Decentralized Reinforcement Learning: Global Decision-Making via Local Economic Transactions

Michael Chang 1  Sidhant Kaushik 1  S. Matthew Weinberg 2  Thomas L. Griffiths 2  Sergey Levine 1

Abstract

This paper seeks to establish a framework for directing a society of simple, specialized, self-interested agents to solve what traditionally are posed as monolithic single-agent sequential decision problems. What makes it challenging to use a decentralized approach to collectively optimize a central objective is the difficulty in characterizing the equilibrium strategy profile of non-cooperative games. To overcome this challenge, we design a mechanism for defining the learning environment of each agent for which we know that the optimal solution for the global objective coincides with a Nash equilibrium strategy profile of the agents optimizing their own local objectives. The society functions as an economy of agents that learn the credit assignment process itself by buying and selling to each other the right to operate on the environment state. We derive a class of decentralized reinforcement learning algorithms that are broadly applicable not only to standard reinforcement learning but also for selecting options in semi-MDPs and dynamically composing computation graphs. Lastly, we demonstrate the potential advantages of a society’s inherent modular structure for more efficient transfer learning.

1. Introduction

Biological processes, corporations, and ecosystems – physically decentralized, yet in some sense functionally unified. A corporation, for example, optimizes for maximizing profits as it were a single rational agent. But this agent abstraction is an illusion: the corporation is simply a collection of human agents, each solving their own optimization problems, most not even knowing the existence of many of their colleagues. But the human as the decision-making agent is also simply an abstraction of the trillions of cells making their own simpler decisions. The society of agents is itself an agent. What mechanisms bridge between these two levels of abstraction, and under what framework can we develop learning algorithms for studying the self-organizing nature of intelligent societies that pervade so much of the world?

Both the monolithic and the multi-agent optimization frameworks in machine learning offer a language for representing only one of the levels of abstraction but not the relation between both. The monolithic framework, the most commonly used in much of modern machine learning, considers a single agent that optimizes a single objective in an environment, whether it be minimizing classification loss or maximizing return. The multi-agent framework considers multiple agents that each optimize their own independent objective and each constitute each other’s learning environments. What distinguishes the multi-agent from the monolithic is the presence of multiple independent optimization problems. The difficulty of interpreting a learner in the monolithic framework as a society of more simpler components is that all components are still globally coupled together by the same optimization problem without independent local optimization problems themselves, as are the weights in a neural network trained by backpropagation. The difficulty of interpreting a multi-agent system under a global optimization problem is the computational difficulty of computing Nash equilibrium (Daskalakis et al., 2009), even for general two-player games (Chen et al., 2009).

To better understand the relationship between the society and the agent, this paper makes four contributions, each of which exist at a different level of abstraction. At the highest level, we define the societal decision-making framework to relate the local optimization problem of the agent to the

You know that everything you think and do is thought and done by you. But what’s a “you”? What kinds of smaller entities cooperate inside your mind to do your work?

(Minsky, 1988)

1Department of Computer Science, University of California, Berkeley, USA. 2Department of Computer Science, Princeton University, USA. Correspondence to: Michael Chang <mbchang@berkeley.edu>.

At the fourth level, we empirically investigate various economic transactions among many individual agents that has gone broke. Kwee et al. (2001) added an agent's bid should be equivalent to the optimal Q-value for executing that agent's transformation and developed the Hayek architecture for introducing new agents and removing agents that have gone broke. Kwee et al. (2001) added external memory to the Hayek architecture.

However, to this date there has been no proof to the best of our knowledge that the bid-updating schemes proposed in these works simultaneously optimize a global objective of the society. Sutton (1988) provides a convergence proof for temporal difference methods that share some properties with the bucket brigade credit assignment scheme, but importantly does not take the competition between the individual agents into account. But it is precisely the competition among the agents in these societal learning methods that make their equilibria nontrivial to characterize (Mazumdar et al., 2019). Our work offers an alternative auction mechanism for which we prove that the optimal solution for the global objective does coincide with a Nash equilibrium of the society. We provide a novel objective for a temporally local credit assignment scheme for optimizing the MDP objective of the society as an emergent consequence of the agents’ optimizing their own auction utilities. These algorithms treat the auction utility as optimization objectives themselves, thereby learning a societal policy that is global in space and time using only credit assignment for learnable parameters that is local in space and time. At the fourth level, we empirically investigate various implementations of the clone Vickrey society under our decentralized reinforcement learning algorithm and find that a particular set of design choices, which we call the credit conserving Vickrey implementation, yields both the best performance at the societal and and agent level.

Finally, we demonstrate that the societal decision making framework, along with its solution, the algorithm that learns the solution, and the implementation of this algorithm, is a broadly applicable perspective on self-organization to not only standard reinforcement learning but also for selecting options in semi-MDPs (Sutton et al., 1999) and for composing functions in dynamic computation graphs. Moreover, we show evidence that the local credit assignment mechanisms of societal decision-making framework produce more efficient learning than the global credit assignment mechanisms of the monolithic framework.

