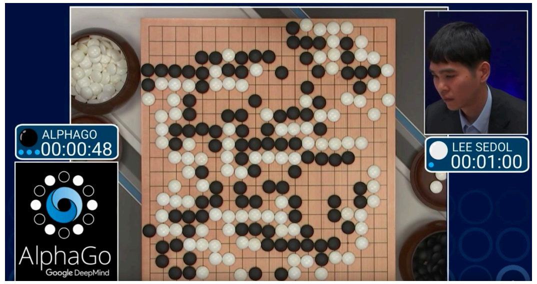
Global Decision-Making via Local Economic Transactions



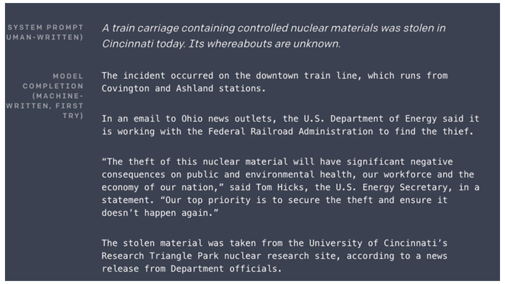


Michael Chang Sid Kaushik Matt Weinberg Tom Griffiths Sergey Levine

Much Success So Far

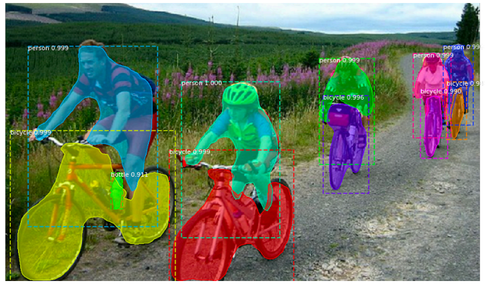


Game Playing
Silver et al. (2016)



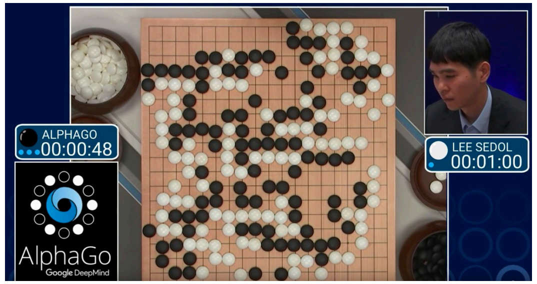
Natural Language Processing
Radford et al. (2019)

Robotics
Levine et al. (2016)

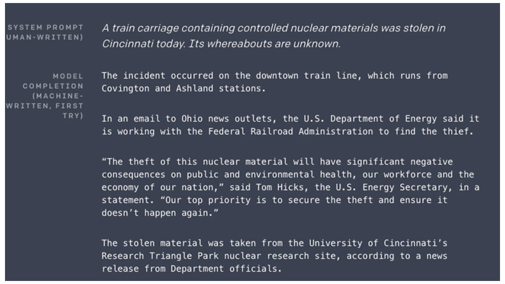


Computer Vision
He et al. (2017)

Much Success So Far: Monolithic Optimization

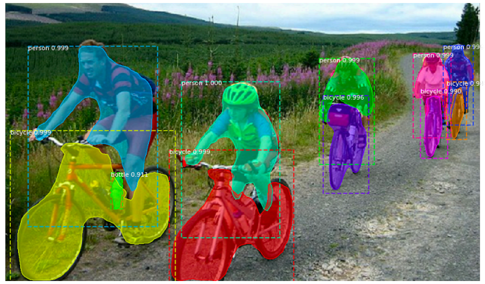


Game Playing
Silver et al. (2016)



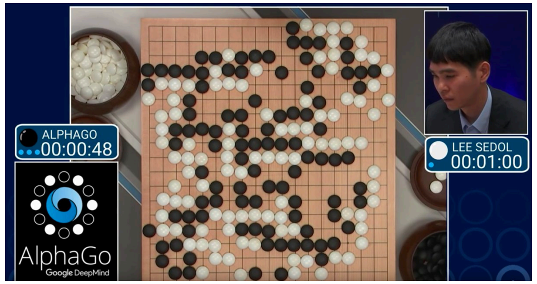
Natural Language Processing
Radford et al. (2019)

Robotics
Levine et al. (2016)

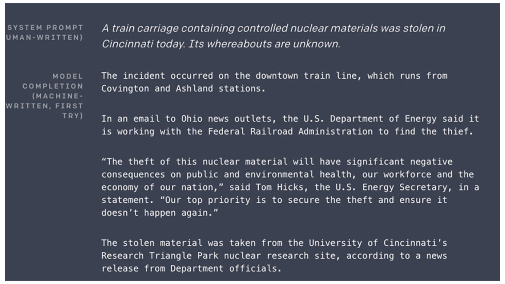


Computer Vision
He et al. (2017)

Much Success So Far: Monolithic Optimization

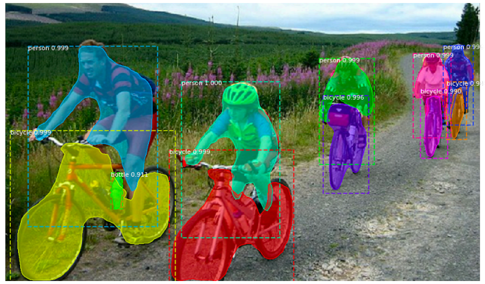


Game Playing
Silver et al. (2016)



Natural Language Processing
Radford et al. (2019)

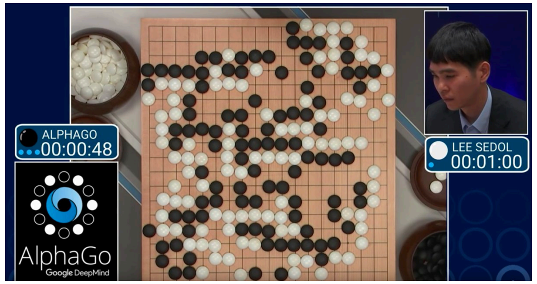
Robotics
Levine et al. (2016)



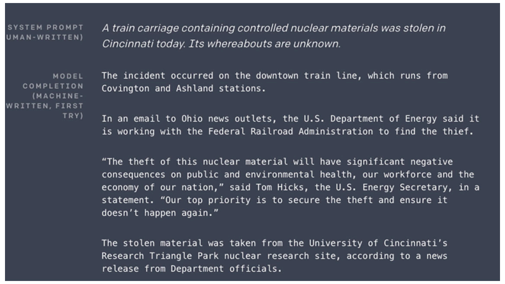
Computer Vision
He et al. (2017)

One optimization problem

Much Success So Far: Monolithic Optimization



Game Playing
Silver et al. (2016)



SYSTEM PROMPT UMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

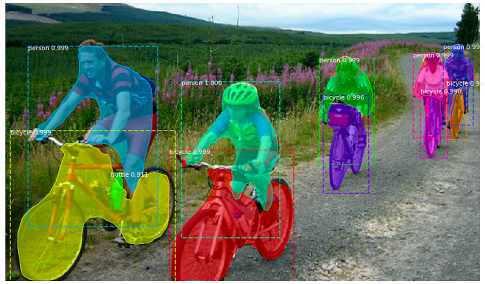
In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

Natural Language Processing
Radford et al. (2019)

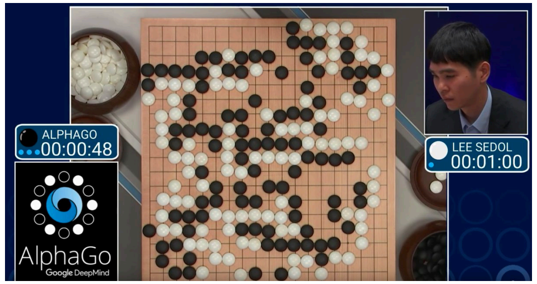
Robotics
Levine et al. (2016)



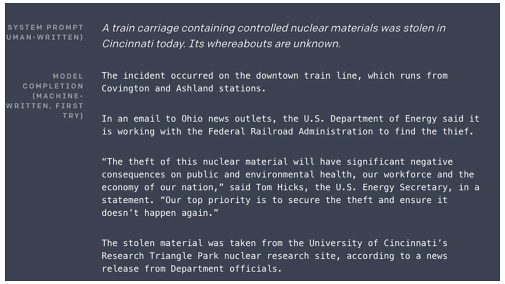
Computer Vision
He et al. (2017)

One optimization problem
One agent

Much Success So Far: Monolithic Optimization



Game Playing
Silver et al. (2016)



SYSTEM PROMPT
UMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL
COMPLETION
(MACHINE-
WRITTEN, FIRST
TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

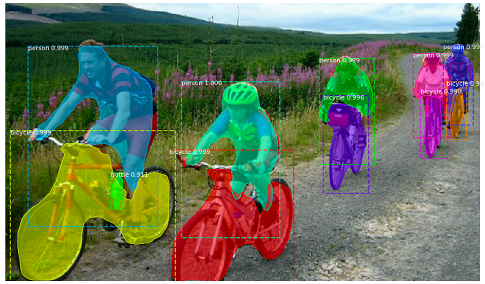
In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

Natural Language Processing
Radford et al. (2019)

Robotics
Levine et al. (2016)



Computer Vision
He et al. (2017)

