a much wider class of functions (see Cybenko, 1989; Barron, 1993). Fourth, neural networks impose very little prior structure as a modeling device.

---

[14]The related result is called the *deterministic policy gradient theorem.* The gradient equals the expected value of $\nabla_\theta \pi(s|\theta) \nabla_a Q^\pi(s,a)|_{a=\pi(s|\theta)}$. $Q^\pi(s,a)$ is the state-action value, which is the expected reward flow of an agent who is at state s, chooses action a, and follows policy π afterward. Policy gradient algorithms (e.g., DDPG; see Section 3.2) usually attempt to estimate Q separately. In practice, this creates another source of instability and requires a very large sample of experiences. However, in the consumption-saving problem this object is more tractable: $Q^\pi(s,a) = u(\tilde{c}_\tau) + \beta u(\tilde{c}_{\tau+1}) + \ldots$, where the consumption path is given by the state $s = (a_\tau, z_\tau)$, the choice of action $a = a_{\tau+1}$, and the policy π. Equation 14 approximates $\nabla_a Q^\pi(s,a)$ with $\beta(1+r)u'(\tilde{c}_{\tau+1}) - u'(\tilde{c}_\tau)$. With this approximation, the agent utilizes her knowledge of her own utility function u and the budget constraint and ignores recurrent derivatives by holding $a_{\tau+2}$ fixed. The latter means that the agent is not considering how a change in her action today would change her actions next period, two periods from now, three periods from now, and so on. This kind of reasoning is likely to be too complicated for a behavioral agent.

[15]Let alone extremely high-dimensional control problems like walking, speaking, and so on and so forth.

Thus, the findings do not come from particular assumptions about the functional form of policy functions.[16] Last but not least, neural networks are systems inspired by human or animal brains.

Properties. The described learning process exhibits the following properties:

1. *Deviations from rational expectations.* The policy learned by an agent does not have to coincide with the saving policy that maximizes her utility.

2. *Past-dependence.* Agents with different trajectories maximize different functions at every step, and thus end up with different network weights and different policy functions. In particular, different income shocks can make otherwise identical agents learn different decision rules for savings.

3. *Local spillovers.* Neural networks are continuous function approximators. Thus, agents should choose quantitatively similar actions in a couple of states if the two states are close to each other. In the considered environment, the optimal policy is continuous too. Thus, to the extent that an agent was able to learn the policy from the states she observed in the past, she should behave reasonably in states that are not very different from her experiences.[17]

4. *Incremental updating.* When a new experience is processed by an agent, this results in a marginal revision of the policy she had before. As a consequence, even periods that are already forgotten still continue to affect an agent's policy.

5. *Forgetting.* Periods from a long time ago eventually disappear from memory and stop influencing policy updates. Their effect on the policy decays over time, as the agent keeps updating the weights based on newer experiences.

6. *Asymptotic rationality.* All the properties listed above are desirable for a model of bounded rationality. However, one can easily think of a random policy-updating scheme that satisfies these properties but is absolutely meaningless. Such a model of learning would never allow an agent to learn anything useful. What makes the suggested model more credible is that it is *asymptotically rational.* In Appendix B I show that agents can learn the optimal policy with very high precision in the long run if they have hyperparameters that are sufficiently favorable and have in their memory stack observations that are spread densely over the state space.[18]

---

[16]For example, the agents are not assumed to know that the optimal decision rule is monotone, approximately linear in the points that are far from the liquidity constraint, etc.

[17]This property should be amplified by the fact that neural networks usually extrapolate in a linear fashion. The optimal policy in the consumption-saving problem has no non-convexities and can be approximated well locally by a linear function.

[18]In other words, the universe of economies with NN agents (approximately) nests the rational expectations economy if agents are set to have favorable learning parameters.

Arguably, all of these properties are shared by human beings. Thus, even if the exact mechanism of the suggested kind of learning might be questioned, it would be a valuable heuristic. It exhibits realistic properties and can be used to investigate the feedback between irrationality and heterogeneous experiences.

Assumptions. A conceptual advantage of this approach is that it assumes that agents use little information. An agent is only assumed to observe her own savings and labor income, the interest rate, and to know her utility function. It is not assumed, for example, that she knows the parameters of the Markov process behind the productivity levels $z_{i,t}$. Another advantage is that an agent is supposed to make only a relatively simple cognitive effort. The agent does not rely directly on any probabilistic calculations, either in Bayesian or in frequentist fashion. She also does not have to solve the Bellman equation or apply the maximum operator.

The policy an agent learns is a result of an ex-post analysis of her life satisfaction. Learning is entirely backward-looking, and the agent essentially is "preparing to fight the last war." One might doubt that an agent can learn much this way, doing no statistical calculations and not having the Bellman equation in mind. However, in a stationary environment the optimal policy is time-invariant. The past is representative of the future, and optimization over the past yields an object that is good for the future too. In Appendix B, I show that an agent with the updating rule from equation 13 typically converges to the optimal policy if her memories are representative about the state space, her learning parameters are not restrictive, and she has enough time to learn.

Generalizability and limitations. The suggested type of learning is meaningful because the environment in which agents operate is stationary, and repeated attempts to improve counterfactual past experience result in a policy that is good for the future as well. This type of logic might be generalized to other economic choices that are repeated and for which the optimal policy is stable. For example, one could incorporate a menu of risky assets into the model in exactly the same way as was done with the risk-free asset in the model. Alternatively, one can add working hours as a control variable. In that case, the utility flow in equation 13 should be augmented with the disutility from labor, and the gradient step in equation 14 would also include the static labor supply optimality condition. In a similar fashion, neural network agents can in principle substitute rational agents in any DSGE model with tractable optimality conditions.

At the same time, this constitutes a challenge to generalize this method to some other economic decisions. If an agent faces a non-stationary environment, the optimal behavior

in the future might be impossible to learn solely based on the past. Similarly, this type of learning cannot be applied to once-in-a-lifetime decisions like the decision to get higher education. An inherent property of such problems is that an agent cannot learn a good policy by reflecting on her own past. The agent needs to incorporate some external information about the problem (e.g., about rewards following the possible choices) and process this information in a way that is cognitively less trivial than the incremental adjustment of parameters (e.g., by calculating payoff expectations or solving a utility maximization problem).

Differences with other learning approaches.   The problem of making choices dynamically and learning a policy that yields a high expected return is at the core of reinforcement learning. The problem of the consumption-saving trade-off is continuous. The bulk of the methods that are popular these days for continuous control problems are model-free actor-critic methods, e.g., DDPG presented in Lillicrap et al. (2015). However, these methods are designed for very different learning environments, for example, the ones related to robotics. That implies some conceptual differences, because rewards and transitions between states typically are not known in advance. In particular, the process of learning the critic network usually only uses the information from the observed flow of rewards. For the consumption-saving problem this would mean that agents are unrealistically ignorant about the environment. Essentially, they would have to estimate their own utility function over time, and are not even assumed to know that the saving they choose today will be their wealth tomorrow. This flexibility comes at a cost: the algorithms are very sample-greedy. Just as happens in Shi (2021) and Chen et al. (2021), learning a good policy might take a very large number of periods (well above any feasible number of periods in a human life). In addition to that, it is not particularly realistic that behavioral agents store the state-action value function, which is a very complex object, in their minds. All of that makes it problematic to use DDPG and other model-free actor-critic algorithms as a modeling device for human behavior in the context I consider here.

3.3.   Environment

Though general equilibrium can be calculated in principle, I restrict attention to partial equilibrium. That should be enough to provide the main insights about the differences between rational agents and neural network agents. I hold the wage w and the interest rate r fixed at the equilibrium levels for the rational expectations economy. In other words, I investigate the life of neural network agents in the world of rational agents.

Generations.   Naturally, for rational agents all of the presented statistics are calculated for the stationary distribution. For NN agents, however, the limiting distribution for a particular agent might neglect some important dynamics on the learning path. Given that in the real world agents do not live infinitely, mistakes specific to the first periods of learning are likely to be more relevant than some "stationary irrationality" in the model. For this reason, I assume a generational structure about NN agents.

Specifically, I assume that each agent lives for 100 periods. When an agent is born, she has some initial policy parameters $\theta_{i,0}$. She starts to learn according to equation 13 but for the first 20 periods she does not make decisions herself. Instead, her decisions are made by her parent, who follows the rational policy. At age 20, she inherits the wealth and the productivity of the parent, and after that, her savings evolve according to her own decisions $a_{i,t+1} = \pi(a_{i,t}, z_{i,t} | \theta_{i,t})$. This approach is motivated by two considerations. For one thing, the first 20 periods should serve as a "burn-in" and decrease the role of the choice for initialization of weights. For another, this allows me to investigate the role of learning in the intergenerational wealth dynamics (see Section 5.5). To calculate aggregate variables, I use the values for adult NN agents only.[19]

Cap on savings.   Some NN agents will save as much as they can, which results in exponential wealth accumulation. As will be discussed later, this behavior has an interpretation. However, to limit the role of these outliers I restrict possible savings in the economy. The upper limit is set to be roughly twice as large as the maximum observed wealth for RE agents.

3.4.   Other components of the model

I find that learning is subject to instabilities, which will be discussed in more detail in this and the next section. These instabilities say something important about the process of learning and might have economic implications. They are also universal for deep reinforcement learning. Typically, some additional components are added to an algorithm that are not directly related to the main logic but serve the purpose of improving the convergence properties. Because a model in which agents only have pathological behavior would be unrealistic and provide little insight, I equip NN agents with some modifications too.

---

[19]Thus, what I obtain are statistics for the subpopulation of NN agents assuming that they are rarely born among rational agents and there is an equal mass of NN agents of each age. RE agents are either immortal or arrive in altruistic generations.

Learning frequency. A single gradient step results only in a small update of the network weights. If the initial policy was far from the optimal policy, it would require a large number of steps to approximate the latter well even if the updates are always in the right direction. To make agents plausibly successful in learning, I assume that they take several learning steps per period. The interpretation for this is that agents are likely to think about their consumption-saving decisions more than once per period. This is especially realistic in a model where one period stands for a year and the corresponding consumption-saving decision should be a result of many smaller consumption decisions that an agent makes within this period. Another interpretation is that agents think only once, but each "round of thinking" consists of multiple learning steps.

Memory structure. In equation 13, the set of episodes E that an agent uses for learning has the following structure. Each agent has a memory buffer that contains her last M experiences. At each learning step, a random subsample of those experiences is selected. This technique is called experience replay and helps to mitigate the problem of correlation in the sample of experiences. See Mnih et al. (2015) for a discussion and a biological interpretation of experience replay. The size of E is called the *batch size* in the literature. Potentially, agents might also use the experiences of other agents for learning updates.