2. Related Work

Describing an intelligent system as the product of interactions among many individual agents dates as far back as the Republic (Plato, 380 B.C.), in which Plato analyzes the human mind via an analogy to a political state. This theme continued into the early foundations of AI in the 1980s and 1990s through cognitive models such as the Society of Mind (Minsky, 1988) and Braitenberg vehicles (Braitenberg, 1986) and engineering successes in robotics (Brooks, 1991) and in visual pattern recognition (Selfridge, 1988).

The closest works to ours were the algorithms developed around that same time period that sought as we do to leverage a multi-agent society for achieving a global objective, starting as early as the bucket brigade algorithm (Holland, 1985), in which agents bid in a first-price auction to operate on the state and auction winners directly paid their bid to the winners from the previous step. Prototypical self-referential learning mechanisms (Schmidhuber, 1987) improved the bucket brigade by imposing credit conservation in the economic transactions. The neural bucket brigade (Schmidhuber, 1989) adapted the bucket brigade to learning neural network weights, where payoffs corresponded to weight changes. Baum (1996) observed that the optimal choice for an agent’s bid should be equivalent to the optimal Q-value for executing that agent’s transformation and developed the Hayek architecture for introducing new agents and removing agents that have gone broke. Kwee et al. (2001) added external memory to the Hayek architecture.

However, to this date there has been no proof to the best of our knowledge that the bid-updating schemes proposed in these works simultaneously optimize a global objective of the society. Sutton (1988) provides a convergence proof for temporal difference methods that share some properties with the bucket brigade credit assignment scheme, but importantly does not take the competition between the individual agents into account. But it is precisely the competition among the agents in these societal learning methods that make their equilibria nontrivial to characterize (Mazumdar et al., 2019). Our work offers an alternative auction mechanism for which we prove that the optimal solution for the global objective does coincide with a Nash equilibrium of the society. We provide a novel objective for a temporally local credit assignment scheme by propagating Q-values through time, which distinguishes this work from recent decentralized learning methods (Goyal et al., 2019; Peng et al., 2019; Ohsawa et al., 2018; Pathak et al., 2019).

3. Preliminaries

To set up a framework for societal decision-making, we relate Markov decision processes (MDP) and auctions under a unifying language. We define an environment as a tuple that specifies an input space, an output space, and additional parameters for specifying an objective. An agent is a function that maps the input space to the output space. An objective is a functional that maps the learner to a real number. Given an environment and objective, the problem the agent solves is to maximize the value of the objective.
In the MDP environment, the input space is the state space $S$ and the output space is the action space $A$. The agent is a policy $\pi: S \rightarrow A$. The transition function $T: S \times A \rightarrow S$, the reward function $r: S \times A \rightarrow \mathbb{R}$, and discount factor $\gamma$ are additional parameters that specify the objective: the return $J(\pi) = \mathbb{E}_{s_0 \sim \rho^p(\cdot)} \left[ \sum_{t=0}^T \gamma^t r(s_t, a_t) \right]$, where $p^\pi(\cdot) = \mathbb{P}(s_0) \prod_{t=0}^T \mathbb{P}(a_t|s_t) \prod_{t=1}^T T(s_{t+1}|s_t, a_t)$. The agent solves the problem of finding $\pi^* = \arg \max_{\pi} J(\pi)$. For any state $s$, the optimal action for maximizing $J(\pi)$ is $\pi^*(s) = \arg \max_a Q^*(s,a)$, where the optimal Q function $Q^*(s,a)$ is recursively defined as $Q^*(s,a) = \mathbb{E}_{s' \sim T(s,a)} [r(s,a) + \gamma \max_{a'} Q^*(s',a')|s,a]$. 

In the auction environments we consider, the input space is a single auction item $s$ and the output space is the bidding space $B$. Instead of a single agent, each of $N$ agents $\psi^{1:N}$ compete to bid for the auction item via its bidding policy $\psi^i: \{s\} \rightarrow B$. Let $b$ be the vector of bids produced by $\psi^i$. The vector $\nu_i$ of each agent’s valuations for auction item $s$ and the auction mechanism – allocation rule $X: B^N \rightarrow \{0,1\}^N$ and pricing rule $P: B^N \rightarrow \mathbb{R}_{\geq 0}^N$ – are additional parameters that specify each agent’s objective: the utility $U_i^\psi(\psi^{1:N}) = \nu_i^T \cdot X_i^\psi(b) - P_i(b)$, where $X_i^\psi(b)$ is the proportion of $s$ allocated to $i$, and $P_i(b)$ is the scalar price $i$ pays. Each agent $i$ independently solves the problem of finding $\psi_i^* = \arg \max_{\psi_i} U_i^\psi(\psi^{1:N})$. The independent optimization of objectives distinguishes a multi-agent problem from a single-agent one and makes multi-agent problems generally difficult to analyze when an agent’s optimal policy depends on the strategies of other agents.