One optimization problem
One agent
One objective

Decentralized Optimization

Corporation

One optimization problem
One agent
One objective

Decentralized Optimization

Corporation

Many optimization problems
Many agents
Many objectives

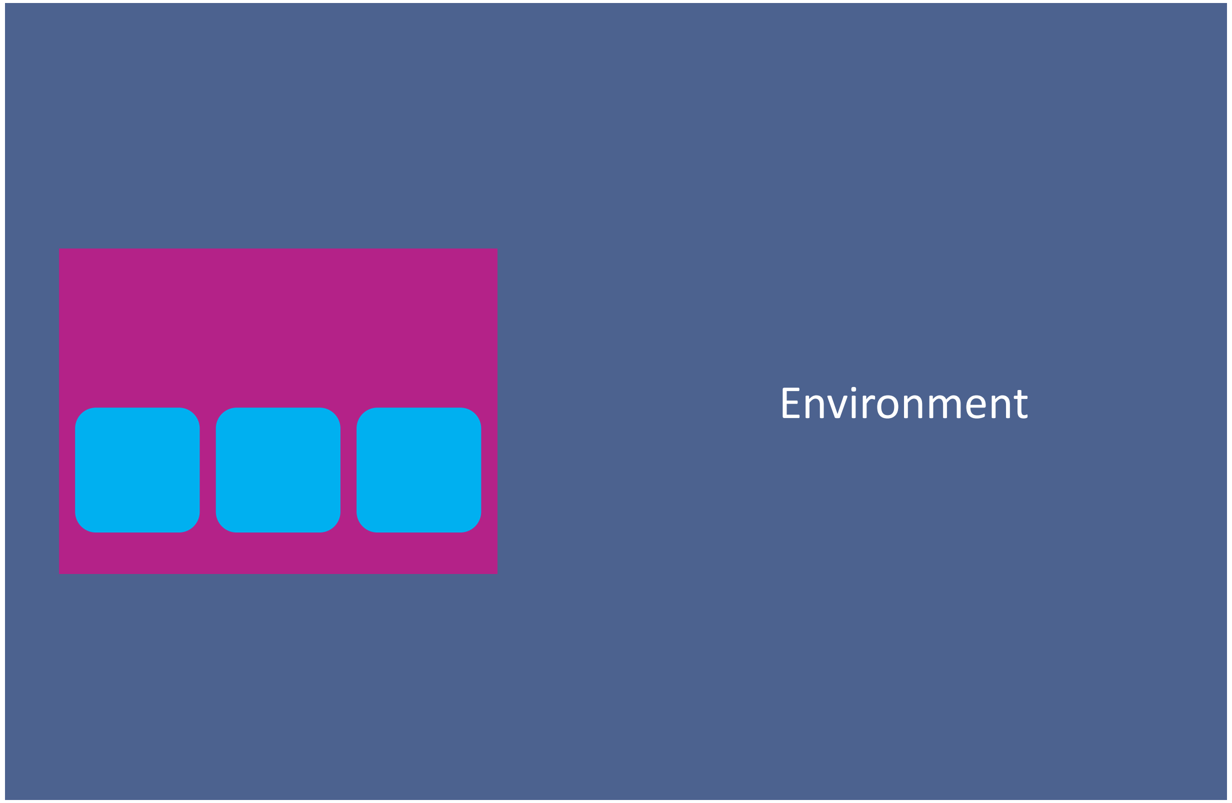
Decentralized Optimization

Many *local* optimization problems
Many *local* agents
Many *local* objectives

→

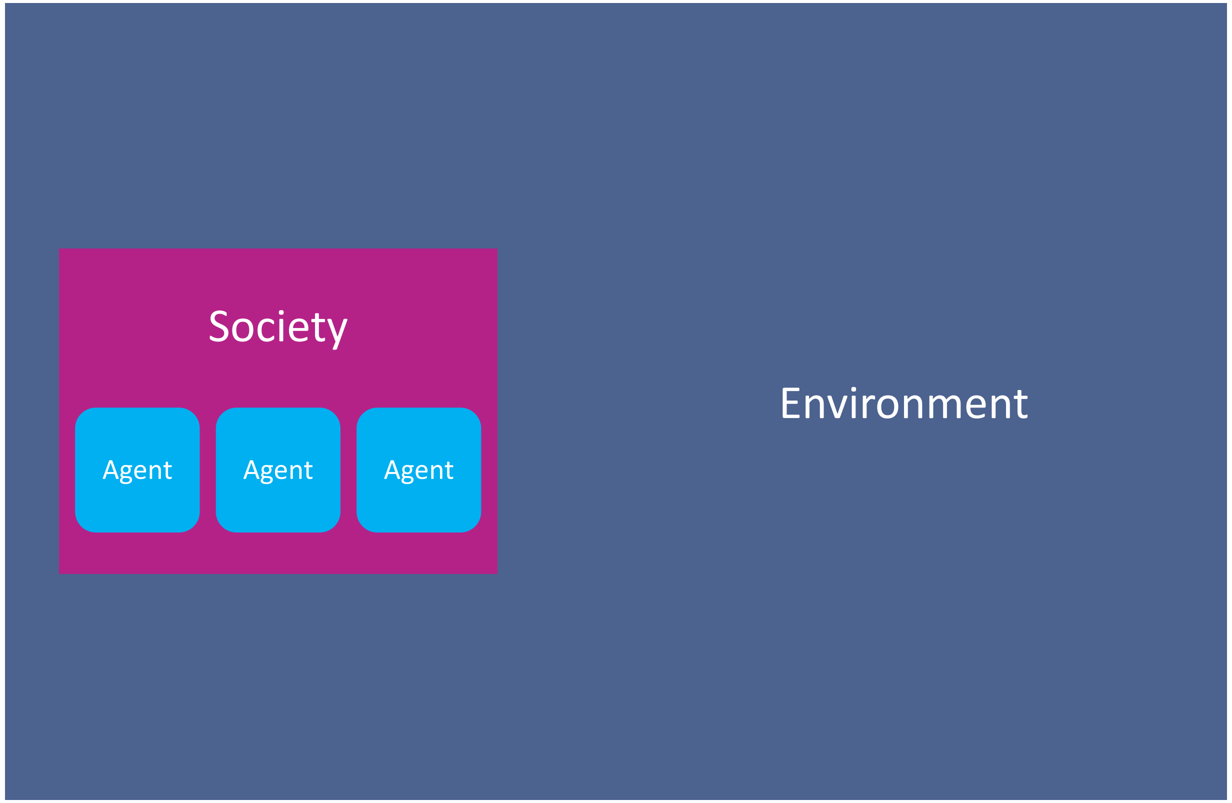
Emergent *global* optimization problem
Emergent *global* agent
Emergent *global* objective