External experiences. I find that stability of learning depends crucially on how informative is the sample of experiences about the state space as a whole.[20] At the same time, the experiences from an agent's past are usually clustered in one particular region of the state space. This is so even for rational agents, as wealth tends to be quite persistent both in the real world and in the Aiyagari economy. For agents on the path of learning, this problem can be exacerbated if their policies are degenerate (e.g., always prescribe no saving) and lead to even less exploration of the state space. I consistently find that these issues can undermine learning with a very high probability.

I assume that each agent considers a fraction of experiences in E that are not her own and instead are generated randomly.[21] There are two complementary reasons for an agent

---

[20]This is not surprising. The problem of each agent is dynamic. To make a good saving decision at some point of the state space, an agent should take into consideration what she would choose in the states in which she might be next period. The accuracy of that knowledge in turn depends on the knowledge about other states, and so on. At the end, having a good idea about all points of the state space that might possibly be reached is important for successful learning.

[21]To take a learning step, an agent needs a triple $(a_\tau, z_\tau, z_{\tau+1})$ that describes a level of wealth and a transition of productivity. Thus, I assume that agents have a fraction of randomly generated experiences (a, z, z') in E. Values of z are drawn from the stationary distribution of Γ, and values of z' are drawn conditional on z. Values of a are drawn from the uniform distribution over $[0, 30]$. The support spans all values of savings that RE agents are likely to reach.

to learn from experiences that are not her own. The first interpretation is that those are the experiences of her friends[22] and other members of society.[23,24] When an agent is informed about the income and wealth of a friend, she starts to think what she would do in his shoes. Similarly, seeing a poor or rich person on the street makes her think about what she would do if she were poorer or richer. The second interpretation is that those are hypothetical experiences that an agent considers for the sake of better learning. These thought experiments are needed to learn successfully and it is natural to assume that they are a part of the human reasoning process.

Initialization.   At the beginning, agents start with policies that are close to the "inertial" decision rule. That is, each agent starts with θ_0 such that $\pi(a, z|\theta_0) \approx a$.[25] An agent following this rule does not exercise any cognitive effort and just copies her past savings. For example, if the agent has some positive wealth, then she simply keeps her amount of savings the same and consumes all her interest income together with her labor income. Such a passive approach to savings seems to be the most realistic candidate for a starting point of learning. It is also a rare example of a simple and natural decision rule that is correctly defined at any state and for any feasible liquidity constraint $\underline{a}$.[26]

Besides its behavioral interpretation, this initialization has the advantage of introducing from little to no bias. First, the optimal policy in the Aiyagari economy is quantitatively close to the diagonal line. Agents save slightly above their previous saving for small values of wealth and slightly below it for large values of wealth. Second, because of the chosen generational structure, NN agents start with the same distribution of wealth as RE agents. Agents who always copy their savings from the past would keep this distribution unchanged.

---

[22]There is empirical evidence that people pay attention to the experiences of their friends while making economic decisions. For example, Bailey et al. (2018) present evidence that this is the case for the decision to buy a house. Bailey et al. (2019) find that this also happens with the decision to buy a new smartphone.

[23]This is also very similar to the idea of learning from experiences generated by several parallel agents, which is exploited by asynchronous reinforcement learning algorithms. See, for example, Mnih et al. (2016).

[24]This is also similar to exploration – adding random noise to actions, which is implemented in many RL algorithms. However, the goal of exploration there is to obtain new information about the environment. The standard exploitation-exploration trade-off is not applicable in the context of this paper as agents know the environment already. However, to make learning more stable, they need to think about the states that are not from their own experience.

[25]This is achieved by setting $\pi(a, z|\theta) = a + \mu\phi(a, z|\theta)$, where $\phi(a, z|\theta)$ is the output of a neural network. Here $\mu = 0.01$ is a small number, and $\mathbb{E}_{\theta_0}\left[\phi(a, z|\theta_0)\right] = 0$ as weights θ_0 are initialized randomly with the standard procedure of PyTorch. Thus, at first all agents have the inertial decision rule with some little idiosyncratic noise. See Paszke et al. (2019) and PyTorch documentation for the details about the random initialization of weights.

[26]Saving a at a state (a, z) is possible when consumption $a(1 + r) + wz - a$ is positive. At the same time, to insure that consumption is always positive, it should be that $\underline{a}(1 + r) + wz_{min} - \underline{a} > 0$. Thus, the first inequality follows from $a \geqslant \underline{a}$. See Aiyagari (1994) for more details.

Thus, if we observe some difference between the wealth distribution of NN agents and the distribution of RE agents, this difference comes from something that NN agents learned and not something they started with.

Target network. For the next-period saving defined in equation 17, $\tilde{\theta}_t$ does not coincide with θ_t. Instead, it slowly tracks the policy network with the updating rule given by $\tilde{\theta}' = (1 - \lambda)\tilde{\theta} + \lambda\theta$, just as happens in Lillicrap et al. (2015). In reinforcement learning, the network with the second set of weights $\tilde{\theta}$ is usually called a target network. In practice, target networks are essential for stability of learning. The intuition behind the need for a target network here is similar to the one discussed in Mnih et al. (2015). Consider the following example. As can be seen from equation 14, a larger value of $\tilde{a}_{\tau+2} = \pi(a_{\tau+1}, z_{\tau+1})|\tilde{\theta})$ implies smaller $\tilde{c}_{\tau+1}$, and thus an update of θ toward larger savings. If $\tilde{\theta}_t$ is set to be the same as θ_t, then an over-saving bias in the policy leads to an update that is also biased toward over-saving. Intuitively, if an agent is certain that she must save a lot for two periods later, then the logic of consumption smoothing requires her to save a lot for tomorrow. Thus, a biased policy can become even more biased after the learning step. The agent might be trapped in an "over-saving chase." Cycles of this kind can lead to divergence or oscillations. A slowly updated target network dramatically mitigates this problem as the implied next-period savings become more stable.

One possible interpretation of this aspect of learning is that it is something like an evolutionary-developed mechanism (which all AI algorithms have to utilize too). Agents are reasonably cautious of the described "learning oscillations" and use a conservative value for the next-period saving in the learning rule given by equation 13. Another interpretation is that agents' cognitive costs make it too hard to simultaneously optimize the choice at some state and to consider the effects of the previous policy updates on the next-period saving. In the extreme case, agents would just optimize over $\pi(a_t, z_t|\theta)$, holding the next-period saving at the same level actually chosen in the past, that is, setting $\tilde{a}_{\tau+2} = a_{\tau+2} = \pi(a_{\tau+1}, z_{\tau+1}|\theta_{\tau+1})$. However, that would be extremely inefficient in terms of learning.[27] A target network can be thought of as adjusting for the fact that the policy was updated but with some delay.

Optimization. Instead of making simple gradient steps of a fixed size, the optimization from equation 13 is done with Adam. Adam is a gradient-based optimization algorithm introduced in Kingma and Ba (2014). It is designed for training neural networks and is

---

[27]The value of $\pi(a_{\tau+1}, z_{\tau+1}|\theta)$ that solves the intertemporal trade-off is close to the rational choice at $(a_{\tau+1}, z_{\tau+1})$ only if $\tilde{a}_{\tau+2}$ is close to the rational choice. In contrast, if the value of $\tilde{a}_{\tau+2}$ is a suboptimal saving choice, it biases the learning update. Thus, if $\tilde{a}_{\tau+2}$ is not updated and always equals $a_{\tau+2}$ chosen with some imperfect policy, then this experience biases learning until it is forgotten.

essentially the standard in modern deep learning. It facilitates stability and increases the speed of convergence.

3.5. Algorithm

Algorithm 1 shows all elements of the model together. In the pseudocode, T stands for the length of life, which is 100 for all agents in the model. The age when childhood ends is 20 for all agents. F stands for the learning frequency. The memory buffer R keeps the last several personal experiences of the agent and has a finite size M. The batch size is $N = |E|$, where E is the set of episodes used for a learning step. The share of external experiences is κ. The values of these hyperparameters are discussed in Section 4.2. The cap on savings a_{max} is 50, the minimum consumption c_{min} equals wz_{min} (see the discussion in Section 3.1).

4. Learning Outcomes

One of the broad conclusions of the paper is that learning a good saving policy dynamically from previous experiences is a hard task. There are several obstacles that make the process of learning intrinsically unstable. These learning issues seem to be inherent to learning as many of them universally arise in most of the modern applications of deep RL. Thus, these or similar issues might affect learning and outcomes of real-world economic agents. In the consumption-saving problem, they might lead to biases or learning traps in which agents save nothing or save excessively. All of that suggests that the assumption of rational expectations in macroeconomics is, in fact, a strong assumption.

4.1. Learning obstacles

The following are the problems that I find negatively affect learning done by neural network agents. Some of them can be mitigated by the modifications described in the previous section.

1. Learning a good policy function requires approximation of a point in an infinitely dimensional space. This implies that in general a good approximation can be achieved only through calibration of a large number of parameters. That in turn requires a lot of updating steps and a lot of experiences to learn from. Thus, a high learning frequency and/or a long learning time are required for successful learning.

2. The experiences from which agents learn are correlated over time. As agents typically do not accumulate or decrease savings sharply, levels of a_t usually are concentrated in

Algorithm 1 Algorithm of neural network agents.

Initialize policy network $\phi(a, z|\theta)$ and target network $\phi(a, z|\tilde{\theta})$

function $\pi(a, z|\theta)$

    $x = a + \mu\phi(a, z|\theta)$

    if $x < \underline{a}$ **then**

        $x = \underline{a}$

    end if

    if $x > a_{max}$ **then**

        $x = a_{max}$

    end if

    if $(1 + r)a + wz - x < c_{min}$ **then**

        $x = (1 + r)a + wz - c_{min}$

    end if

    return x

end function

Initialize memory buffer R of size M

Draw initial saving a_1 from the stationary distribution of wealth for RE agents

for $t = 1, \ldots, T$ **do**

    Draw z_t from Markov process Γ

    if $t \leqslant$ (end of childhood age) **then**

        Select action $a_{t+1} = \pi^{RE}(a_t, z_t)$

    else

        Select action $a_{t+1} = \pi(a_t, z_t|\theta)$

    end if

    Store episode (a_{t-1}, z_{t-1}, z_t) in R

    for step $= 1, \ldots, F$ **do**

        Reset learning buffer E

        Draw the number of external experiences k from the binomial distribution $B(N, \kappa)$

        Draw k artificial experiences (a, z, z'), put them into E

        Sample $N - k$ random experiences from R, put them into E

        For each episode $e = (a, z, z') \in E$, set

$$\tilde{c}_{e,1} = (1 + r)a + wz - \pi(a, z|\theta),$$

$$\tilde{c}_{e,2} = (1 + r)\pi(a, z|\theta) + wz' - \tilde{a},$$

        where $\tilde{a} = \pi\big(z', \pi(a, z|\theta)|\tilde{\theta}\big)$ is treated as a constant.