However, if an auction is dominant strategy incentive compatible (DSIC), bidding one’s own valuation is optimal, independent of other players’ bidding strategies. That is, truthful bidding is the unique dominant strategy. Notably, the Vickrey auction (Vickrey, 1961), which sets $P_i(b)$ to be the second highest bid $\max_j, j \neq i b_j$ and $X_i^\psi(b) = 1$ if $i$ wins and $0$ and $0$ respectively if $i$ loses, is DSIC, which means the dominant strategy equilibrium occurs when every agent bids truthfully, making the Vickrey auction straightforward to analyze. Another attractive property of the Vickrey auction is that the dominant strategy equilibrium automatically maximizes the social welfare $\sum_{i=1}^N v_i \cdot X_i^\psi(b)$ (Roughgarden, 2016), which selects the bidder with the highest valuation as winner. The existence of dominant strategies in the Vickrey auction removes the need for agents to recursively model others, giving the Vickrey auction the practical benefit of running in linear time (Roughgarden, 2016).

4. Societal Decision-Making

The perspective of this paper is that a society of agents can be abstracted as an agent that itself solves an optimization problem at a global level as an emergent consequence of the optimization problems its constituent agents solve at the local level. To make this abstraction precise, we now introduce the societal decision-making framework for analyzing and developing algorithms that relate the global decision problem of a society to the local decision problems of its constituent agents. We use primitive and society to distinguish between the agents at the local and global levels, respectively, which we define in the context of their local and global environments and objectives:

Definition 4.1. A primitive $\omega$ is a tuple $(\psi, \phi_T)$ of a bidding policy $\psi: S \rightarrow B$ and transformation $\phi_T: S \rightarrow S$.

Definition 4.2. A society $\Omega$ is a set of primitives $\omega^{1:N}$.

The global environment is an MDP that we call the global MDP, with state space $S$ and discrete action space $A = \{1, \ldots, N\}$ that indexes the primitives $\omega^{1:N}$. The local environment is an auction that we call the local auction with auction item $s \in S$ and bidding space $B = [0, \infty)$.

The connection between the local and global environments is as follows. Each state in the global MDP is an auction item for a different local auction. The winning primitive $\omega$ of the auction at state $s$ transforms $s$ into the next state $s'$ of the global MDP using its transformation $\phi_T$, parameterized by the global MDP’s transition function $T$. For each primitive $i$ at each state $s$, its local objective is the utility $U_i^\psi(\psi^{1:N})$. Its local problem is to maximize $U_i^\psi(\psi^{1:N})$. The global objective is the return $J(\pi_\Omega)$ in the global MDP of the global policy $\pi_\Omega$. The global problem for the society is to maximize $J(\pi_\Omega)$. We define the optimal societal Q function $Q_\Omega^*(s,\omega)$ as the expected return received from $\omega$ invoking its transformation $\phi_T$ on $s$ and the society activating primitives optimally with respect to $J(\pi_\Omega)$ afterward.

Since all decisions made at the societal level are an emergent consequence of decisions made at the primitive level, the societal decision-making framework is a self-organization perspective on a broad range of sequential decision problems. If each transformation $\phi_T$ specifies a literal action, then societal decision-making is a decentralized re-framing of standard reinforcement learning (RL). Societal decision-making also encompasses the decision problem of choosing $\phi_T$s as options in semi-MDPs (Sutton et al., 1999) as well as choosing $\phi_T$s as functions in a computation graph (Chang et al., 2018; Rosenbaum et al., 2017; Alet et al., 2018).

We are interested in auction mechanisms and learning algorithms for optimizing the global objective as an emergent consequence of optimizing the local objectives. Translating problems from one level of abstraction to another in this way would enable the design of decentralized learning algorithms that are inherently modular. It would also provide a recipe for engineering a multi-agent system to achieve a desired global outcome and permit theoretical expectations on the nature of the equilibrium of the society. To this end,
we present in the next section an auction mechanism for which the dominant strategy equilibrium of the primitives coincides with the optimal policy of the society, which we develop into a class of decentralized RL algorithms in the later sections.

5. Mechanism Design for the Society

We first observe that to produce the optimal global policy, the optimal bidding strategy for each primitive at each local auction must be to bid their societal Q-value. By defining each primitive’s valuation of a state as its optimal societal Q-value at that state, we show that the Vickrey auction ensures the dominant strategy equilibrium profile of the primitives coincides with the optimal global policy. Then we show that a market economy perspective on societal decision-making overcomes the need to assume knowledge of optimal Q-values, although weakens the dominant strategy equilibrium to a Nash equilibrium. Lastly, we explain that adding redundant primitives to the society mitigates market bubbles by enforcing credit conservation. Proofs are in the Appendix.