Decentralized Optimization

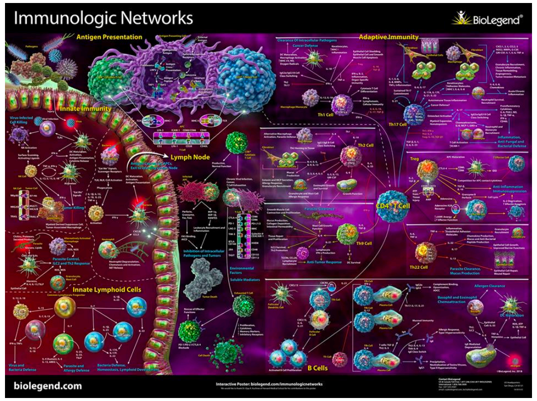


Decentralized Optimization

Decentralized Optimization



Decentralized Optimization

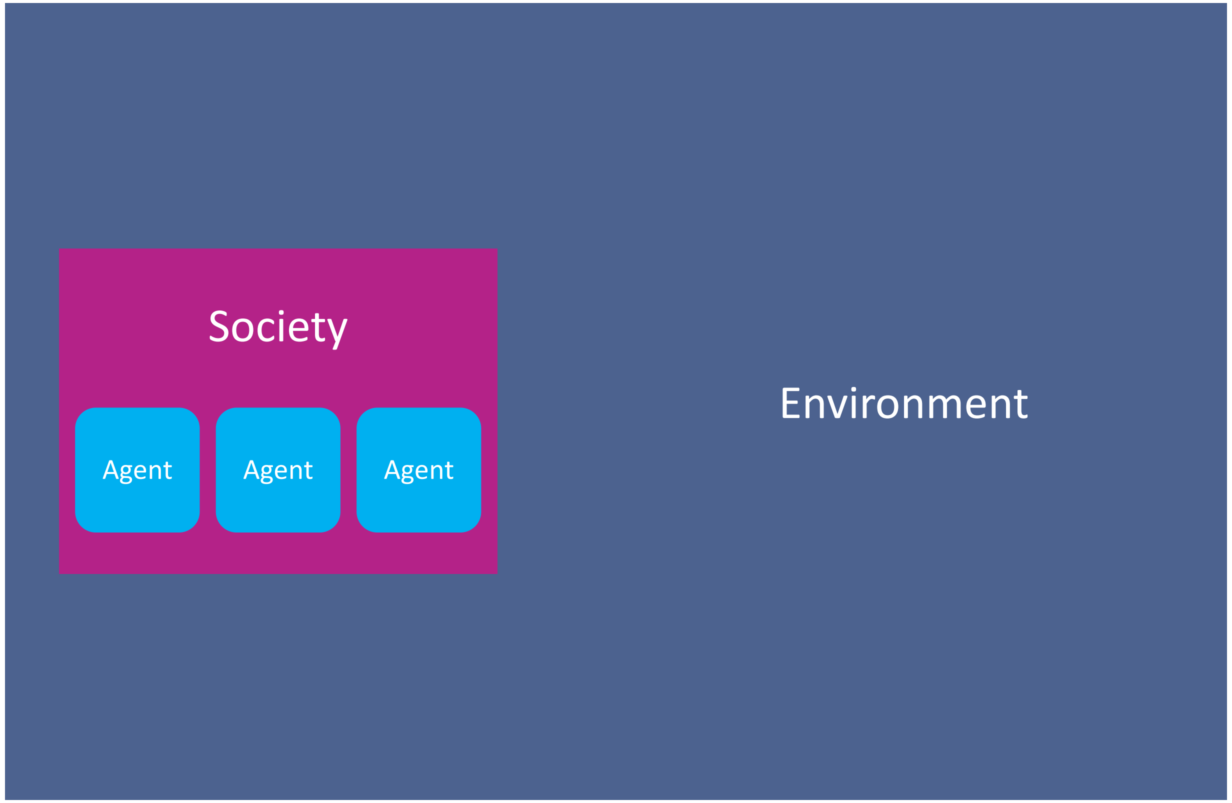


Biological Processes

Ecosystems

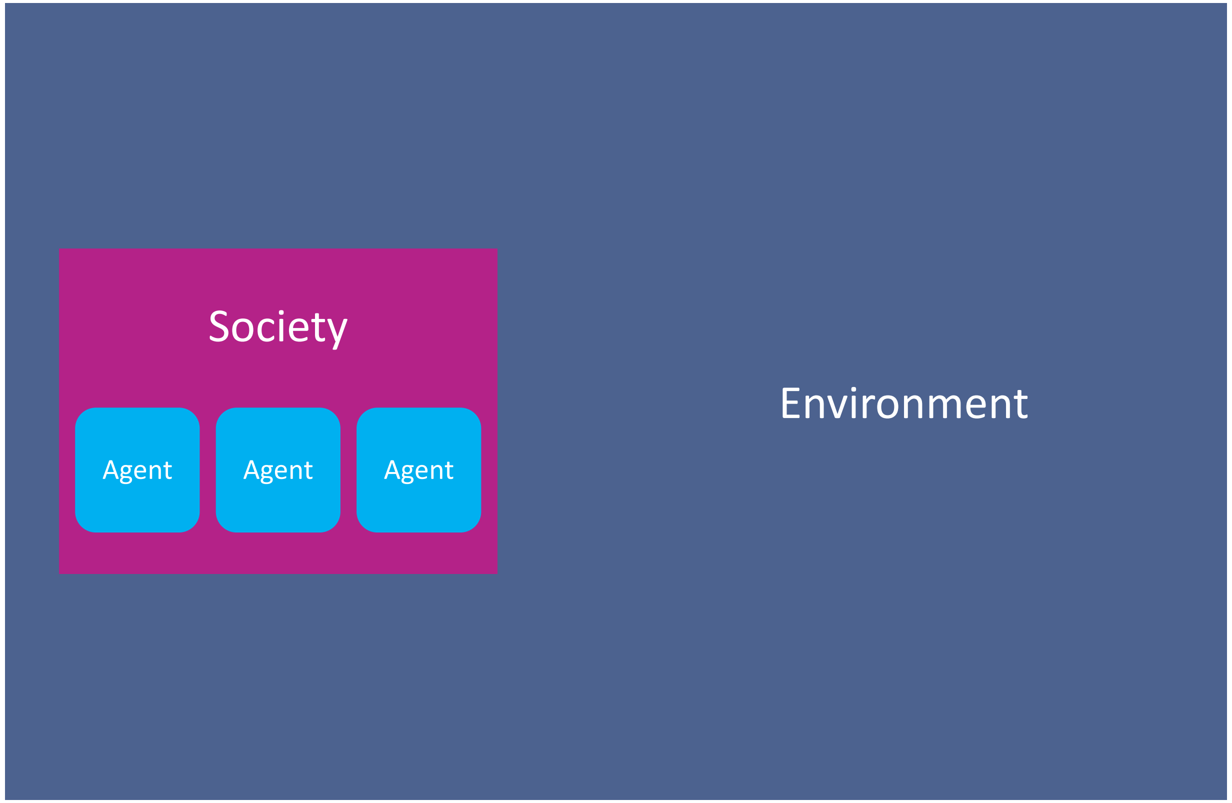
Economies

Organizations



Optimization at Two Levels of Abstraction

Challenge
How can we build machine learning algorithms that relate the global level of the society and the local level of the agent?

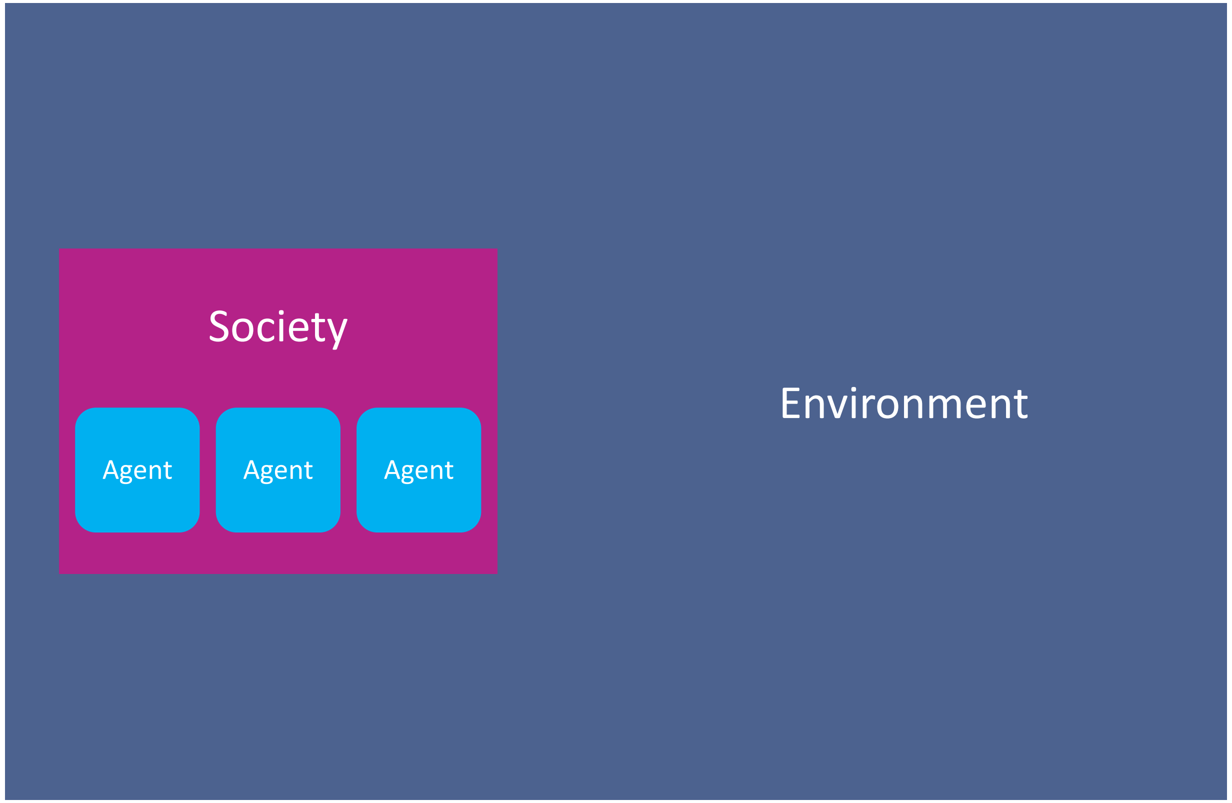


Optimization at Two Levels of Abstraction

Challenge
How can we build machine learning algorithms that relate the global level of the society and the local level of the agent?

Implications

- Enable the design of learning algorithms that are inherently modular

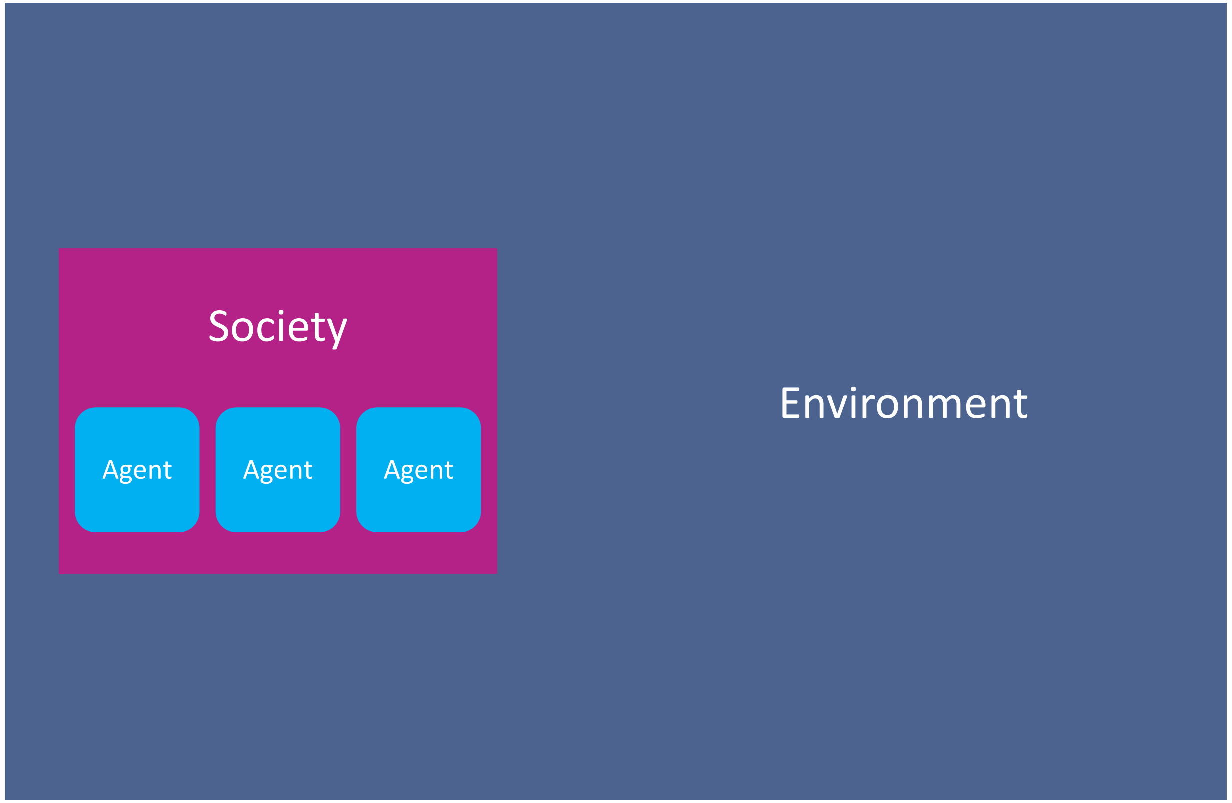


Optimization at Two Levels of Abstraction

Challenge
How can we build machine learning algorithms that relate the global level of the society and the local level of the agent?