        Update weights θ by maximizing

$$\sum_{e \in E} \left[u(\tilde{c}_{e,1}) + \beta u(\tilde{c}_{e,2})\right]$$

        Update target network:
$$\tilde{\theta}' \leftarrow (1 - \lambda)\tilde{\theta} + \lambda\theta$$

    end for

end for

---

one region of the state space. An agent can mitigate this problem by using external experiences and randomizing experiences from her memory.

3. The experiences depend on the policies that agents learned. The wealth that an agent has at some period depends on the saving decisions she made before. If the current policy is degenerate (e.g., implies always saving nothing), then the sample to learn from also becomes degenerate (e.g., it consists only of experiences with zero wealth). Again, an agent can mitigate this problem by using external experiences.

4. The policy learned before an update affects the update. This opens room for self-perpetuating biases, like the over-saving chase described in Section 3.4. Target networks help to mitigate this problem.

5. Learning cannot happen on the boundaries of the action space. If $\pi(a, z|\theta)$ is too small or too large, it must be clipped to be a valid action. This can happen, for example, if an agent always wants to consume more than her cash-on-hand. The gradient with respect to θ is zero, and there is no update of the weights. If the policy $\pi(\cdot|\theta)$ is very pathological and this happens at every considered state, the agent is stuck without learning. Though this particular mechanism is somewhat specific to neural networks, it also reflects something about learning in general. Successful learning requires an informative feedback about the quality of the decisions made. When an agent has no chance to meaningfully experiment because the choices she makes are infeasible or too suboptimal, this might result in a learning trap. It is either very hard or impossible to learn if the current policy is degenerate.

6. Learning is slow for high values of consumption. In equation 14, marginal utilities $u'(\tilde{c}_\tau)$ and $u'(\tilde{c}_{\tau+1})$ and consequently the Euler equation error become small if $\tilde{c}_\tau$ and $\tilde{c}_{\tau+1}$ are both large. Then, even if the policy update happens in the right direction, it affects the policy very little. This happens if an agent has accumulated a lot of wealth and consumes from the high interest income she has. Intuitively, for wealthy high consumers, consumption mistakes are less noticeable and make them learn less. Though it makes perfect sense for agents to pay more attention to situations when consumption mistakes affect utility a lot, this mechanism can result in persistence of over-saving behavior.

Before I present the simulation results, I will discuss how these obstacles become relevant through the choice of the hyperparameters that make agents boundedly rational.

4.2. Margins of rationality

Algorithms based on neural networks almost unavoidably require a relatively large set of hyperparameters to define. Some of those have a natural interpretation; others are less interpretable. At the end, I use the following research strategy. I choose hyperparameters in a way that provides good learning results. Then, I choose some parameters of the model that I experimentally find important for performance and that are interpretable as "margins of rationality." I investigate the effect of making agents less rational on these margins.

Deviations from rationality might take different forms. Very generally, agents might have limited ability to collect, store, or process information. The first happens in the model, e.g., because an agent learns from a sample of correlated experiences. The second happens if she only can remember a few periods from her past. The third happens if the agent is limited in her ability to convert memories into a decision rule, e.g., because few learning steps are taken each period. As there are many reasons for irrationality, there are many hyperparameters that positively affect learning. I found the following ones to be important.

1. **Learning frequency**. The higher the number of learning steps per period, the faster agents develop policies that are close to the optimal saving rule.

2. **Share of external experiences**. The higher the share of random external experiences, the fewer the agents who converge to degenerate policies. As was discussed in the previous section, awareness about outcomes of others is important for successful learning.[28]

3. **Network architecture**. I find that if the neural architecture is too simple, then the model is not expressive enough and it is impossible to approximate well the optimal policy.[29]

4. **Memory size.** Memory size positively affects learning results, but the effect is moderate. This is because some information from the experiences that are already dropped from memory is still kept in the weights of the network.

Since all of these hyperparameters are sensible margins of (ir)rationality, I do not choose a single one. To investigate the role of bounded rationality and provide some related comparative statics, I consider two types of NN agents. The second type (*high rationality*) is supposed to strictly dominate the first type (*low rationality*). The parameters for the two

---

[28]At the same time, larger awareness might in principle be harmful as it restricts agents from focusing on the experiences that are more relevant for them. For example, an agent with low wealth might want to prioritize updating policy in the region of the state space with low savings.

[29]However, there is no monotone dependence, as the standard bias-variance trade-off logic applies. A neural network with a complicated architecture and many neurons introduces higher variance.

types are presented in Table 1. The default option for plots is the first configuration, as it provides stronger insights about the role of bounded rationality.

	Low rationality	High rationality
Neurons in hidden layers	(4,4)	(8,8)
Learning frequency (F)	50	200
Share of external experiences (κ)	25%	50%
Memory size (M)	50	100
Batch size (N)	10	
α	0.01	
λ	0.01	

Table 1: Configurations of NN agents.

The other hyperparameters are less interpretable as margins of rationality and are taken to be the same for both agents. The batch size is 10, which means that at each learning step, agents think about the 10 periods selected in E. This number seems to be intuitively plausible. At the same time, I find that batch size affects the results very little. The value of the learning rate is a common choice in applications of reinforcement learning to simple control problems (other common choices would be 10^{-3} and 10^{-4}). I find this learning rate to be a good choice for the trade-off between quality and speed of learning. A learning rate that is too high would result in the standard gradient descent problems like "overshooting." A learning rate that is too small would make learning very slow (that could be somewhat compensated by increasing the learning frequency, but a very high learning frequency does not seem plausible for human agents). The weight λ for updating the target network is also among the standard choices in the literature. A smaller value of λ would slow down learning, while a larger value would decrease stability. The current choice implies that NN agents with low rationality almost entirely forget their policy in 3-5 periods (for example, $0.99^{5 \times 50} \approx 0.08$), which seems to be realistic. The main logic behind these choices is that they can provide a good approximation to rationality. I keep them fixed and focus on the effect of making agents less rational on the dimensions with clear interpretation.
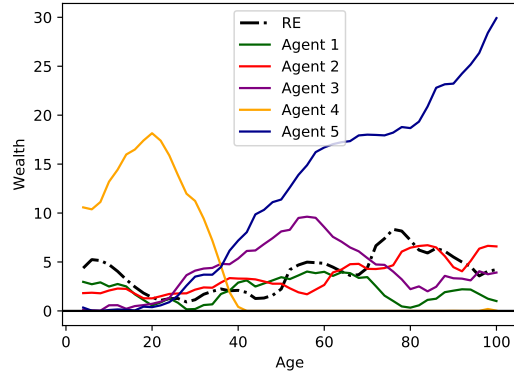
4.3. Examples of learning

I simulate a large number[30] of neural network agents over their lifetime. Figure 2 illustrates the model and some of the common learning outcomes.
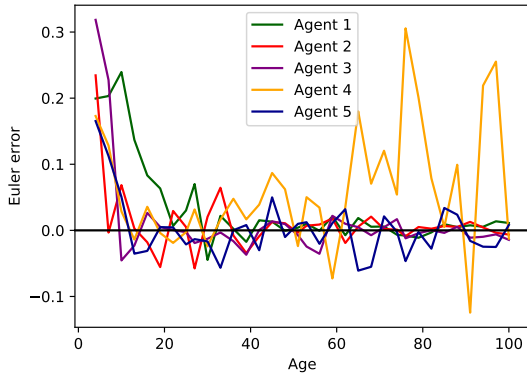
---

[30]There are between 10,000 and 30,000 NN agents in each simulation. One simulation requires between 5 and 20 hours when performed on a laptop.
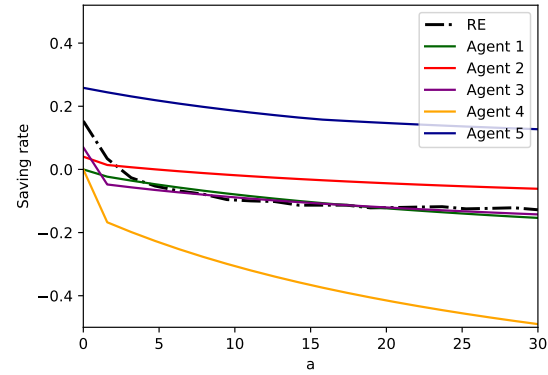
(a) Diversion from RE wealth trajectories. Shaded area has width equal to $2\times$(standard deviation of diversion).

(b) Wealth trajectories, example.

(c) Histories of learning errors, example. Euler error is defined as $\frac{\beta(1+r)u'(c_{t+1})}{u'(c_t)} - 1$.

(d) Policies in the last period, example. Saving rate is defined as $\frac{a_{t+1}-a_t}{ra_t+wz_t}$, z fixed at the mean level.

Fig. 2. Illustration of learning outcomes.

For Figure 2(a), I simulate alternative wealth trajectories for all NN agents in the economy. Those are the trajectories they would have under the same productivity shocks but if they followed the RE policy instead of the policies they learned. I calculate the difference between these trajectories and the actual trajectories. Figure 2(a) shows how the mean and variance of diversion evolve over time. For the first 20 years, decisions are made by rational parents, and there is no diversion. Since childhood is over, agents start to deviate from the RE paths. The dispersion is growing over time, as diversions accumulate. The mean diversion is positive, as NN agents tend to have more savings on average than RE agents.