5.1. Optimal Bidding

We state what was observed informally in (Baum, 1996):

**Proposition 5.1.** Assume at each state $s$ the local auction allocates $X^t(b) = 1$ if $i$ wins and $X^t(b) = 0$ if $i$ loses. Then all primitives $\omega^i$ bidding their optimal societal Q-values $Q^i_{\Omega}(s, \omega^i)$ collectively induce an optimal global policy.

This proposition makes the problem of self-organization concrete: getting the optimal behavior in the global MDP to emerge from the optimal behavior in the local auctions can be reduced to incentivizing the primitives to bid their optimal societal Q-value at every state.

5.2. Dominant Strategies for Optimal Bidding

To incentivize the primitives to bid optimally, we propose to define the primitives’ valuations $v^i_s$ for each state $s$ as their optimal societal Q-values $Q^i_{\Omega}(s, \omega^i)$ and use the Vickrey auction mechanism for each local auction.

**Theorem 5.2.** If the valuations $v^i_s$ for each state $s$ are the optimal societal Q-values $Q^i_{\Omega}(s, \omega^i)$, then the society’s optimal global policy coincides with the primitives’ unique dominant strategy equilibrium under the Vickrey mechanism.

Then, the utility $U^i_s(\omega^{1:N})$ at each state $s$ that induces the optimal global policy, which we refer equivalently as $U^i_s(\omega^{1:N})$ for the winning primitive $\omega^i$ and $U^j_s(\omega^{1:N})$ for losing primitives $\omega^j$, is given by

$$U^i_s(\omega^{1:N}) = Q^i_{\Omega}(s, \omega^i) - \max_{j \neq i} b^k_s$$

and by $U^j_s(\omega^{1:N}) = 0$ for losing primitives.

5.3. Economic Transactions for Propagating $Q^*_{\Omega}$

We have so far defined optimal bidding with respect to societal decision-making and characterized the utilities as functions of $Q^i_{\Omega}$ for which such bidding is a dominant strategy. We now propose to redefine the utilities without knowledge of $Q^i_{\Omega}$ by viewing the society as a market economy.

Monolithic frameworks for solving MDPs, such as directly optimizing the policy $J(\pi)$ with policy gradient methods, are analogous to command-economies, in which all production – the transformation of past states $s_t$ into future states $s_{t+1}$ – and wealth distribution – the credit assignment of reward signals to parameters – derive directly from single central authority – the MDP objective. In contrast, under the societal decision-making framework, the optimal global policy does not derive directly from the MDP objective, but rather emerges implicitly as the equilibrium of the primitives optimizing their own local objectives. We thus redefine the valuations $v^i_t$ following the analogy of a market economy, in which production and wealth distribution are governed by the economic transactions between the primitives.

Specifically, we couple the local auctions at consecutive time-steps in the same game by defining the valuation $v^i_t$ of primitive $\omega^i$ for winning the auction item $s_t$ as the revenue it can receive in the auction at the next time-step by selling the product $s_{t+1}$ of executing its transformation $\omega^i_{t+1}$ on $s_t$. This compensation comes as the environment reward plus the (discounted) winning bid at the next time-step:

$$U^i_t(s_t) = r(s_t, \omega^i_t) + \gamma \max_{k} b^{k}_{s_{t+1}} - \max_{j \neq i} b^j_s,$$ \hspace{1cm} (2)

Analogous to a market economy, the revenue $\omega^i$ receives for producing $s_{t+1}$ from $s_t$ depends on the price the winning primitive $\omega^k$ at $t+1$ is willing to bid for $s_{t+1}$. In turn, $\omega^k$ sells $s_{t+2}$ to the winning primitive at $t+2$, and so on. Ultimately currency is grounded in the reward. Wealth is distributed based on what future primitives decide to bid for the fruits of the labor of information processing carried out by past primitives transforming one state to another.

**Definition 5.1.** A Market MDP is a global MDP in which all utilities are defined as in Equation 2.

As valuations now depend on the strategies of future primitives, the dominant strategy equilibrium from Theorem 5.2 must be weakened to a Nash equilibrium in the general case:

**Proposition 5.3.** In a Market MDP, it is a Nash equilibrium for every primitive to bid $Q^i_{\Omega}(s, \omega^i)$. Moreover, if the Market MDP is finite horizon, then bidding $Q^i_{\Omega}(s, \omega^i)$ is the unique Nash equilibrium that survives iterated deletion of weakly dominated strategies.
Decentralized Reinforcement Learning: Global Decision-Making via Local Economic Transactions

5.4. Redundancy for Credit Conservation

In general, credit is not conserved in the Market MDP: the winning primitive at time \( t = 1 \) gets paid an amount equal to the highest bid at time \( t \), but the winner at time \( t \) only pays an amount equal to the second highest bid at time \( t \). At time \( t \), \( \omega^i \) could bid arbitrarily higher than \( \max_j \neq i b^j \) without penalty, which distorts the valuations for primitives that bid before time \( t \), creating a market bubble.