Implications

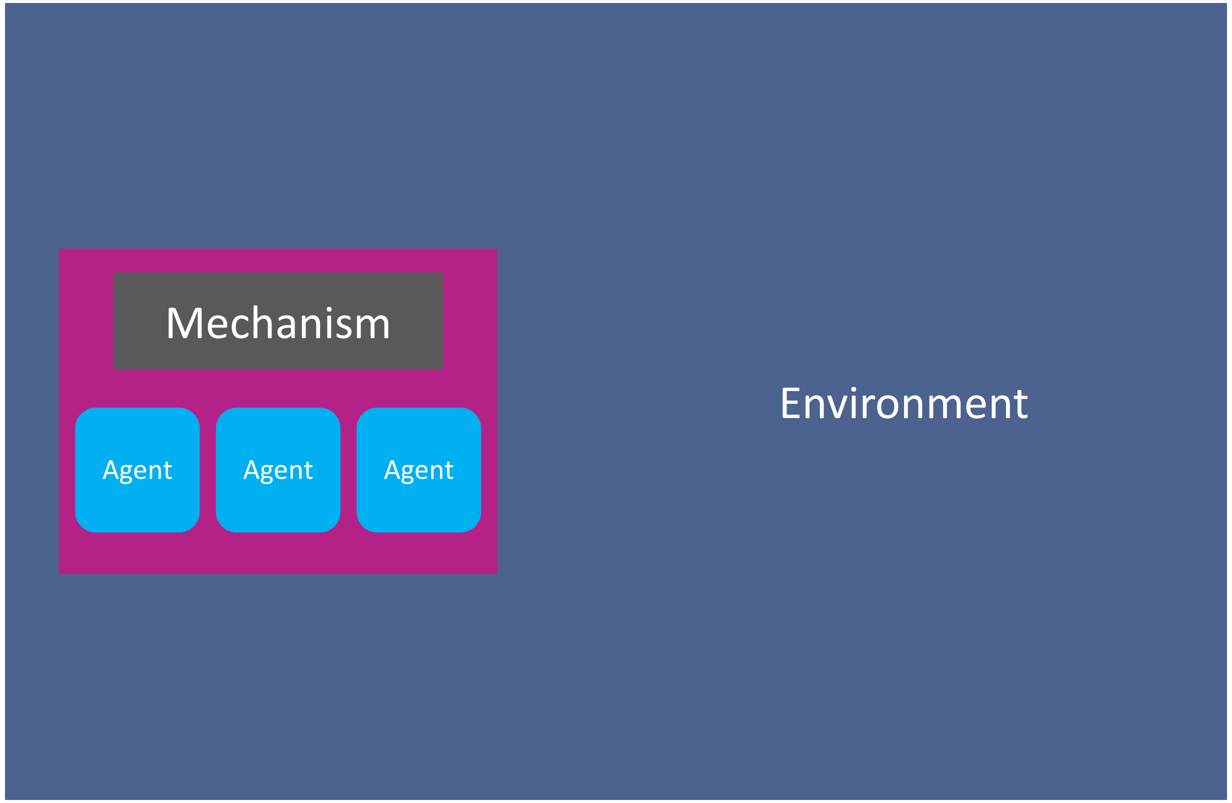
- Enable the design of learning algorithms that are inherently modular
- Provide a recipe for engineering and analyzing a multi-agent system to achieve a desired global outcome



Optimization at Two Levels of Abstraction

Challenge
How can we build machine learning algorithms that relate the global level of the society and the local level of the agent?

Implications
- Enable the design of learning algorithms that are inherently modular
- Provide a recipe for engineering and analyzing a multi-agent system to achieve a desired global outcome

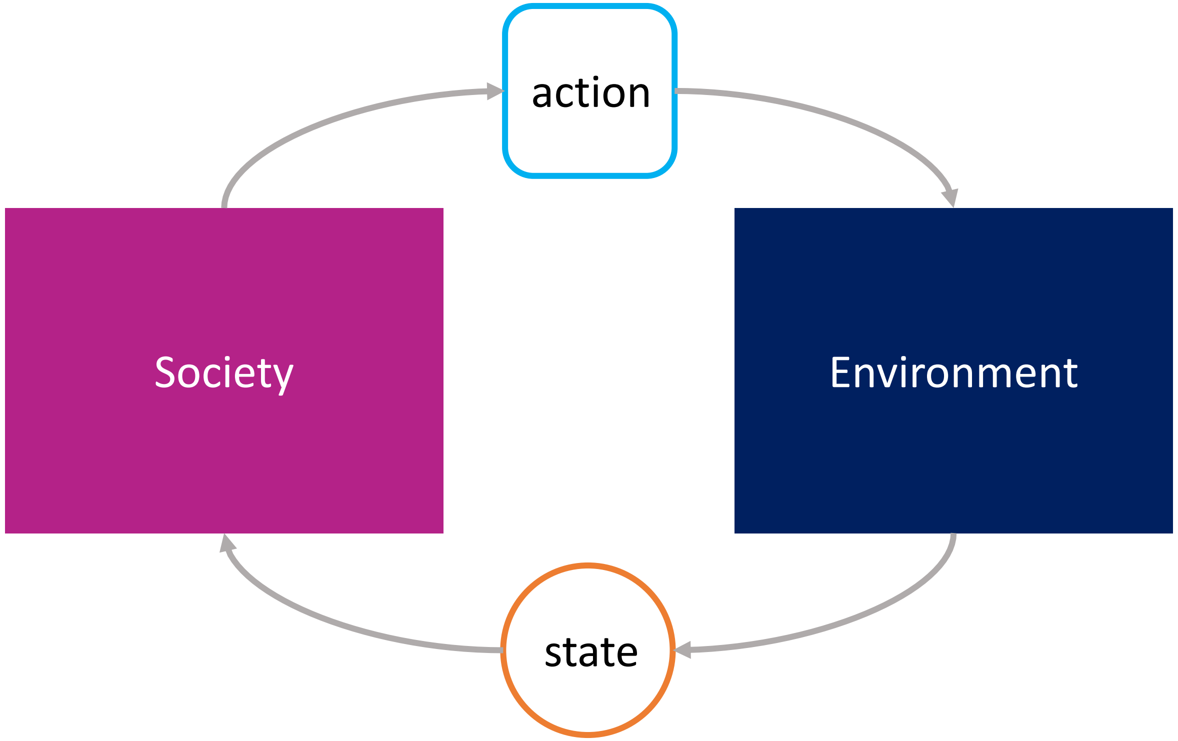


This Paper

This Paper: Assumptions

This Paper: Assumptions

Assumptions
- Sequential decision-making setting

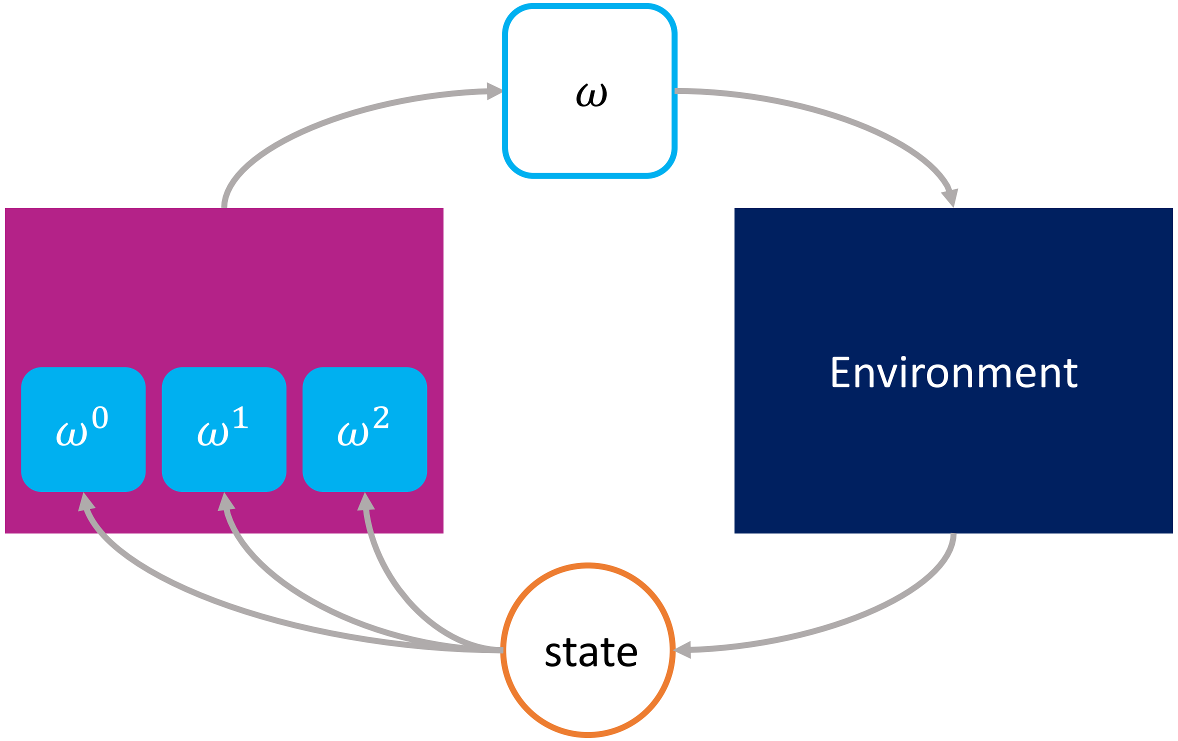


Global MDP

This Paper: Assumptions

Assumptions

- Sequential decision making setting
- Each agent produces a specialized transformation to the state (e.g. a literal action)