Figure 2(b) presents examples of realized wealth paths for five NN agents and one RE agent. There are two sources of variation in outcomes. The first one is standard for hetero-

geneous agents models: the agents had different productivity histories z_t over their lifetimes. In addition to that, for NN agents the difference in outcomes also comes from the difference in learned decision rules. Decision rules are different because agents learned from different experiences and because of the randomness in learning.[31]
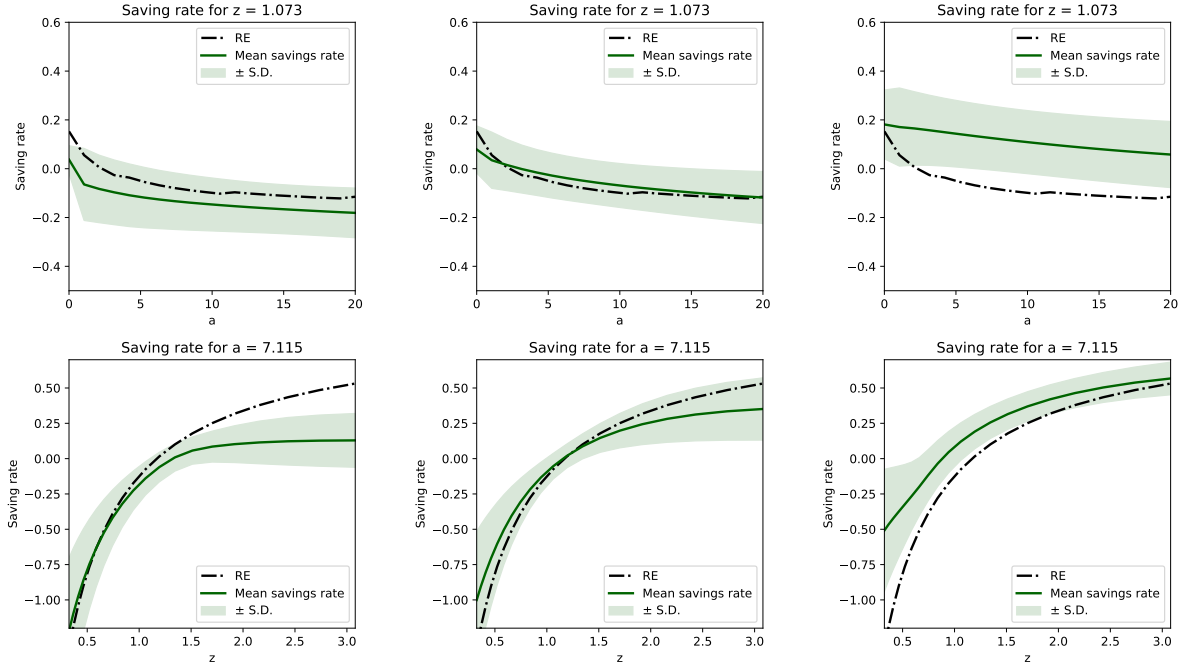
The trajectories of agents 1, 2, and 3 seem to have a nature similar to that of the presented RE trajectory. In contrast, agents 4 and 5 are in a *learning trap*. For the last half of her life, agent 4 has no savings at all. The trap of agent 5 is very different in terms of her consumption-saving behavior. This agent steadily increases the amount of her savings well above levels that rational agents are likely to reach.

The process of learning can be illustrated by the average Euler equation error the same agents got for the experiences in E when they were updating their policies. As can be seen in Figure 2(c), all agents have relatively large errors at the first steps of their lives. However, agents 1, 2, and 3 successfully update their policies, and over time their Euler errors start to fluctuate around 0. Agent 4, however, is in a save-nothing learning trap. Her policy tells her to save below the liquidity constraint, her policy gradient is zero, she is not able to update her policy, and she is constantly making consumption errors. Agent 5, despite being in the "trap" of saving too much, still has small Euler errors as her consumption is large. Because her Euler error is small, her learning updates are small, which prevents her from quickly fixing her consumption behavior.

Finally, the difference between the agents can also be understood by looking at their policy functions.[32] Agents 1, 2, and 3 learned policies that are close to the optimal one. In particular, they correctly learned to save less when their wealth is larger. These agents might save from 5 to 15 percentage points less or more than is optimal, which intuitively seems to be a plausible scale of irrationality. In contrast, agent 4 felt into a learning trap. She has a policy that implies negative saving rates for all values of a. Her neural network has output (at least for wealth 0, where she is) that is below the liquidity constraint. She can neither implement this choice nor learn from it, as she always has a zero policy gradient. She is permanently stuck with no savings. Finally, agent 5 learned an insensible policy of constantly increasing her savings.

---

[31]Neural network agents have different random initializations of their networks, different random batch draws at any learning step, and different random external experiences.

[32]To illustrate policies, I plot the saving rate of agents. This is for practical reasons, as it is hard to see any difference if plain policy values are to be plotted (see Figure 11 for an example).

(a) Agents with small wealth.   (b) Agents with medium wealth.   (c) Agents with large wealth.

Fig. 3. Average policies by wealth.

Saving rate is defined as $\frac{a_{t+1}-a_t}{ra_t+wz_t}$. When a is varied, z is fixed at the average level. When z is varied, a is fixed at the average level. Width of shaded area equals $2\times$(standard deviation of policies).

4.4. Policies and wealth

To provide a more general illustration of how policies drive wealth in the economy, I split agents into three groups. One group includes all agents with last-period wealth below the 5th percentile of wealth in the RE economy. Another group includes all agents with last-period wealth above the 95th percentile of wealth in the RE economy. The last group includes everyone else. Figure 3 presents the average of the policies for each of the groups.

As Figure 3(a) shows, agents with low wealth have policies that are biased in the direction of under-saving. Many of them, just like agent 4 from Section 4.3, have a policy that implies a negative saving rate everywhere, and are stuck with zero savings. In contrast, agents from the medium-wealth group have policies that are approximately unbiased on average, as can be seen in Figure 3(b). Finally, agents with high wealth significantly over-save, as can be seen in Figure 3(c). As will be discussed in Section 4.5, their biased policies drive their high wealth.
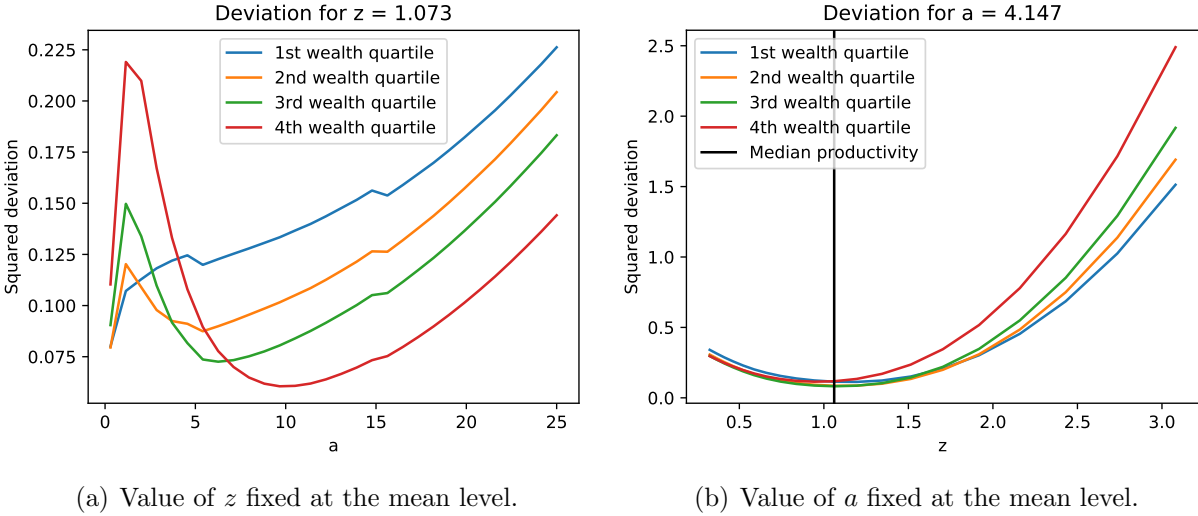
27

(a) Value of z fixed at the mean level. (b) Value of a fixed at the mean level.

Fig. 4. Mean squared difference from RE policy.

4.5.   Familiarity with the state space

The previous exercise shows that there is a statistical association between being rich and saving too much or having low wealth and saving too little. Because there is feedback between policies and experiences, the direction of causality can be unclear. For example, does wealth make the wealthy make some particular over-saving mistakes, or did they become wealthy because of the tendency to over-save? To separate the effect of experiences, I consider the policies that agents have just after their childhood is over. Because before that agents only learned from the experiences generated by their rational parents, all the difference in wealth is independent of their policies and is essentially exogenous. In Figure 4, I plot the average squared difference[33] from the RE policy as a function of the state.

I split agents into four groups by their wealth quartiles. As can be seen from Figure 4(a), for high values of a the average difference from rationality is larger for agents with smaller wealth. Those agents learned how to act in the region of low savings, and their policies are such that they would make a lot of mistakes if they suddenly became rich. However, the same happens with the wealthiest agents in the region of low wealth, as they would make bad choices there. Because for small values of a_t the range of possible choices of a_{t+1} is small, mechanically all squared differences are small too. However, for some agents these mistakes might have important consequences. As will be discussed in Section 5.5, I find that descendants of parents with low wealth are more likely to accumulate wealth excessively.

---

[33]Squared difference from the RE policy is equal, up to a multiplier, to the second-order approximation to the utility loss from choosing a suboptimal action at a particular state.

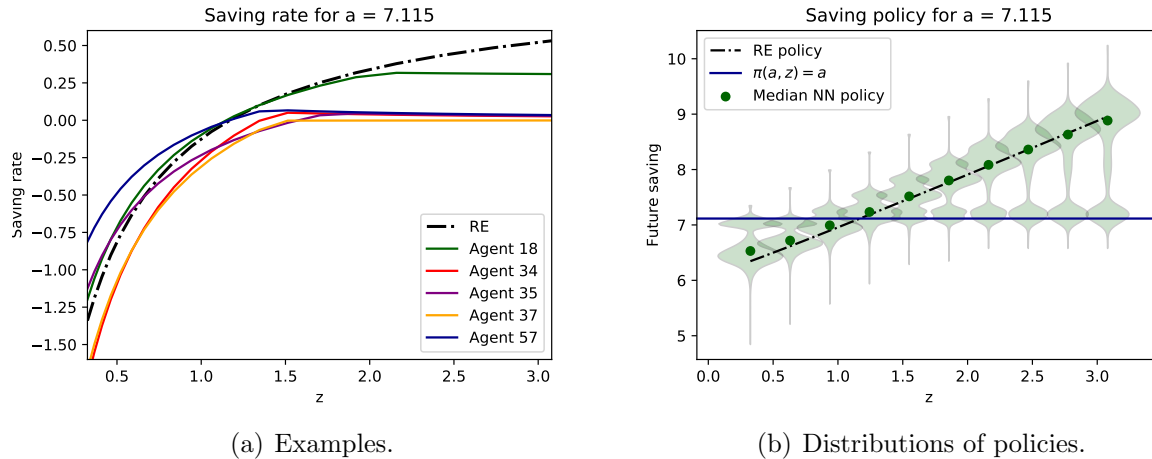(a) Examples.   (b) Distributions of policies.

Fig. 5. Failing to learn reaction to shocks.

Descendants of parents with high wealth, on the contrary, are more likely to have no savings in the last period.