To prevent market bubbles, we propose a modification to the society, which we will call a cloned society, for enforcing credit conservation – for all transformations \( \phi_T \) initialize at least two primitives that share the same \( \phi_T \):

**Lemma 5.4.** For a cloned society, at the Nash equilibrium specified in Proposition 5.3, what the winning primitive \( \hat{\omega}^i \) at time \( t \) receives from the winning primitive \( \hat{\omega}^k \) at \( t + 1 \) is exactly what \( \omega^k \) pays: \( b^k_{st+1} \).

We now state our main result.

**Theorem 5.5.** Define a cloned Vickrey society as a cloned society that solves a Market MDP. Then it is a Nash equilibrium for every primitive in the cloned Vickrey society to bid \( Q^*_t(s, \omega^i) \). In addition, the price that the winning primitive pays for winning is equivalent to what it bid.

The significance of Theorem 5.5 is that guaranteeing truthful bidding of societal Q-values decouples the analysis of the local problem within each time-step from that of the global problem across time-steps, which opens the possibility for designing learners to reach a Nash equilibrium that we know exists. Without such a separation of these two levels of abstraction the entire society must be analyzed as a repeated game through time – a non-trivial challenge.

6. From Equilibria to Learning Objectives

So far the discussion has centered around the quality of the equilibria, which assumes the primitives know their own valuations. We now propose a class of decentralized RL algorithms for learning optimal bidding behavior without assuming knowledge of the primitives’ valuations.

Instead of learning to optimize the MDP objective directly, we propose to train the primitives’ parameterized bidding policies to optimize their utilities in Equation 2, yielding a class of decentralized RL algorithms for optimizing the global RL objective that is agnostic to the choice of RL algorithm used to train each primitive. By Theorem 5.5, truthful bidding for all agents is one global optimum to all agent’s local learning problems that also serves as a global optimum for the society as whole. In the special case where the transformation \( \phi_T \) is a literal action, this class of decentralized RL algorithms can serve as an alternative to any standard algorithm for solving discrete-action MDPs. An on-policy learning procedure is described in Algorithm 1.

![Image](image_url)

**Figure 1.** The cloned Vickrey society. In this market economy of primitive agents, wealth is distributed not directly from the global MDP objective but based on what future primitives decide to bid for the fruits of the labor of information processing carried out by past primitives transforming one state to another. The primitive \( \omega^i \) that wins the auction at time \( t \) receives an environment reward \( r(s_t, \omega^i) \) as well as payment \( b^i_{st+1} \) from \( \omega^i_{t+1} \) for transforming \( s_t \) to \( s_{t+1} \). By the Vickrey auction, the price \( \omega^j_{t+1} \) pays to transform \( s_{t+1} \) is the second highest bid \( b^j_{st+1} \) at time \( t + 1 \). Because each primitive \( \omega^i \) and its clone \( \hat{\omega}^i \) have the same valuations, their bids are equivalent and so credit is conserved through time.

**Algorithm 1 On-Policy Decentralized RL**

1: Initialize: Primitives \( \omega^{1:N} \), Memory \( m^{1:N} \), RL update rule \( f \)
2: while True do
3: \( \triangleright \) Sample Episode
4: \( \triangleright \) while episode has not terminated do
5: \( \omega^{1:N} \) observe state \( s \)
6: \( \omega^{1:N} \) produce bids \( b^1, \ldots, b^N \)
7: Auction selects winner \( \hat{\omega} \) with transformation \( \hat{\phi}_T \)
8: \( \hat{\omega} \) produces \( s' = \hat{\phi}_T(s) \)
9: Record environment reward \( r(s, \hat{\omega}) \)
10: send \( s \rightarrow s' \)
11: end while
12: \( \triangleright \) Compute Utilities
13: for each time-step \( t \) until the end of sampled episode do
14: \( \omega_t \) gets \( U^1_t (\omega_t^{1:N}) = r(s_t, \omega_t) + \gamma \cdot \max_{b^j_{st+1}} - \max_{j \neq i} b^j_t \)
15: Losers \( \omega'_t \) get \( U^2_t (\omega_t^{1:N}) = 0 \)
16: for all primitives \( \omega^i \) do
17: Primitive \( \omega'_t \) stores \( (s_t, b^i_{st}, st+1, U^1_t (\omega'_t)) \) into \( m^i \)
18: end for
19: end for
20: \( \triangleright \) Update
21: if time to update then
22: for all primitives \( \omega^i \) do
23: Update primitive \( \omega^i \) with update rule \( f \) with memory \( m^i \)
24: end for
25: end if
26: end while

6.1. Local Credit Assignment in Space and Time

The global problem requires a solution that is global in space, because the society must collectively work together, and global in time, because the society must maximize expected return across time-steps. But an interesting property of using the auction utility in Equation 2 as an RL objective is that given redundant primitives it takes the form of the
Bellman equation, thereby implicitly coupling all primitives together in space and time. Thus each primitive need only optimize for its immediate utility at each time-step without needing to optimize for its own future utilities. This class of decentralized RL algorithms thus implicitly finds a global solution in space and time using only credit assignment that is \textit{local in space}, because transactions are only between individual agents, and \textit{local in time}, because each primitive need only solve a contextual bandit problem at each time-step.