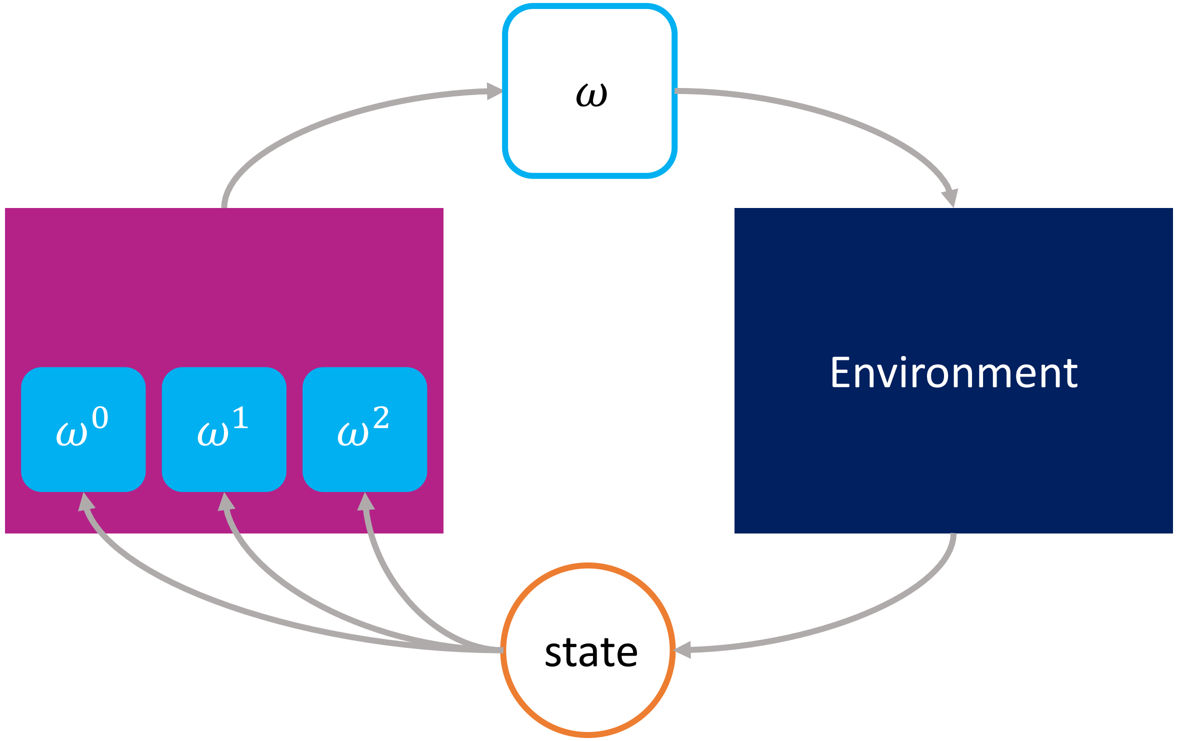


"Global MDP"

This Paper: Assumptions

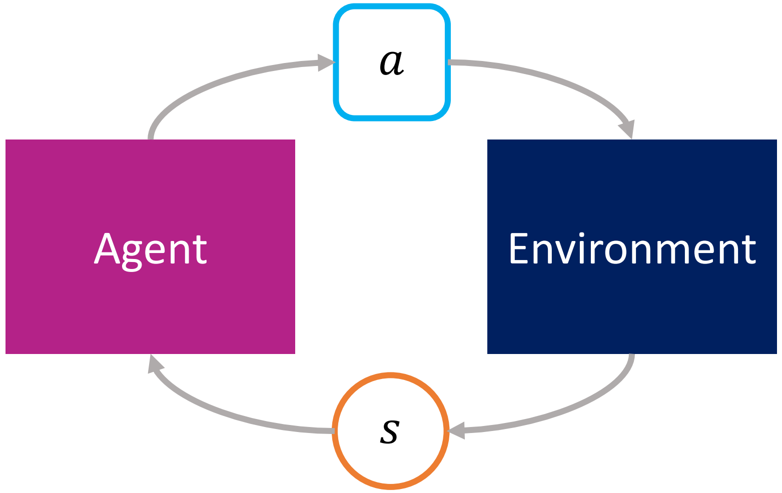
Assumptions

- Sequential decision making setting
- Each agent produces a specialized transformation to the state (e.g. a literal action)
- Only one agent activates at each time step

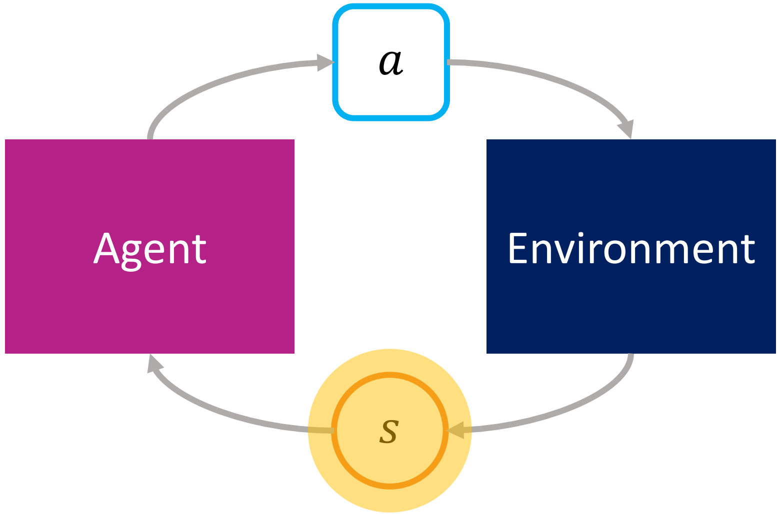


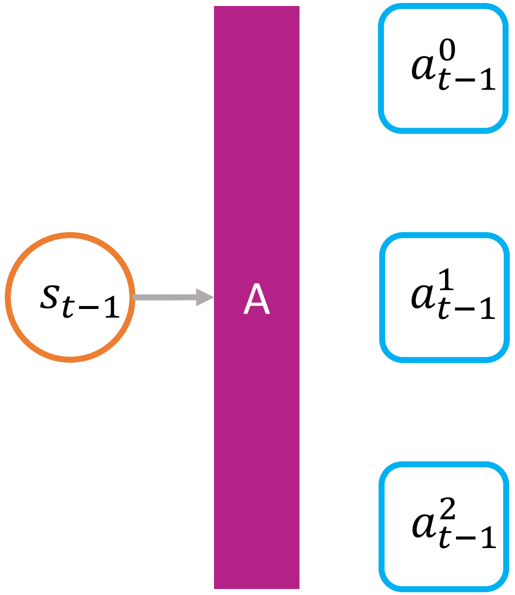
“Global MDP”