Similarly, Figure 4(b) shows a noticeable U-shaped pattern of mistakes that is centered roughly around the median value of productivity z. Agents make larger mistakes at the states of productivity that are rarely realized. That happens because those states are largely unknown to the agents. I conclude that for neural network agents having experience from a particular region of the state space helps to make better choices there.

4.6.   Reaction to productivity shocks

Another class of learning failures I find is inattention to productivity shocks. Some agents fail to learn to react either to all changes in productivity or to productivity shocks with high and rare values. In the model, this is equivalent to passively consuming all labor income after some income threshold. This generates excess sensitivity of consumption in the economy, which will be discussed in Section 5.4.

Figure 5(a) presents several examples of how policies might look in the case when agents do not react to high shocks only. In Figure 5(b), I plot the distributions of the policies for several values of z. As can be seen, median policies are quite close to the optimal decision rule, but there is always a mode around the inertial policy that corresponds to the agents who ignore shocks. Thus, agents who are inattentive to shocks account for a significant part of the under-savings biases that can be seen in the bottom row of Figure 3. In particular, the bias is strongest in Figure 3(a), where agents with low wealth under-save a lot on average for high values of z. For middle-wealth agents, inattention occurs less frequently, and it is essentially absent for agents with high wealth (this can also be seen if I reproduce Figure 5(b)

29

for different wealth groups).

Intuitively, some agents might fail to start thinking about the volatility of shocks. They are stuck around the inertial initialization or update their policies but only in terms of the dependence on the other state variable a. Also, high levels of productivity are too infrequent, and many agents do not have a proper sample to learn in that region of the state space. Though this behavior might depend on the choice of initialization (which assumes that agents completely ignore shocks) and the updating framework, it is also both plausible and quite intuitive that some agents would fail to learn that productivity levels from the Markov process Γ are correlated or can be used to smooth consumption. This finding resembles the behavior of agents in Gabaix (2016). There, agents choose to pay partial attention or no attention at all to some state variables.

Though it is not modeled here formally, this phenomenon might become especially important in times of high volatility. A considerable fraction of agents in the model react to large shocks incorrectly and in the same way. This can have a significant effect on the economy if there is a structural change that makes large shocks more likely. In addition to that, an inability to learn any reaction to shocks might have macroeconomic consequences in a model with a richer structure, say, with monetary and fiscal shocks. If many agents fail to account for the shocks, this can affect the persistency of GDP cycles and the efficiency of monetary or fiscal policy.

4.7. Cost of irrationality

Making suboptimal saving choices leads to smaller expected lifetime utility. How much does an agent who is born irrational lose ex-ante? I find the compensating variation CV as the solution for

$$\mathbb{E}\left[\sum_{t=20}^{100} \beta^t u\big(c_{i,t}^{RE}\big)\right] = \mathbb{E}\left[\sum_{t=20}^{100} \beta^t u\big((1+CV)c_{i,t}^{NN}\big)\right],$$

where $c_{i,t}^{RE}$ and $c_{i,t}^{NN}$ stand for consumption of RE and NN agents respectively. The compensating variation is the fraction that should be added to the consumption of NN agents to make their expected adulthood utility the same as that of RE agents.[34] Similarly, I find the equivalent variation EV as the solution for

$$\mathbb{E}\left[\sum_{t=20}^{100} \beta^t u\big((1-EV)c_{i,t}^{RE}\big)\right] = \mathbb{E}\left[\sum_{t=20}^{100} \beta^t u\big(c_{i,t}^{NN}\big)\right].$$

---

[34]For simplicity, I ignore the value of consumption after the last period (though agents take it into consideration). It should not significantly affect the results as β^{100} is a very small number.
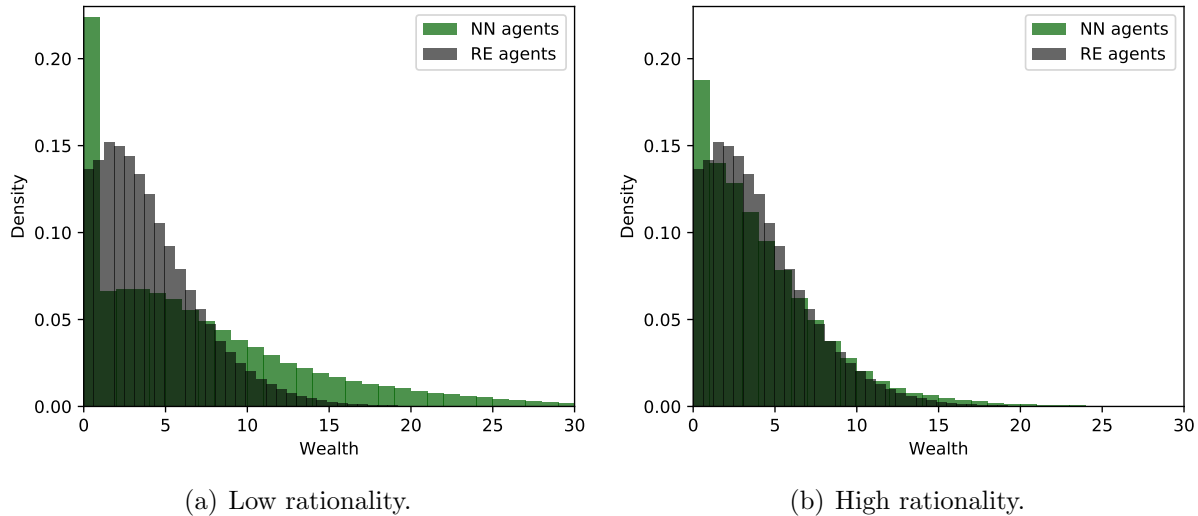
(a) Low rationality.   (b) High rationality.

Fig. 6. Distribution of savings in the economy.

This is the fraction that RE agents would be ready to give up ex-ante to keep being rational.

Table 2 presents the results for the two types of NN agents. Not surprisingly, having low rationality is harmful for utility. The cost of that is more than one-tenth of consumption. Having high rationality leads to a smaller disadvantage relative to full rationality, with the loss roughly equivalent to 4% of consumption.

	Low rationality	High rationality
CV	13.6%	4.3%
EV	12.0%	4.2%

Table 2: Equivalent and compensating variations for NN agents vs. RE agents.

5. Macroeconomic Implications

Figure 6 shows the simulated distributions of savings for rational agents and the two types of neural network agents. As can be seen in Figure 6(b), under high rationality the results closely approximate the outcomes under rational expectations. However, as is clear from Figure 6(a), learning and bounded rationality might have important effects on the macroeconomy.

5.1. Hand-to-mouth agents

Models with liquidity constraints typically do not generate a mass of agents around the minimum wealth. This happens because rational agents save precautionarily in order not to find themselves in the position where they would not be able to borrow to smooth their consumption. However, in the real world many households have essentially no savings. Following Zeldes (1989), I say that an agent has hand-to-mouth status if her wealth is smaller than her labor income over two months.[35] In the rational expectations equilibrium, I find that only 2.6% of agents are hand-to-mouth. In contrast, Aguiar et al. (2020) report that 22.7% of the households surveyed for the Panel Study of Income Dynamics (PSID) are hand-to-mouth according to this definition. Similar to Ilut and Valchev (2020), I test whether policy learning can explain the discrepancy between the data and the model with rational agents.

As can be seen from Figure 6, neural network agents have a tendency to accumulate around the lower bound for assets. The first row of Table 3 presents the fraction of hand-to-mouth agents in the simulation. Both types of NN agents are substantially more likely to become hand-to-mouth than agents who comply with rational expectations. As before, the configuration with low rationality delivers larger differences. In the model, hand-to-mouth NN agents most often are in a save-nothing learning trap. This suggests that behavioral patterns related to trap-like learning outcomes might contribute to the large share of hand-to-mouth agents observed empirically.

	PSID	Low rationality	High rationality	RE
$\mathbb{P}(\text{H2M}_t)$	0.227	0.157	0.072	0.026
$\mathbb{P}(\text{H2M}_{t+2}\vert \text{H2M}_t)$	0.648	0.897	0.686	0.511
$\mathbb{P}(\text{H2M}_{t+4}\vert \text{H2M}_t)$	0.584	0.845	0.526	0.321

Table 3: Frequency and persistence of hand-to-mouth status. Estimates for PSID are from Aguiar et al. (2020).

I also calculate the probability that agents stay hand-to-mouth after two periods and four periods in the RE and NN economies. The last two rows of Table 3 present the results. For the same reason that hand-to-mouth status is rare in RE models, it is also not particularly persistent. After hitting the liquidity constraint, agents usually quickly accumulate some amount of assets. In the data, however, the status is more persistent, and this is also true in the NN economy. The hand-to-mouth neural network agents mostly are the ones in a learning trap, which often means that agents learned to save nothing and do not update

---

[35]Formally, agent i has hand-to-mouth status at period t if $a_{i,t} < \frac{wz_{i,t}}{6}$.

policies anymore. Hand-to-mouth status is thus extremely persistent (in fact, the model implies larger persistence than observed in the data).

5.2. Wealth inequality

It is hard to fit the empirically observed wealth distribution with a typical heterogeneous agents model, in particular, because we observe thick upper tales in the data (see Benhabib and Bisin, 2018, for a review). As is clear from Figure 6, neural network agents end up with a wealth distribution more dispersed than that of rational agents. Table 4 shows that NN agents indeed have various measures of inequality closer to those obtained from the PSID and reported in Krueger et al. (2016) than the rational expectations equilibrium.

	PSID	Low rationality	High rationality	RE
Gini	0.77	0.535	0.486	0.411
Top 1% share	30.9	5.8	5.2	3.9
Top 5% share	53.7	20.1	18.2	15.2
Top 20% share	82.7	53.5	48.6	44.2

Table 4: Wealth inequality. Estimates for PSID are from Krueger et al. (2016). Shares of wealth are in percentage points.

Broadly, the main reason for higher inequality in an economy of neural networks is that the heterogeneity in thinking adds up to the heterogeneity from idiosyncratic shocks. There are two particular extreme cases of how this happens. The first one is the higher share of agents stuck with no savings. Those are the agents who at some point learned a policy that keeps them with no wealth and that does not allow for improving learning. The second one is the high number of agents who save excessively and end up in the upper tail of the distribution. Thus, the upper tail consists of agents who learned a policy that is biased toward high saving. Those agents might have small updates because of high consumption or upward-biased policy updates, or because they are not able to learn further at all.

Though it seems unlikely that in the real world all of the wealthiest (or all of the least well-off) are the ones who just failed to learn a good decision rule, this finding is still suggestive. I consistently find agents diverging to save-nothing or save-everything policies for different set-ups and different hyperparameters. The main explanations for this are the self-perpetuating biases and learning traps mentioned above. These mechanisms might also be present in the real world, generating extreme learning outcomes and contributing to inequality in the economy. The model predicts a moderate but non-negligible effect of learning on inequality.