6.2. Redundancy for Avoiding Suboptimal Equilibria
A benefit of casting local auction utilities as RL objectives is the practical development of learning algorithms that need not assume oracle knowledge of valuations, but unless care is taken with each primitives’ learning environments, the society may not converge to the globally optimal Nash equilibrium described in Section 5. As an example of a suboptimal equilibrium, in a Market MDP with two primitives, even if \( v^1 = 1, v^2 = 2 \), it is a Nash equilibrium for \( b^1 = 100, b^2 = 0 \). Without sufficient competitive pressure to not overbid or underbid, a rogue winner lacks the risk of losing to other primitives with similarly close valuations. Fortunately, the redundancy of a cloned Vickrey society serves a dual purpose of not only preventing market bubbles but also introducing competitive pressure in the primitives’ learning environments in the form of other clones.

7. Experiments
Now, we study how well the cloned Vickrey society can recover the optimal societal Q-function as its Nash equilibrium. We compare several implementations of the cloned Vickrey society against baselines across simple tabular environments in Section 7.1, where the transformations \( \phi_T \) are literal actions. Then in Section 7.2 we demonstrate the broad applicability of the cloned Vickrey society for learning to select options in semi-MDPs and composing functions in a computation graph. Moreover, we show evidence for the advantages of the societal decision making framework over its monolithic counterpart in transferring to new tasks.

Although our characterization of the equilibrium and the learning algorithm are agnostic to the implementation of the primitive, in our bidding policy \( \psi \) is implemented as a neural network that maps the state to the parameters of a Beta distribution, from which a bid is sampled. We used proximal policy optimization (PPO) (Schulman et al., 2017) to optimize the bidding policy parameters.

7.1. Numerical Simulations
In the following simulations we ask: (1) How closely do the bids the primitives learn match their optimal societal Q-values? (2) Does the solution to the global objective emerge from the competition among the primitives? (3) How does redundancy affect the solutions the primitives converge to?

We first consider the Market Bandit, the simplest global MDP with one state, comparing the cloned Vickrey society with societies with different auction mechanisms. Next we consider the Chain environment, a sparse-reward multi-step global MDP designed to study market bubbles. Last we consider the Duality environment, a multi-step global MDP designed to study suboptimal equilibria. We use solitary society to refer to a society without redundant clones.

7.1.1. Market Bandits and Auction Mechanisms
The simulation primarily studies question (1) by comparing the auction mechanism of the cloned Vickrey society against other mechanisms in eliciting bids that match optimal societal Q-values. We compare against first price auction, based on Holland (1985), a solitary society that uses the first price auction mechanism for the local auction, in which the winning primitive pays a price of their bid, rather the second highest bid; against Vickrey Auction, a solitary version of the cloned Vickrey society; and against Environment Reward, a baseline solitary society whose utility function uses only the environment reward, with no price term \( P(b) \).

The environment is a four-armed Market Bandit whose arms correspond to transformations \( \phi \) that deterministically yield reward values of 0.2, 0.4, 0.6, and 0.8. Solitary societies thus would have four primitives, while cloned societies eight. Since we are mainly concerned with question (1), we stochastically drop out a subset of primitives at every round to give each primitive a chance to win and learn its value. In the Market Bandit, the arm rewards directly specify the primitives’ valuations, so if the primitives successfully learned their valuations, we would expect them to learn to
Figure 3. **Implementations.** The table shows the bid price that temporally consecutive winners $\omega_{t+1}$ and $\omega_t$ pay and receive based on three possible implementations of the cloned Vickrey society: $CCV$, $BB$, $V$, with tradeoffs depicted in the Venn diagram. We use $b_{t+1}$ and $b_t$ to denote the highest and second highest bids at time $t$ and $t+1$ respectively.

<table>
<thead>
<tr>
<th></th>
<th>$\omega_t$ Receives</th>
<th>$\omega_{t+1}$ Pays</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CCV$</td>
<td>$b'_t$</td>
<td>$b_{t+1}$</td>
</tr>
<tr>
<td>$BB$</td>
<td>$b'_t$</td>
<td>$b_{t+1}$</td>
</tr>
<tr>
<td>$V$</td>
<td>$b_{t+1}$</td>
<td>$b_{t+1}$</td>
</tr>
</tbody>
</table>

Figure 4. **Learned Bidding Strategies for Chain.** We organize the analysis by distinguishing between the credit-conserving ($CCV$ and $BB$) and the non-credit-conserving ($V$) implementations. The solitary $CCV$ (a) and $BB$ (b) implementations learn to bid very close to 0: $CCV$ because the valuation for a primitive at $t \geq 1$ is only the second-highest bid at $t+1$, resulting in a rapid decay in the valuations leftwards down the chain; $BB$ because each primitive is incentivized to pay as low of a price for winning as possible. The cloned $CCV$ (d) and $BB$ (e) implementations learn to implement a form of return decomposition (Arjona-Medina et al., 2019) that redistributes the terminal reward into a series of positive payoffs back through the chain, each agent getting paid for contributing to moving the society closer to the goal state, where the $CCV$ implementation’s bids are closer to the optimal societal Q-value than those of the $BB$ implementation. Because both the solitary (c) and cloned (f) versions of the $V$ implementations do not conserve credit, they learn to bid close to the optimal societal Q-value, but both suffer from market bubbles where the primitive for going left bids higher than the primitive for going right, even though the optimal global policy is to keep moving right.