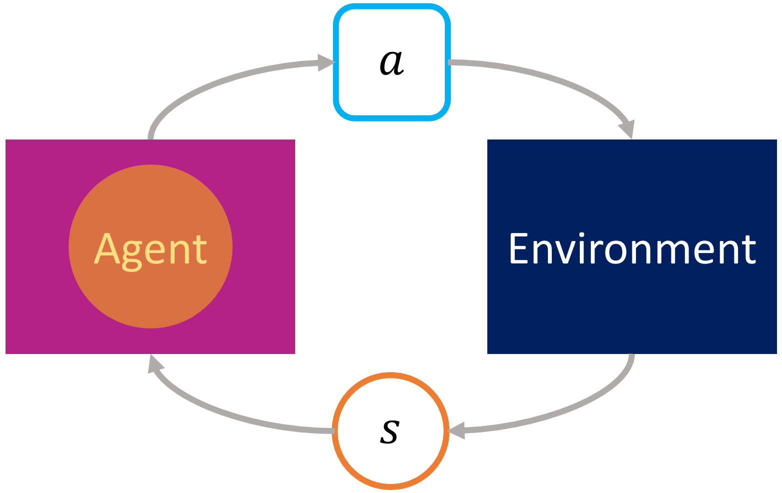
Intuition

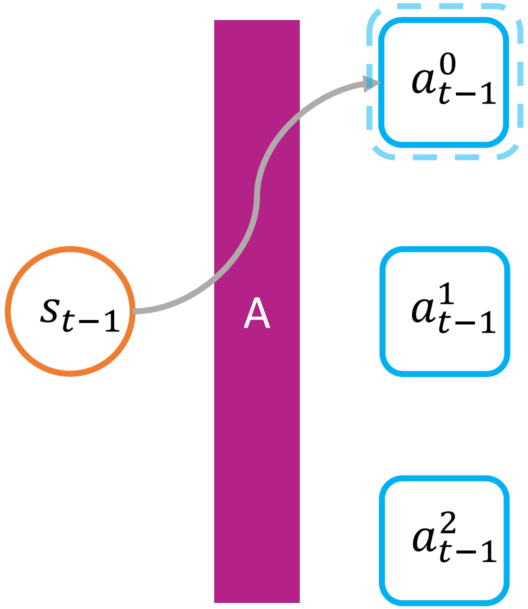


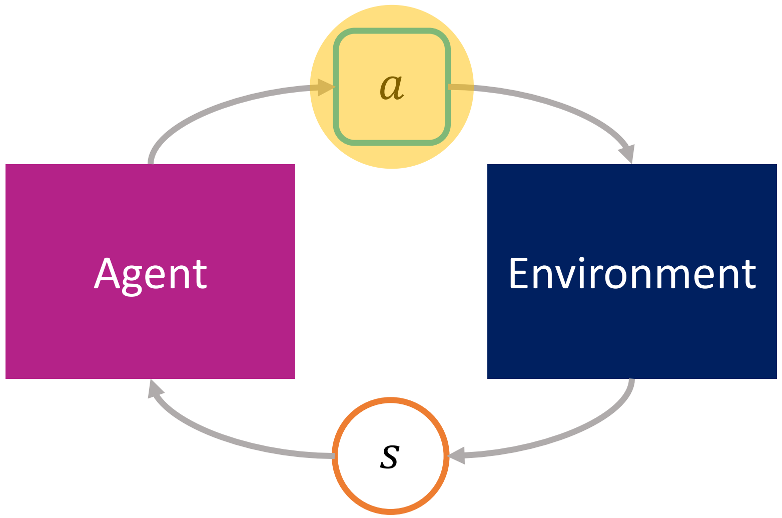

s_{t-1}

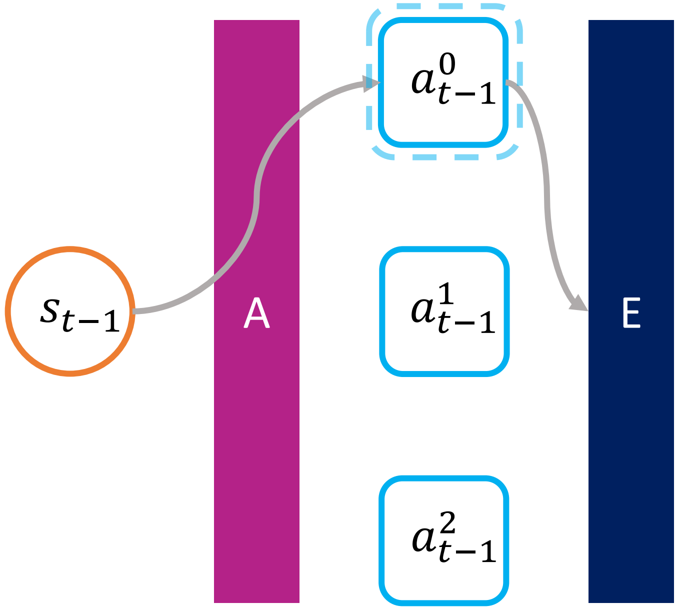

a
Agent
Environment
s


s_{t-1}
A
a_{t-1}^0
a_{t-1}^1
a_{t-1}^2

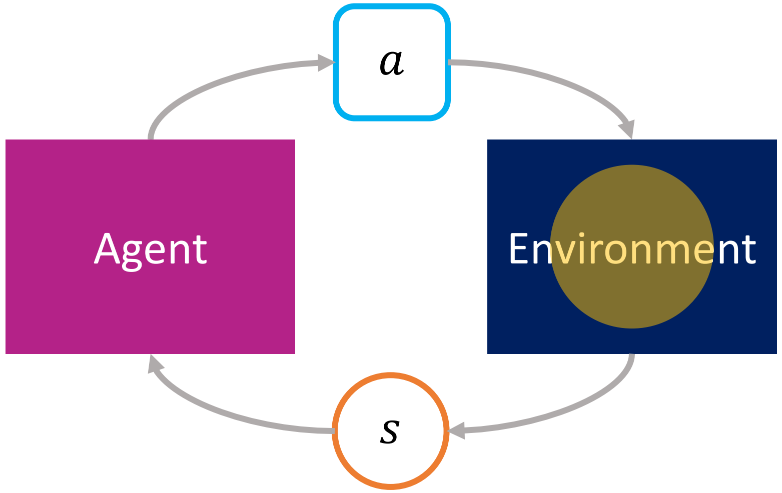

a
Agent
Environment
s


A
a_{t-1}^0
s_{t-1}
a_{t-1}^1
a_{t-1}^2

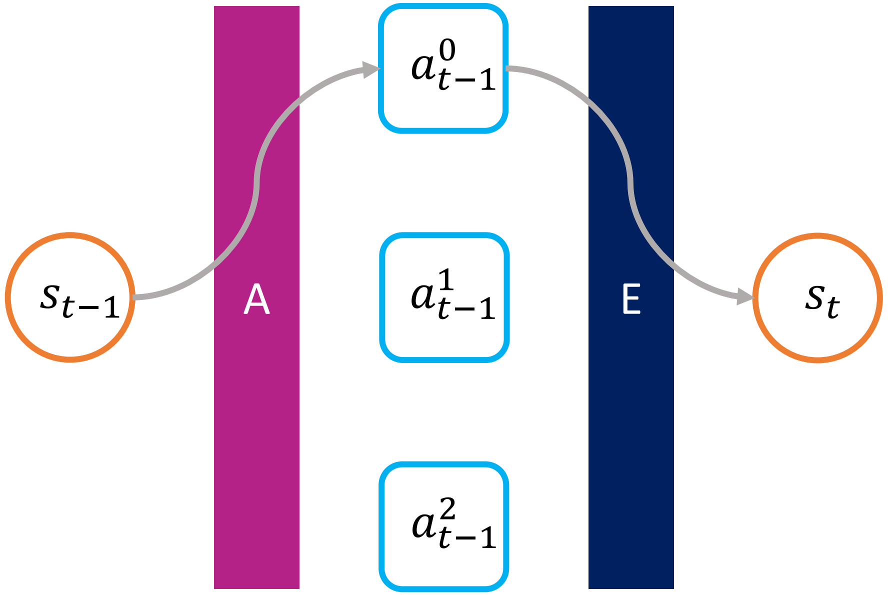

a
Agent
Environment
s

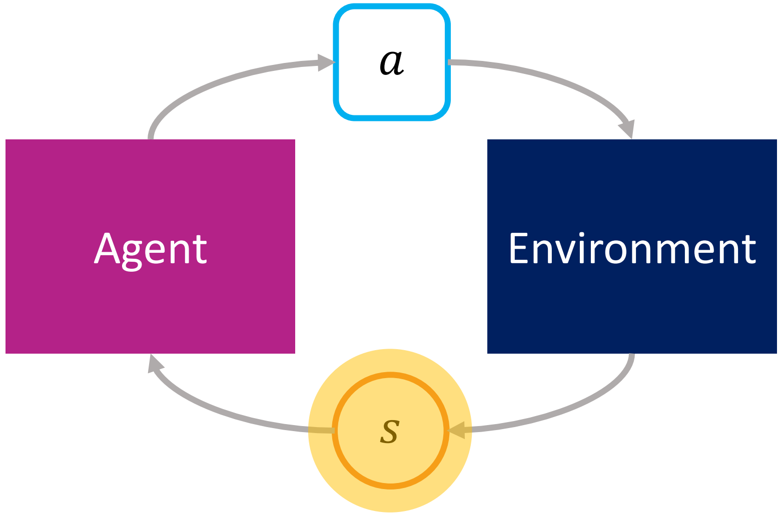


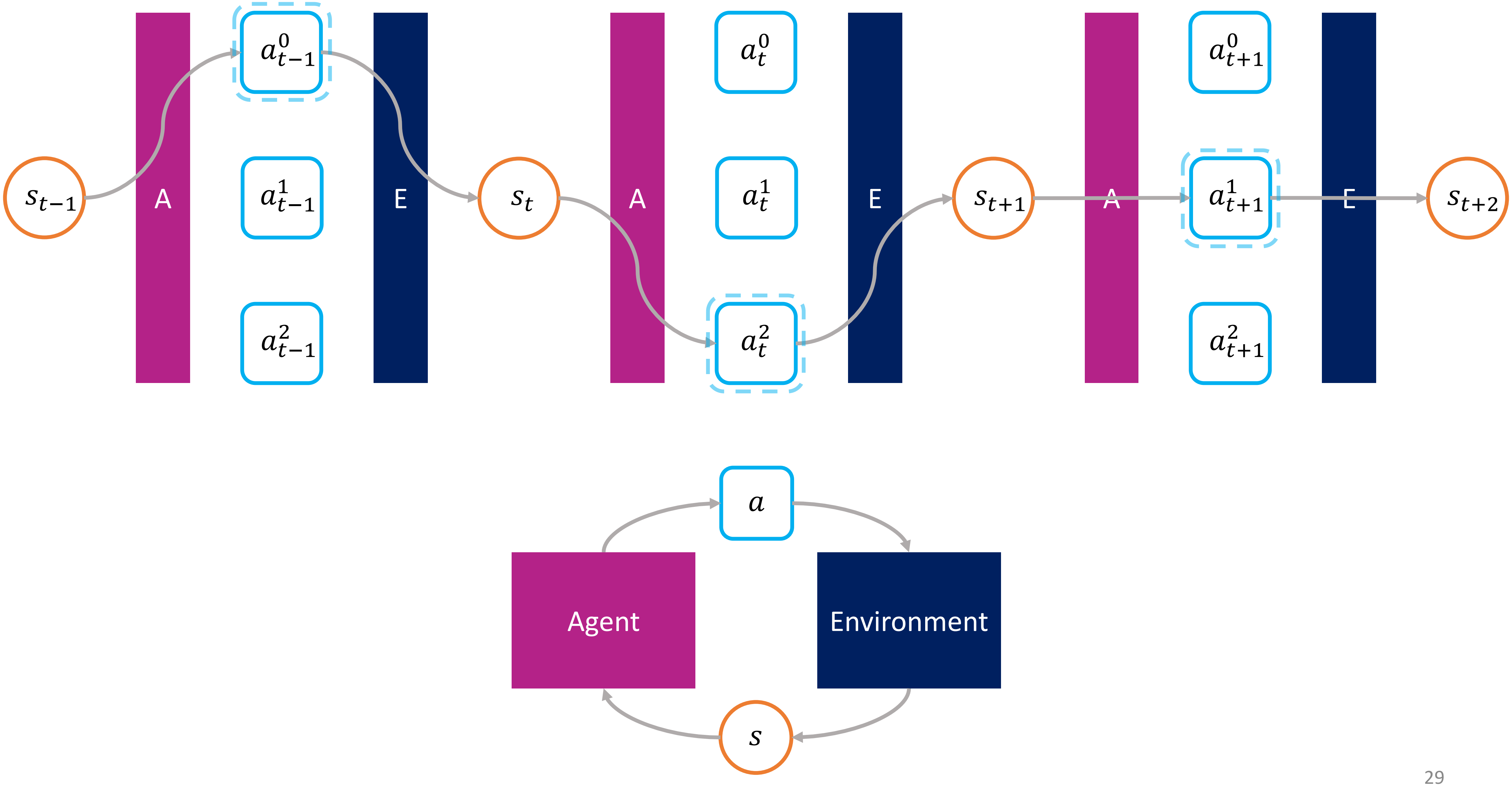
s_{t-1}
A
a^0_{t-1}
a^1_{t-1}
a^2_{t-1}
E