5.3. Marginal propensity to consume

The baseline rational expectations model fails to deliver an average marginal propensity to consume close to the empirical estimates. The latter tend to vary from 20% to 60% (see a summary of the empirical literature in Carroll et al., 2017). In contrast, the RE equilibrium has an average MPC around 7.7%. I define the marginal propensity to consume out of transitory income as

$$MPC(a, z) = 1 - \frac{1}{1+r}\frac{\partial}{\partial a}\pi(a, z),$$

which for a rational agent is the same as the share consumed out of a marginal increase in cash-on-hand coming from a transitory source.

Data	Low rationality	High rationality	RE
$[0.2, 0.6]$	0.17	0.11	0.08

Table 5: Marginal propensity to consume. The range of empirical estimates is from Carroll et al. (2017).
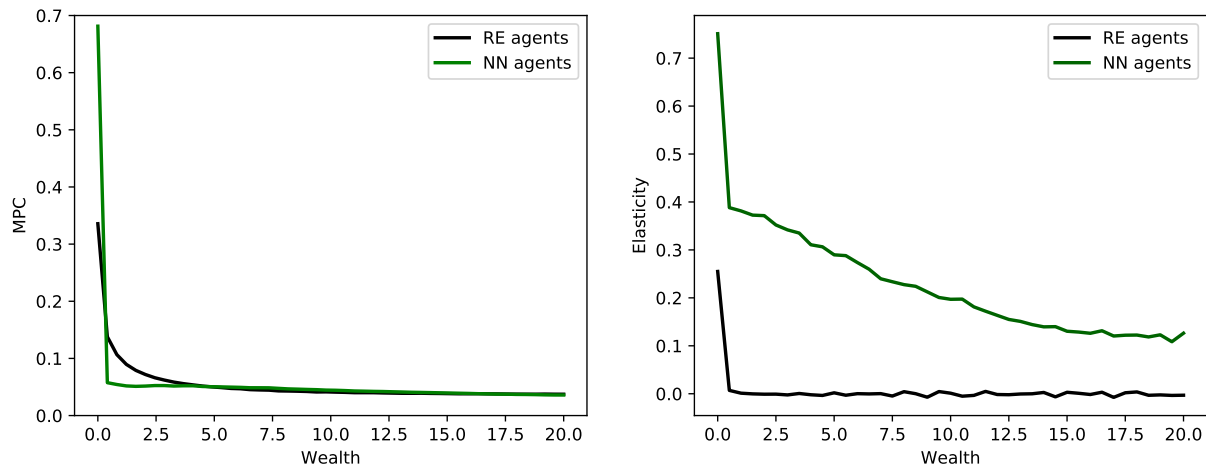
I calculate the average marginal propensity to consume in the simulated data. Table 5 shows that the economy with low rationality has an average MPC roughly twice that of the Aiyagari economy with rational agents. Further, I split the generated data into brackets by current asset holdings. Figure 7(a) presents the dependence of the average MPC on wealth. As can be seen, the difference mainly comes from the agents near the liquidity constraint. The NN economy generates a larger mass of hand-to-mouth agents. Those agents have a higher MPC even in the rational expectations model (Carroll and Kimball, 1996), but in the NN economy those often are the agents in a save-nothing learning trap. The MPC of those agents is extremely high (and usually is equal to 1).

5.4. Sensitivity of consumption

Another puzzle against which NN agents can be tested is the ability of expected changes in income to predict future changes in consumption. With perfect capital markets and under the permanent income hypothesis, predictable changes should not matter. One way to test this is to consider the regression

$$\Delta \log c_{i,t+1} = \alpha + \beta \log y_{i,t} + \epsilon_{i,t.}$$

If $\log y_{i,t}$ follows an AR(1) process with autocorrelation ρ, then $\mathbb{E}_t \Delta \log y_{i,t+1} = -(1 - \rho)\log y_{i,t}$, and $-\frac{\beta}{1-\rho}$ is the elasticity of changes in consumption with respect to predictable

(a) Marginal propensity to consume in simulation, by wealth.

(b) Elasticity of changes in consumption $\Delta \log c_{i,t+1}$ with respect to predictable changes in income $\mathbb{E}_t \Delta \log y_{i,t+1}$, by wealth.

Fig. 7. Consumption behavior of NN agents.

changes in income. This approach was first implemented by Flavin (1981), who found evidence against the PIH in the data. After that, some studies also found a strong rejection (see Zeldes, 1989; Lusardi, 1996), while others found limited or no evidence of that (see, for example, Blundell et al., 2008; Hsieh, 2003).

I calculate the elasticity in the simulated data. Table 6 shows that the NN economy has a much larger elasticity than the Aiyagari economy with rational agents. Further, I split the generated data into brackets by current asset holdings. Figure 7(b) presents the dependence of elasticity on wealth. As can be expected, for rational agents, deviations from the PIH happen only around the liquidity constraint. In contrast, NN agents exhibit excess sensitivity of consumption for all values of wealth.

Data	Low rationality	High rationality	RE
[0,0.4]	0.339	0.063	0.026

Table 6: Elasticity of changes in consumption $\Delta \log c_{i,t+1}$ with respect to predictable changes in income $\mathbb{E}_t \Delta \log y_{i,t+1}$. Upper bound for empirical estimates is from Lusardi (1996).

Divergence from rational expectations is likely to make agents fail to perfectly extract the predictable component from labor income fluctuations. One particular mechanism behind the high sensitivity in the economy was described in Section 4.6, where I show that some agents do not learn to react to productivity shocks at all, and some ignore increases in productivity

after certain thresholds. Those agents can change consumption one-to-one after a change in income. As was discussed, they mostly have low wealth. This is consistent with Figure 7(b), which shows that richer NN agents have a smaller elasticity of consumption.

5.5. Intergenerational mobility

In this part, I investigate the effect of learning and bounded rationality on intergenerational mobility. There are many channels through which parental outcomes are transmitted to children's outcomes: transfers, bequests, education, ability, health, social capital, norms, behavioral patterns, and so on (see Black and Devereux, 2010, for a review). Among those channels, two are related to the framework considered in this project. Waldkirch et al. (2004) found a correlation between the consumption patterns of parents and children in the PSID. Charles and Hurst (2003) use the PSID to investigate the relationship between the child's wealth and parental wealth. They found that a part of the relationship can be explained by parents and children having similar savings propensities, preferences for risk, and investment behavior. All of that suggests that learning related to the consumption-saving problem might play some role in the intergenerational wealth dynamics.

One reason why mobility might be different for neural network agents is bounded rationality. It affects the lifetime wealth trajectories and might contribute to convergence or further divergence of agents who started from unequal wealth levels. The direction of this effect is ambiguous. For example, if agents just make random consumption mistakes, it would increase mobility as it would introduce noise that makes starting conditions less important. Alternatively, if all agents follow policies close to the inertial rule $\pi(a, z) = a$, then they would preserve their inherited wealth and there would be no wealth mobility at all. Another reason for different mobility comes from learning. In the first periods of life, learning depends not on the experiences generated by an agent but on the experiences generated by her parent. Thus, learning provides an additional channel through which parental outcomes determine the outcomes of the child.

I find that learning can either contribute to higher mobility or suppress it. First, let us see the case when learning serves as an intergenerational equalizer. A particularly interesting mechanism for that is the following. In the model, descendants of parents with smaller savings more often accumulate extreme levels of wealth (defined as being above the 99th wealth percentile in the RE economy); see Figure 8(a). Similarly, descendants of parents with higher savings more often are hand-to-mouth in the last period, as can be seen in Figure 8(b). All of that, of course, is not the case for rational agents.[36] This is likely to

---

[36]Here and below, for RE agents I consider the wealth of the same agent but the corresponding number

(a) Probability of having extreme wealth.   (b) Probability of being hand-to-mouth.
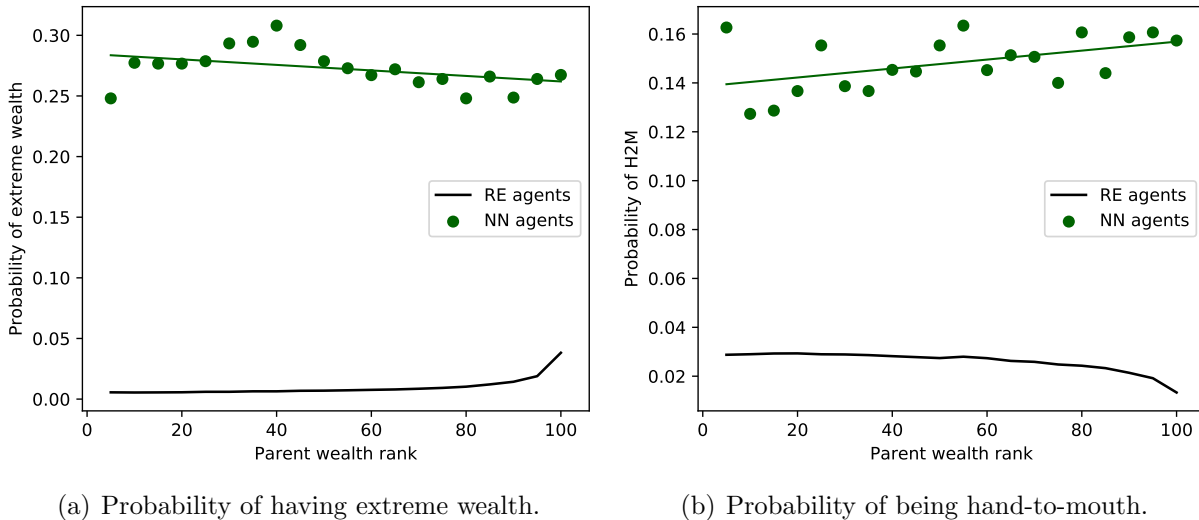
Fig. 8. Probability of extreme outcomes vs. parental wealth rank.

Extreme wealth is defined as being above the 99th wealth percentile in the RE economy in the last period. Parents are RE agents.

happen because of the learning pattern observed in Section 4.5. Those with low-wealth (high-wealth) parents are less aware of the right behavior in the region of high (low) savings, and are more likely to make mistakes if a sequence of shocks brings them there. In addition to this mechanism, consumption mistakes coming from imperfect learning generally might increase mobility as they can work as a noise that distorts the role of initial wealth. All in all, parental wealth has a slightly negative effect on the wealth of neural network agents at the end of their lives. This can be seen in Figure 9(a), where I present rank-rank plots in the fashion of Chetty et al. (2014) for both NN and RE agents.