7.1.2. **Market MDPs**

Now we consider MDPs that involve multiple time-steps to understand how learning is affected when primitives’ valuations are defined by the bids of future primitives. Redundancy theoretically makes what the winner $\omega_t$ receives and $\omega_{t+1}$ equivalent, but the stochasticity of the bid distribution yields various possible implementations for the cloned Vickrey society in the market MDP—bucket brigade ($BB$), Vickrey ($V$), and credit conserving Vickrey ($CCV$), summarized in Figure 3, with different pros and cons. Note that the solitary bucket brigade society is a multiple time-step analog of the first price auction in Section 7.1.1. We aim to empirically test which of the implementations of the cloned Vickrey society yields best local and global performance.

The **Chain environment and market bubbles.** A primary purpose of the Chain environment (Figure 5a) is to study the effect of redundancy on mitigating market bubbles in the bidding behavior and how that affects global optimality. The Chain environment is a sparse reward finite-horizon environment which initializes the society at $s_0$ on the left, yields a terminal reward of 0.8 if the society enters the goal state $s_3$ on the right, and ends the episode after 20 steps if the society does not reach the goal. Chain thus tests the ability of the society to propagate utility from future agents to past agents in the absence of an immediate environment reward signal. We compare the bidding behaviors of the $BB$, $V$, and $CCV$ implementations of the cloned Vickrey society, as well as those of their solitary counterparts, in Figure 5 and the society’s global learning curve in Figure 7a.

To answer question (1), we observe that redundancy indeed prevents market bubbles, with the cloned $CCV$ implementation bidding closest to the optimal societal Q values. Details are in the caption of Figure 4. When we consider the society’s global learning curve in Figure 7a, the answers to questions (2) and (3) go hand-in-hand: the solitary societies fail to find the globally optimal policy and the cloned $CCV$ implementation has the highest sample efficiency.

The **Duality environment and suboptimal equilibria.** The Chain environment experiments suggested a connection between lack of redundancy and globally suboptimal equilibria, the subtleties of which we explore further in the Duality environment (Figure 5b). The $CCV$ implementation has yielded the best performance so far but does not guarantee Bellman optimality (Figure 3) without redundant primitives. We show that in the Duality environment, without redundant primitives the dominant strategy equilibrium would lead the society to get stuck in a self-loop at $s_1$ indefinitely, even though the global optimal solution would be
Figure 5. Multi-Step MDPs. (a) In the Chain environment, the society starts at state $s_0$ and the goal state is $s_5$. Only activating primitive $\omega_1$ at state $s_4$ yields reward. The optimal global policy is to directly move right by continually activating $\omega_1$. Without credit conservation, the society may get stuck going back and forth between $s_0$ and $s_4$ without reaching the goal. (b) In the Duality environment, the society starts at state $s_0$. $s_{-1}$ is an absorbing state with perpetual negative rewards. The optimal societal policy is to cycle between $s_0$ and $s_1$ to receive unbounded reward, but without redundant primitives, the society may end up in a suboptimal perpetual self-loop at $s_1$.

Figure 6. CCV Bidding Curves for Duality. Each column shows the bidding curves of the solitary (top row) and cloned (bottom row) CCV societies for states $s_{-1}$, $s_0$, and $s_1$. Without redundant primitives to force the second-highest and highest valuations to be equal, the dominant strategy of truthful bidding may not coincide with the globally optimal policy because the solitary CCV implementation does not guarantee Bellman optimality. The bidding curves in (c) show that $\omega^1$ learns a best response of bidding higher than primitive $\omega^0$ at state $s_1$, even though it would be globally optimal for the society if $\omega^0$ wins at $s_1$. Adding redundant primitives causes the second-highest and highest valuations to be equal, causing $\omega^0$ to learn to bid highly as well at $s_1$, which results in a more optimal return as shown in Figure 7b.

Figure 7. Multi-Step MDP Global Learning Curves. We observe that cloned societies are more robust against suboptimal equilibria than solitary societies. Furthermore the cloned CCV implementation achieves the best sample efficiency, suggesting that truthful bidding and credit-conservation are important properties to enforce for enabling the optimal global policy to emerge.