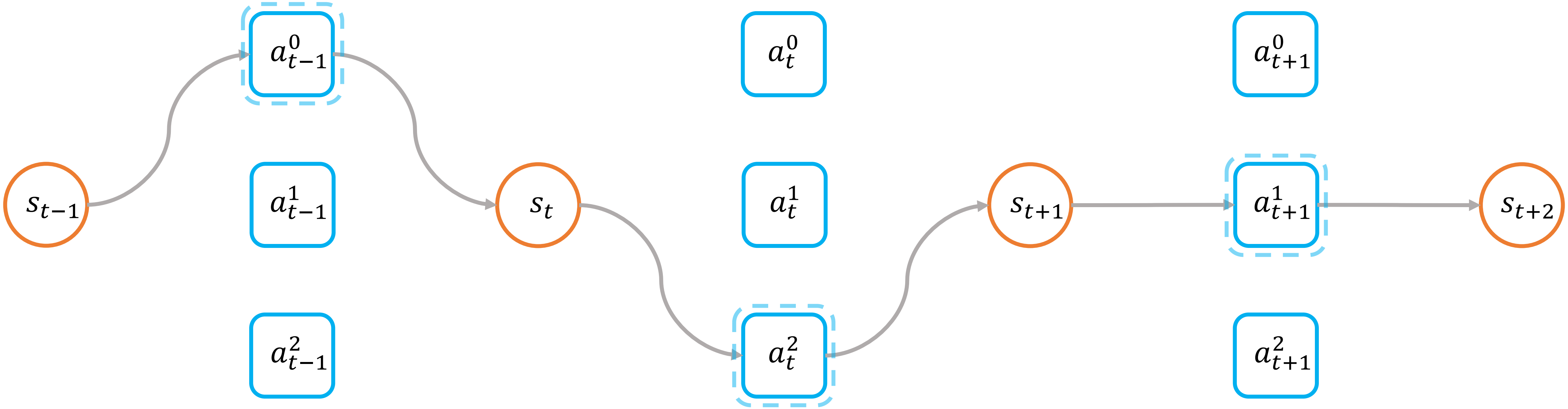


a
Agent
Environment
s

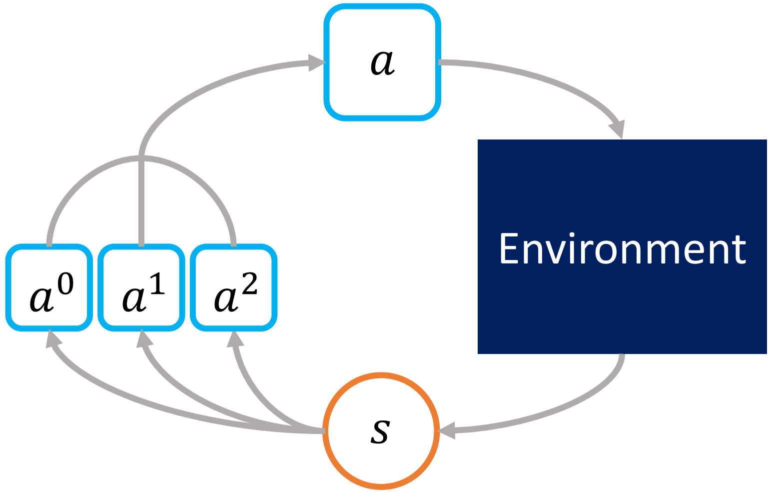

s_{t-1}
A
a_{t-1}^0
a_{t-1}^1
a_{t-1}^2
E
s_t


a
Agent
Environment
s

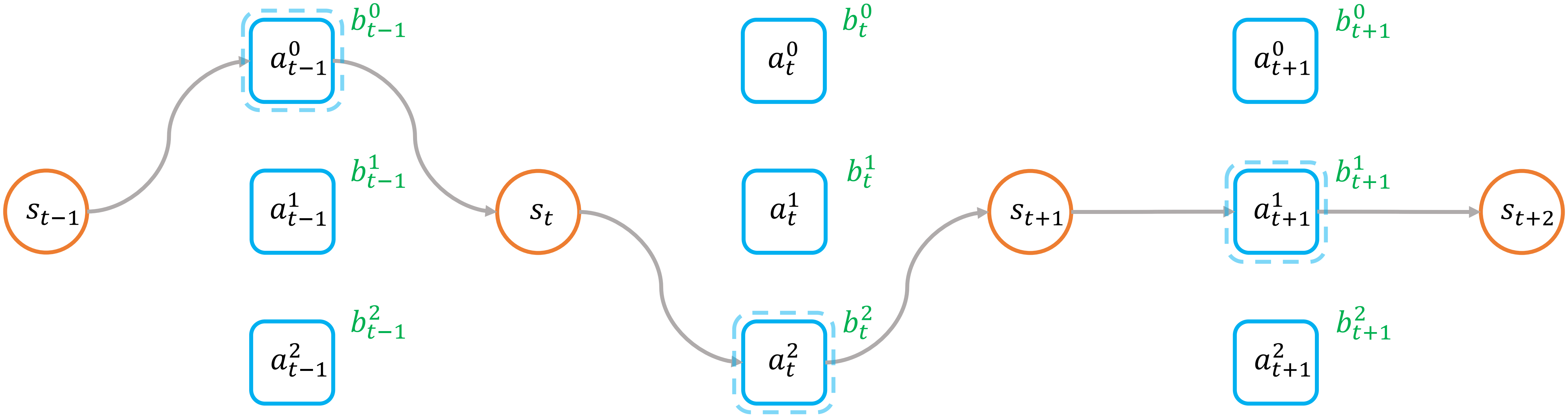

s_{t-1}
A
a^0_{t-1}
a^1_{t-1}
a^2_{t-1}
E
s_t
A
a^0_t
a^1_t
a^2_t
E
s_{t+1}
A
a^0_{t+1}
a^1_{t+1}
a^2_{t+1}
E
s_{t+2}
a
Agent
Environment
s

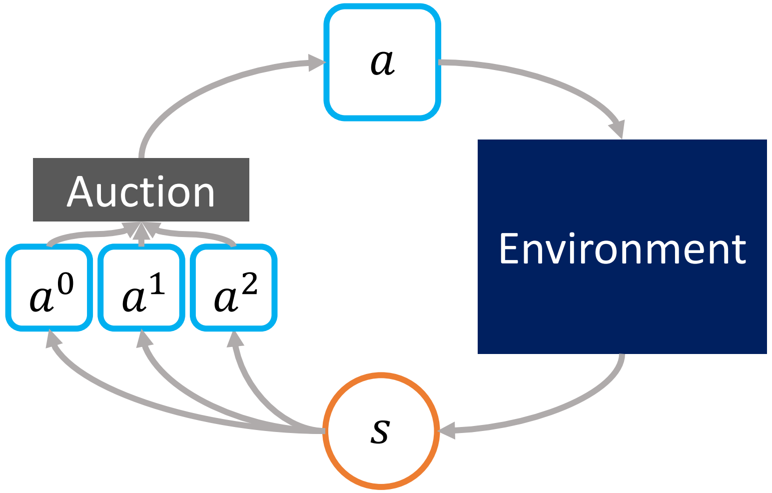


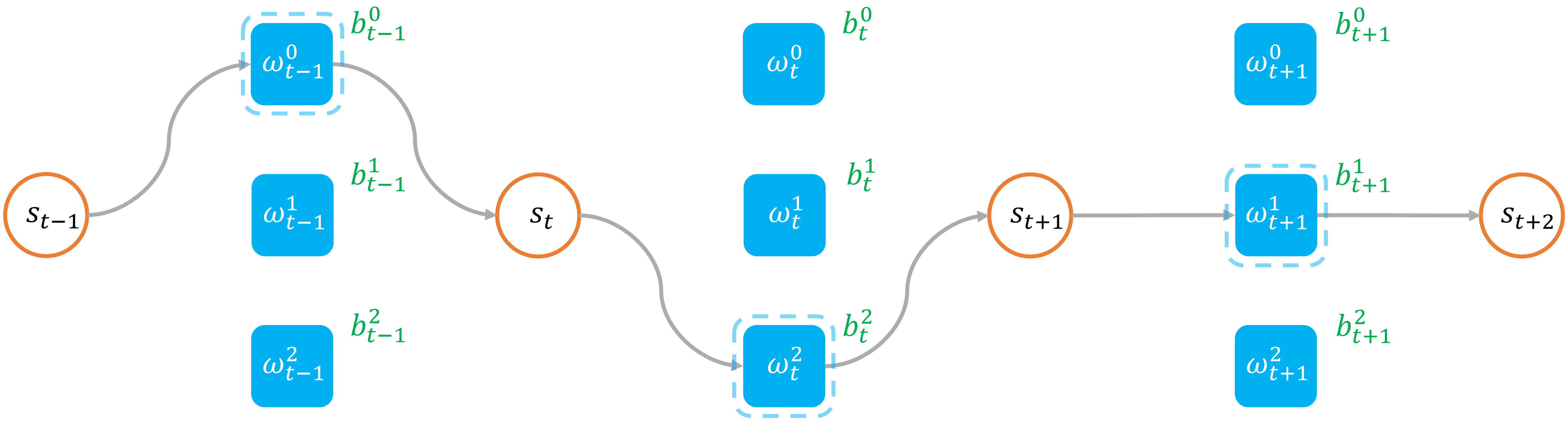
s_{t-1}
a_{t-1}^0
a_{t-1}^1
a_{t-1}^2
s_t
a_t^0
a_t^1
a_t^2
s_{t+1}
a_{t+1}^0
a_{t+1}^1
a_{t+1}^2
s_{t+2}