However, this discussion misses another potential channel of how learning might affect mobility. If a parent was not rational as in the model but instead an agent stuck in an under- or over-saving learning trap, then this trap is likely to substantially affect learning by the descendant, and, consequently, her economic outcomes. To see whether this is the case, I simulate the second generation of neural network agents. Those agents learn from the last 20 experiences of their neural network parents before becoming adults, and then inherit their wealth. Figure 9(b) shows that this indeed makes wealth much more persistent across generations. In particular, it can be seen that wealthier agents account for most of the effect. A descendant of a wealthy parent might keep the over-saving behavior, e.g., because learning is slow for high values of consumption. Then, she would increase the inherited

of periods later.

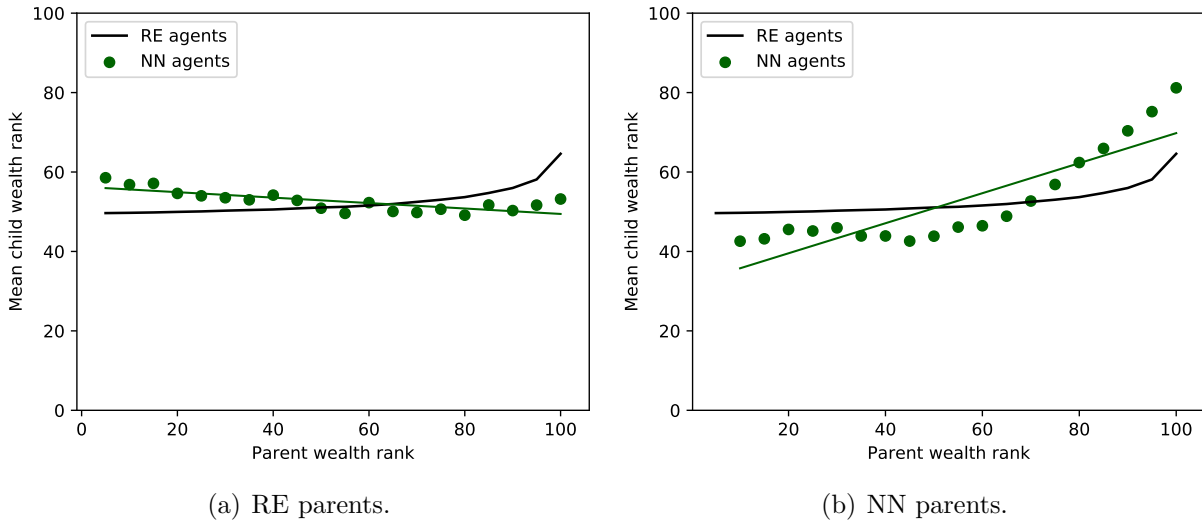(a) RE parents.

(b) NN parents.

Fig. 9. Association between children's and parents' percentile wealth rank. The outcomes are in the last period.

extreme wealth even further instead of dissaving it (the latter is what a rational agent would do). I find that this mechanism generates a lot of immobility in the economy. This is also consistent with the findings of Charles and Hurst (2003), who find large wealth to be more persistent across generations.

To provide some connection with the data, I compare the results of the simulation with the estimates of Chetty et al. (2014) and Charles and Hurst (2003). In Chetty et al. (2014), the child is approximately 30 years old, and thus I use income after 30 periods for comparison. In Charles and Hurst (2003), the average age of children is 37.5 years, and thus I use wealth after 38 periods. However, the comparison with these empirical estimates is only suggestive. Charles and Hurst (2003) consider the wealth of the child before bequests, and in Chetty et al. (2014) bequests are also unlikely because of the young age. The Aiyagari framework does not allow for a good counterpart of their estimates, as it essentially assumes bequests.[37] At the same time, all channels of intergenerational immobility other than wealth are not represented well in the model. For example, there is no transmission of human capital as productivity shocks have a relatively transitory effect on income. Thus, a direct comparison of estimates from the model and the papers might be impossible.

Table 7 presents the results. For the baseline NN economy, the mobility is much lower, just as Figure 9(b) suggests. The moments generated by the model better fit the correlation between parental and child incomes than the RE baseline. Most of the difference in average

[37]That said, bequests in the real world might have limited importance for most agents except for the richest ones. It has also been documented that the latter tend to start transferring wealth earlier.

incomes comes from the difference in interest income. Accordingly, the model also predicts a smaller mobility of wealth. The predictions are too small compared to the estimates of Charles and Hurst (2003) because of the large role of bequests in the model.

		Data	Low rationality	High rationality	RE
Income					
	Rank-rank slope	0.341	0.370	0.132	0.125
	Intergenerational elasticity	0.45	0.35	0.13	0.13
Wealth					
	Shorrocks index	0.88	0.56	0.89	0.81
	Intergenerational elasticity	0.37	0.49	0.32	0.46

Table 7: Intergenerational mobility for NN agents with NN parents.

Income is defined as $ra_t + wz_t$. For income, the gap between parents and children is 10 years. The empirical estimates are from Chetty et al. (2014). For wealth, the gap between parents and children is 18 years. The empirical estimates are from Charles and Hurst (2003). The Shorrocks index is defined as $\frac{n - \sum_{i=1}^{n} a_{i,i}}{n-1}$, where $(a_{i,j})$ is the transition matrix between quintiles. A smaller index indicates lower mobility. To find intergenerational elasticity of wealth, I exclude agents with no savings.

Interestingly, for agents with high rationality I observe mobility even slightly higher than that in the RE economy. The mechanisms discussed above, e.g., the role of mistakes as an equalizing noise, outweigh the channel of over-saving by the children of extremely wealthy agents (who are almost absent among NN agents with high rationality; see the distribution in Figure 6(b)), and learning again becomes an intergenerational equalizer.

6.   Conclusion

Neural networks provide a new approach to model behavioral agents in macroeconomics. Its credibility comes from the fact that neural networks can successfully perform a large variety of human-like tasks. I focus on a particular advantage of neural networks: they can be used to model dynamic learning of a decision rule. This is a kind of learning that is not traditionally considered in economics, but it might be important as agents face it often in the real world. For the consumption-saving problem, the approach reveals a connection between the wealth experience and the quality of choices made. It also reveals learning traps in which agents become permanently poor (in terms of savings) or suboptimally rich. That might explain several patterns observed in the wealth distribution. This model can be scaled to multiple assets and more general economic control problems, potentially revealing some learning patterns and traps there too.

References

Adusumilli, K., Geiecke, F., Schilter, C., 2019. Dynamically optimal treatment allocation using reinforcement learning. arXiv:1904.01047.

Aguiar, M. A., Bils, M., Boar, C., 2020. Who are the hand-to-mouth? Working Paper 26643, National Bureau of Economic Research.

Aiyagari, S. R., 1994. Uninsured idiosyncratic risk and aggregate saving. The Quarterly Journal of Economics 109, 659–684.

Arifovic, J., 1996. The behavior of the exchange rate in the genetic algorithm and experimental economies. Journal of Political Economy 104, 510–541.

Athey, S., Wager, S., 2021. Policy learning with observational data. Econometrica 89, 133–161.

Azinovic, M., Gaegauf, L., Scheidegger, S., 2019. Deep equilibrium nets. Available at SSRN 3393482 .

Ba, L. J., Caruana, R., 2013. Do deep nets really need to be deep? arXiv:1312.6184.

Bach, F., 2017. Breaking the curse of dimensionality with convex neural networks. The Journal of Machine Learning Research 18, 629–681.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., 2018. The economic effects of social networks: Evidence from the housing market. Journal of Political Economy 126, 2224–2276.

Bailey, M., Johnston, D. M., Kuchler, T., Stroebel, J., Wong, A., 2019. Peer effects in product adoption. Working Paper 25843, National Bureau of Economic Research.

Barron, A. R., 1993. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory 39, 930–945.

Benhabib, J., Bisin, A., 2018. Skewed wealth distributions: Theory and empirics. Journal of Economic Literature 56, 1261–91.

Black, S. E., Devereux, P. J., 2010. Recent developments in intergenerational mobility. Working Paper 15889, National Bureau of Economic Research.

Blundell, R., Pistaferri, L., Preston, I., 2008. Consumption inequality and partial insurance. American Economic Review 98, 1887–1921.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., Kurth-Nelson, Z., 2020. Deep reinforcement learning and its neuroscientific implications. Neuron 107, 603–616.

Bullard, J., Duffy, J., 1998. A model of learning and emulation with artificial adaptive agents. Journal of Economic Dynamics and Control 22, 179–207.

Carroll, C., Slacalek, J., Tokuoka, K., White, M. N., 2017. The distribution of wealth and the marginal propensity to consume. Quantitative Economics 8, 977–1020.

Carroll, C. D., Kimball, M. S., 1996. On the concavity of the consumption function. Econometrica pp. 981–992.

Charles, K. K., Hurst, E., 2003. The correlation of wealth across generations. Journal of Political Economy 111, 1155–1182.

Chen, M., Joseph, A., Kumhof, M., Pan, X., Shi, R., Zhou, X., 2021. Deep reinforcement learning in a monetary model. arXiv:2104.09368.

Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N., 2014. Is the United States still a land of opportunity? Recent trends in intergenerational mobility. American Economic Review 104, 141–47.

Choi, J. J., Laibson, D., Madrian, B. C., Metrick, A., 2009. Reinforcement learning and savings behavior. The Journal of Finance 64, 2515–2534.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, 303–314.

Duarte, V., 2018. Machine learning for continuous-time economics. Available at SSRN 3012602 .

Farhi, E., Werning, I., 2019. Monetary policy, bounded rationality, and incomplete markets. American Economic Review 109, 3887–3928.

Fernández-Villaverde, J., Nuno, G., Sorg-Langhans, G., Vogler, M., 2020. Solving high-dimensional dynamic programming problems using deep learning. Working paper.

Fernández-Villaverde, J., Hurtado, S., Nuño, G., 2019. Financial frictions and the wealth distribution. Working Paper 26302, National Bureau of Economic Research.

Flavin, M. A., 1981. The adjustment of consumption to changing expectations about future income. Journal of Political Economy 89, 974–1009.

Gabaix, X., 2016. Behavioral macroeconomics via sparse dynamic programming. Working Paper 21848, National Bureau of Economic Research.

García-Schmidt, M., Woodford, M., 2019. Are low interest rates deflationary? A paradox of perfect-foresight analysis. American Economic Review 109, 86–120.