7.2. Semi-MDPs and Computation Graphs

One benefit of framing global decision-making from the perspective of local economic transactions is that the same societal decision-making framework and learning algorithms can be used regardless of the type of transformation $\phi_T$. We now demonstrate in the Two Rooms environment that the cloned Vickrey society can learn more efficiently than a monolithic counterpart to select among pre-trained options to solve a gym-minigrid (Chevalier-Boisvert et al., 2018) navigation task that involves two rooms separated by a bottleneck state. We also demonstrate in the Mental Rotation environment that the cloned Vickrey society can also learn to dynamically compose computation graphs of pre-specified affine transformations for classifying spatially transformed MNIST digits. We use the cloned CCV society for these experiments.

7.2.1. Transferring with Options in Semi-MDPs

We construct a two room environment, Two Rooms which requires opening the red door and reaching either a green goal or blue goal. The transformations $\phi_T$ are subpolicies that have been pre-trained to open the red door, reach the green goal, and reach the blue goal. In the pre-training task, only reaching the green goal gives a non-zero terminal reward and reaching the blue goal does not give reward. In the transfer task, the rewards for reaching the green and blue goals are switched. We compared the cloned Vickrey society against a non-hierarchical monolithic baseline that selects among the low-level gym-minigrid actions as well as a hierarchical monolithic baseline that selects among the pre-trained subpolicies. Both baseline policies sample from a Categorical distribution and are trained with PPO. The cloned Vickrey society is more sample efficient in learning on the pre-training task and significantly faster in adapting
Figure 8. Mental Rotation. The cloned Vickrey society learns to transform the image into a form that can be classified correctly with a pre-trained classifier by composing two of six possible affine transformations of rotation and translation. Clones are indicated by an apostrophe. In this example, the society activated primitive $\omega_2'$ to translate the digit up then primitive $\omega_1$ to rotate the digit clockwise. Though the bidding policies $\psi_i$ and $\psi_i'$ of the clones $\omega_i$ and $\omega_i'$ have the same parameters, their sampled bids may be different because the bidding policies are stochastic.

Figure 9. Two Rooms. The cloned Vickrey society adapts more quickly than the hierarchical monolithic baseline in both the pre-training and the transfer tasks. We hypothesize this better learning efficiency is due to the inherent modularity in the credit assignment mechanism of decentralized reinforcement learning. This task is difficult enough that the non-hierarchical monolithic baseline does not learn to solve the task from scratch.

to the transfer task (Figure 9). Our hypothesized explanation for this efficiency is that the local credit assignment mechanisms of a society parallelize learning across space and time, a property that is not true of the global credit assignment mechanisms of a monolithic learner.

7.2.2. COMPOSING DYNAMIC COMPUTATION GRAPHS

We adapt the Image Transformations task from Chang et al. (2018) as what we call the Mental Rotation environment (Figure 8), in which MNIST images have been transformed with a composition of one of two rotations and one of four translations. There are $60,000 \times 8 = 240,000$ possible unique inputs, meaning learning an optimal global policy involves training primitives across 240,000 local auctions.

The transformations $\phi_T$ are pre-specified affine transformations. The society must transform the image into a form that can be classified correctly with a pre-trained classifier, with a terminal reward of 1 if the predicted label is correct and 0 otherwise. The cloned Vickrey society converges to a mean return of 0.933 with a standard deviation of 0.014.

8. Conclusion

This work formally defines the societal decision-making framework and proves the optimality of a society – the cloned Vickrey society – for solving a problem that was first posed in the AI literature in the 1980s: to specify the incentive structure that causes the global solution of a society to emerge as the equilibrium strategy profile of self-interested agents. For training the society, we further proposed a class of decentralized reinforcement learning algorithms whose global objective decouples in space and time into simpler local objectives. We have demonstrated the generality of this framework for selecting actions, options, and computations as well as its potential advantages for transfer learning.

The generality of the societal decision-making framework opens much opportunity for future work in decentralized reinforcement learning. A society’s inherent modular structure suggests the potential of reformulating problems with global credit-assignment paths into problems with much more local credit-assignment paths that offer more parallelism and independence in the training of different components of a learner. Understanding the learning dynamics of multi-agent societies continues to be an open problem of research. It would be exciting to explore algorithms for constructing and learning societies in which the primitives are also societies themselves. We hope that the societal decision-making framework and its associated decentralized reinforcement learning algorithms provide a foundation for future work exploring the potential of creating AI systems that reflect the collective intelligence of multi-agent societies.
Acknowledgements

The authors are grateful to the anonymous reviewers, Michael Janner, and Glen Berseth for their feedback on this paper, as well as Aviral Kumar, Sam Toyer, Marco Casuanno-Towner, Abhishek Gupta, and Dinesh Jayaraman for helpful discussions. This research was supported by AFOSR grant FA9550-18-1-0077, NSF CCF-1717899, and Google Cloud Platform. MC is supported by the National Science Foundation Graduate Research Fellowship Program. SK is supported by Berkeley Engineering Student Services.

References


Plato. Republic. 380 B.C.