Goldfayn-Frank, O., Wohlfart, J., 2020. Expectation formation in a new environment: Evidence from the German reunification. Journal of Monetary Economics 115, 301–320.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., Seung, H. S., 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405, 947–951.

Hertwig, R., Barron, G., Weber, E. U., Erev, I., 2004. Decisions from experience and the effect of rare events in risky choice. Psychological Science 15, 534–539.

Holland, J. H., Miller, J. H., 1991. Artificial adaptive agents in economic theory. The American Economic Review 81, 365–370.

Hsieh, C.-T., 2003. Do consumers react to anticipated income changes? Evidence from the Alaska permanent fund. American Economic Review 93, 397–405.

Ilut, C., Valchev, R., Vincent, N., 2020. Paralyzed by fear: Rigid and discrete pricing under demand uncertainty. Econometrica 88, 1899–1938.

Ilut, C. L., Valchev, R., 2020. Economic agents as imperfect problem solvers. Working Paper 27820, National Bureau of Economic Research.

Kahou, M. E., Fernández-Villaverde, J., Perla, J., Sood, A., 2021. Exploiting symmetry in high-dimensional dynamic programming. Working Paper 28981, National Bureau of Economic Research.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Krueger, D., Mitman, K., Perri, F., 2016. Macroeconomics and household heterogeneity. In: *Handbook of Macroeconomics*, Elsevier, vol. 2, pp. 843–921.

Kuchler, T., Zafar, B., 2019. Personal experiences and expectations about aggregate outcomes. The Journal of Finance 74, 2491–2542.

Lettau, M., Uhlig, H., 1999. Rules of thumb versus dynamic programming. American Economic Review 89, 148–174.

Lian, C., 2020. Mistakes in future consumption, high MPCs now. Working paper, MIT.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv:1509.02971.

Lusardi, A., 1996. Permanent income, current income, and consumption: Evidence from two panel data sets. Journal of Business & Economic Statistics 14, 81–90.

Mackowiak, B., Wiederholt, M., 2009. Optimal sticky prices under rational inattention. American Economic Review 99, 769–803.

Maliar, L., Maliar, S., Winant, P., 2019. Will artificial intelligence replace computational economists any time soon? Discussion Paper DP14024, CEPR.

Malmendier, U., Nagel, S., 2011. Depression babies: Do macroeconomic experiences affect risk taking? The Quarterly Journal of Economics 126, 373–416.

Malmendier, U., Nagel, S., 2016. Learning from inflation experiences. The Quarterly Journal of Economics 131, 53–87.

Mankiw, N. G., Reis, R., 2002. Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips curve. The Quarterly Journal of Economics 117, 1295–1328.

Marimon, R., McGrattan, E., Sargent, T. J., 1990. Money as a medium of exchange in an economy with artificially intelligent agents. Journal of Economic Dynamics and Control 14, 329–373.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning. In: *International Conference on Machine Learning*, PMLR, pp. 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

Nagel, S., Xu, Z., 2019. Asset pricing with fading memory. Working Paper 26255, National Bureau of Economic Research.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch:

An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 8024–8035.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Shi, R. A., 2021. Learning to make consumption-saving decisions in a changing environment: an AI approach. arXiv:2105.10099.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M., 2014. Deterministic policy gradient algorithms. In: *International conference on machine learning*, PMLR, pp. 387–395.

Sims, C. A., 2003. Implications of rational inattention. Journal of Monetary Economics 50, 665–690.

Stevens, L., 2020. Coarse pricing policies. The Review of Economic Studies 87, 420–453.

Sutton, R. S., Barto, A. G., 2018. Reinforcement learning: An introduction. MIT Press.

Tauchen, G., 1986. Finite state Markov-chain approximations to univariate and vector autoregressions. Economics Letters 20, 177–181.

Valchev, R., Ilut, C., 2017. Economic agents as imperfect problem solvers. In: *2017 Meeting Papers*, Society for Economic Dynamics, no. 1285.

Waldkirch, A., Ng, S., Cox, D., 2004. Intergenerational linkages in consumption behavior. Journal of Human Resources 39, 355–381.

Zeldes, S. P., 1989. Consumption and liquidity constraints: an empirical investigation. Journal of Political Economy 97, 305–346.

Zhang, K., Yang, Z., Başar, T., 2019. Multi-agent reinforcement learning: A selective overview of theories and algorithms. arXiv:1911.10635.

Zhao, C., Ke, S., Wang, Z., Hsieh, S.-L., 2020. Behavioral neural networks. Available at SSRN 3633548 .

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., Socher, R., 2020. The AI economist: Improving equality and productivity with AI-driven tax policies. arXiv:2004.13332.

A.   Neural Networks

This section is supposed to briefly introduce the notion of a neural network but cannot serve as a complete introduction to the subject. An (artificial) neural network is a result of stacking several layers of neurons. Each neuron has an output, which is given by applying an activation function f to a linear combination of the outputs from the neurons in the previous layer (for the first hidden layer, a linear combination of the original inputs). Figure 10 provides an example of a neural network relevant for this paper. This network has a two-dimensional input and a one-dimensional output. It has two hidden layers with $n_1 = n_2 = 4$ neurons in each of them. Formally, the system is given by

$$h_j^{(1)} = f\big(\theta_{j,0}^{(1)} + \theta_{j,1}^{(1)} a_t + \theta_{j,2}^{(1)} z_t\big),$$

$$h_j^{(2)} = f\Big(\theta_{j,0}^{(2)} + \sum_{i=1}^{n_1} \theta_{j,i}^{(2)} h_i^{(1)}\Big),$$

$$a_{t+1} = \theta_0^{(3)} + \sum_{i=1}^{n_2} \theta_i^{(3)} h_i^{(2)}.$$

An activation function f is needed to make the system non-linear, otherwise, the output would be simply a linear function of the original inputs. These days, the standard choice for f is the ReLU activation function, which is given by $f(x) = \max\{0, x\}$. It also has a biological motivation, see Hahnloser et al. (2000).
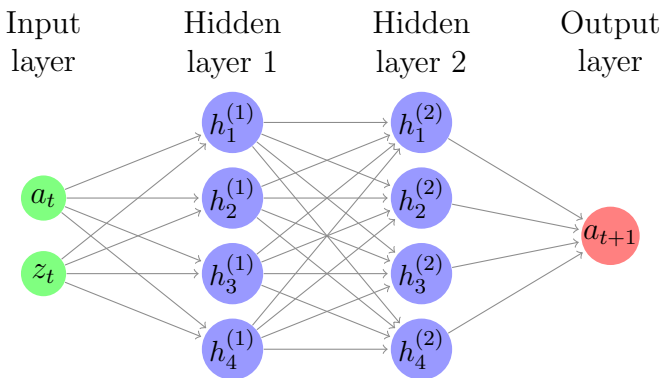
Fig. 10. An example of a neural network with two hidden layers.

There are many reasons why neural networks became extremely successful in practical applications. Some of these reasons are yet to be understood from the theoretical viewpoint, but some are already established. Rumelhart et al. (1986) introduced back-propagation, the procedure for efficient differentiation with respect to the weights of a network. Cybenko

(1989), Barron (1993), Bach (2017), and others emphasize the good approximation properties. The modern strand of the literature, including Ba and Caruana (2013), emphasizes the better ability of deep neural networks to *learn* while having expressivity similar to that of networks with just one hidden layer.

There is a feature of neural networks that is related to the described learning process. Neural networks are models with many parameters and high expressivity. In terms of the bias-variance trade-off, this means that they usually have low bias but a very high variance. For example, a typical neural network used for a classification task could perfectly fit the data set it is trained on. To avoid overfitting, training is stopped after some number of gradient descent steps. Similarly, neural networks behind saving policy functions are updated incrementally. The same approach is implemented in all modern deep reinforcement learning algorithms.

B. Asymptotic Rationality

In the following, I provide evidence that the suggested model of learning is asymptotically rational. Though no formal statement about convergence is provided, I show that agents with hyperparameters that reflect few cognitive limitations can learn the optimal policy with a very high degree of precision. This suggests that the learning described in this paper is indeed directed toward the optimal policy, and the class of neural network agents nests very good approximations to rational agents. If this property was not satisfied, that would cast doubt on the credibility of the model, as too many updating schemes satisfy the other properties listed on page 13.

For this experiment, I use a more expressive network architecture with 4 hidden layers that have 4, 8, 8, and 4 neurons respectively. I also use a smaller learning rate $\alpha = 0.001$. I simulate several agents who learn from external experiences only (values of a_t are distributed uniformly over the savings dimension of the state space) and do that at a high frequency, making 10,000 learning steps per period. Figure 11 illustrates policies of the first 5 agents at the end of their lives. As can be seen, the degree of precision is very high even when policies are plotted as saving rates, let alone plain values of savings.
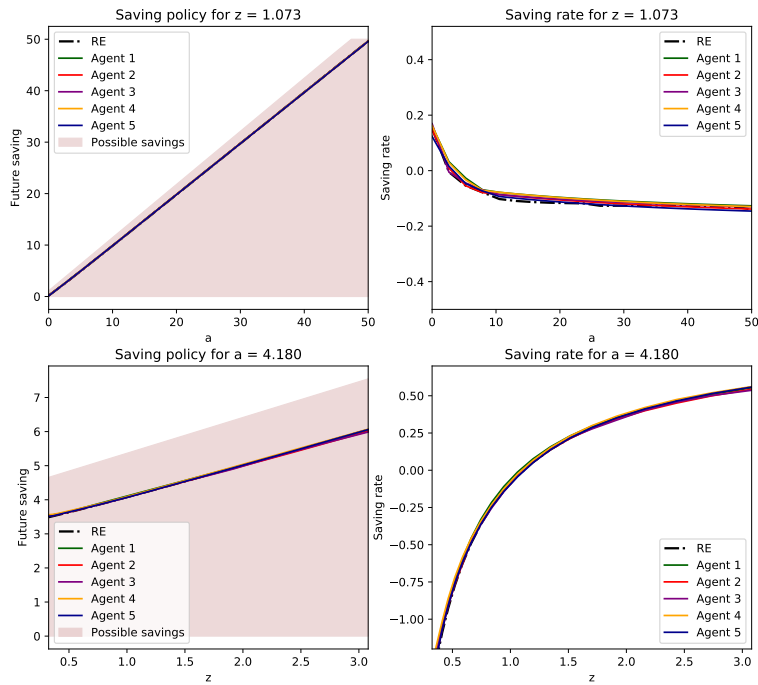
Fig. 11. Examples of successful learning.

In the first row, z is fixed at the mean level. In the second row, a is fixed at the mean level for the RE economy. Saving rate is defined as $\frac{a_{t+1}-a_t}{ra_t+wz_t}$.