a
Environment
a^0
a^1
a^2
s

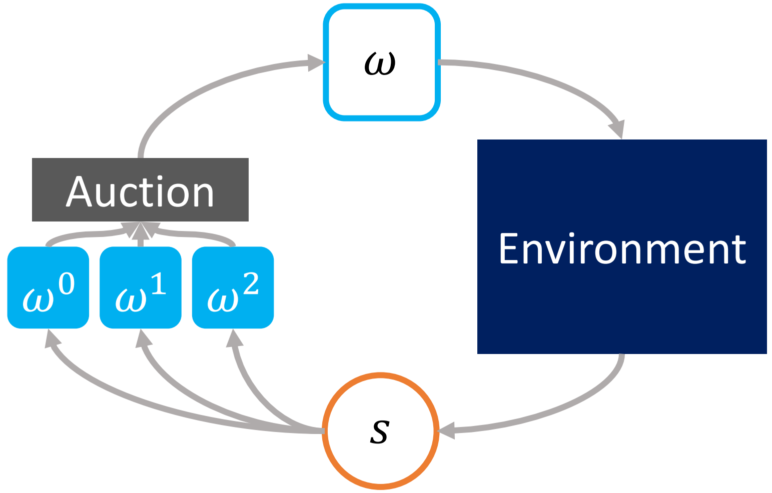

s_{t-1}
a^0_{t-1} b^0_{t-1}
a^1_{t-1} b^1_{t-1}
a^2_{t-1} b^2_{t-1}
s_t
a^0_t b^0_t
a^1_t b^1_t
a^2_t b^2_t
s_{t+1}
a^0_{t+1} b^0_{t+1}
a^1_{t+1} b^1_{t+1}
a^2_{t+1} b^2_{t+1}
s_{t+2}

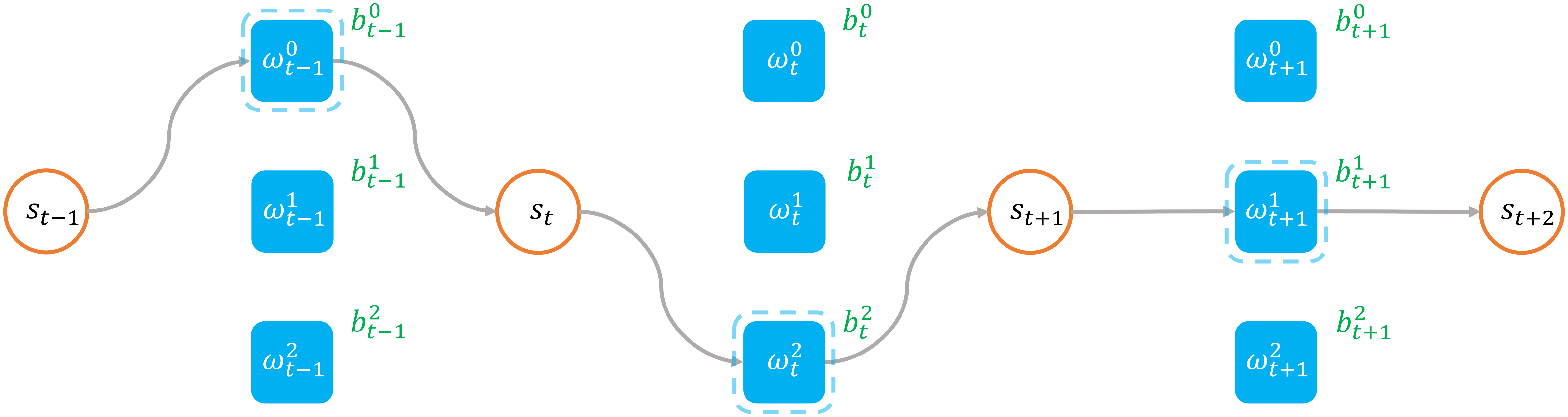

a
Auction
Environment
a^0 a^1 a^2
s

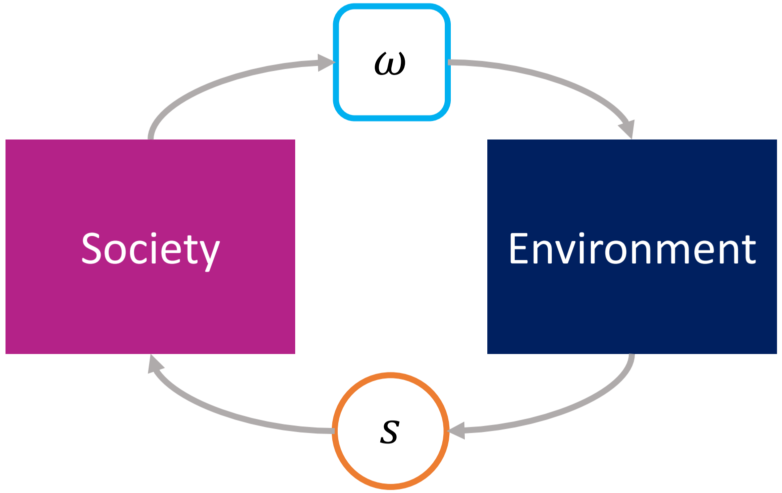


Local Auction

Action Space: bids b
Objective: optimize utility in auction

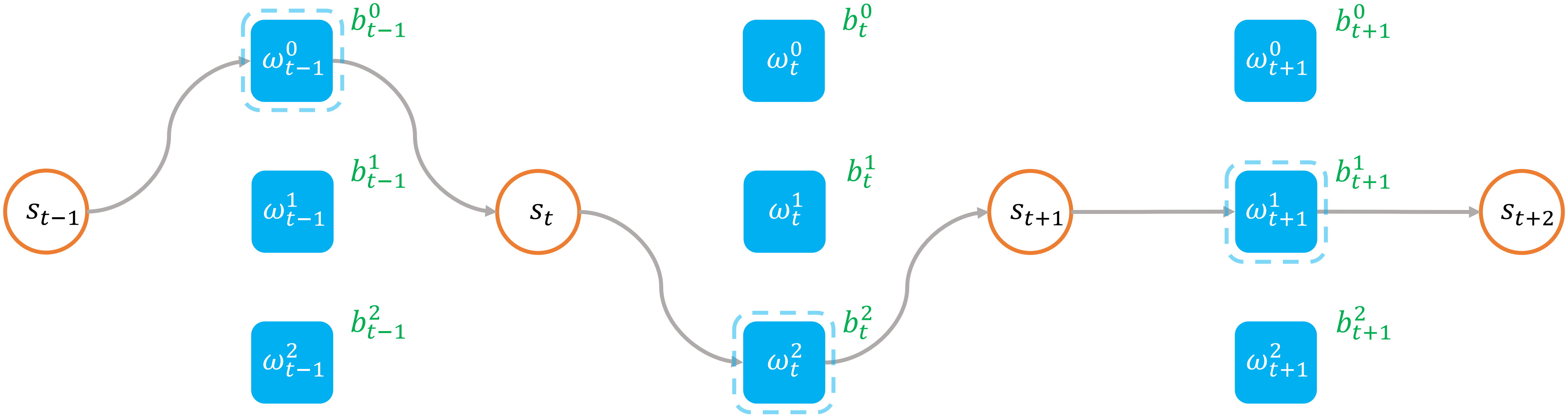






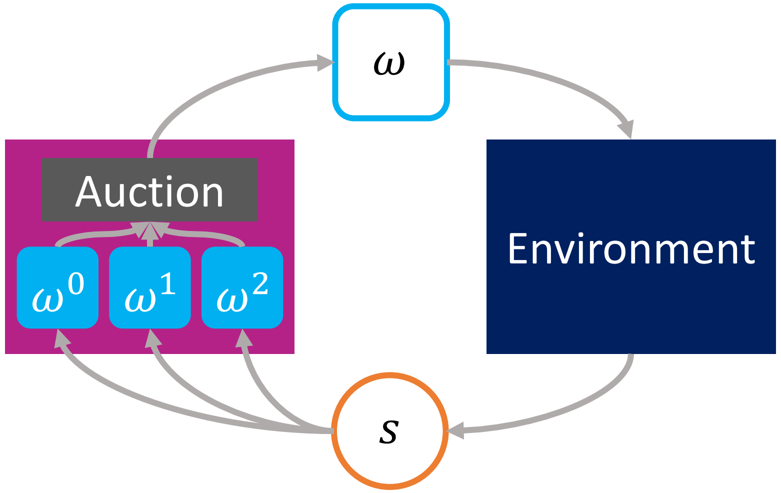
Global MDP

Action Space: agents ω
Objective: optimize return in environment



Local Auction

Action Space: bids b
Objective: optimize utility in auction



Global MDP

Action Space: agents ω
Objective: optimize return in environment

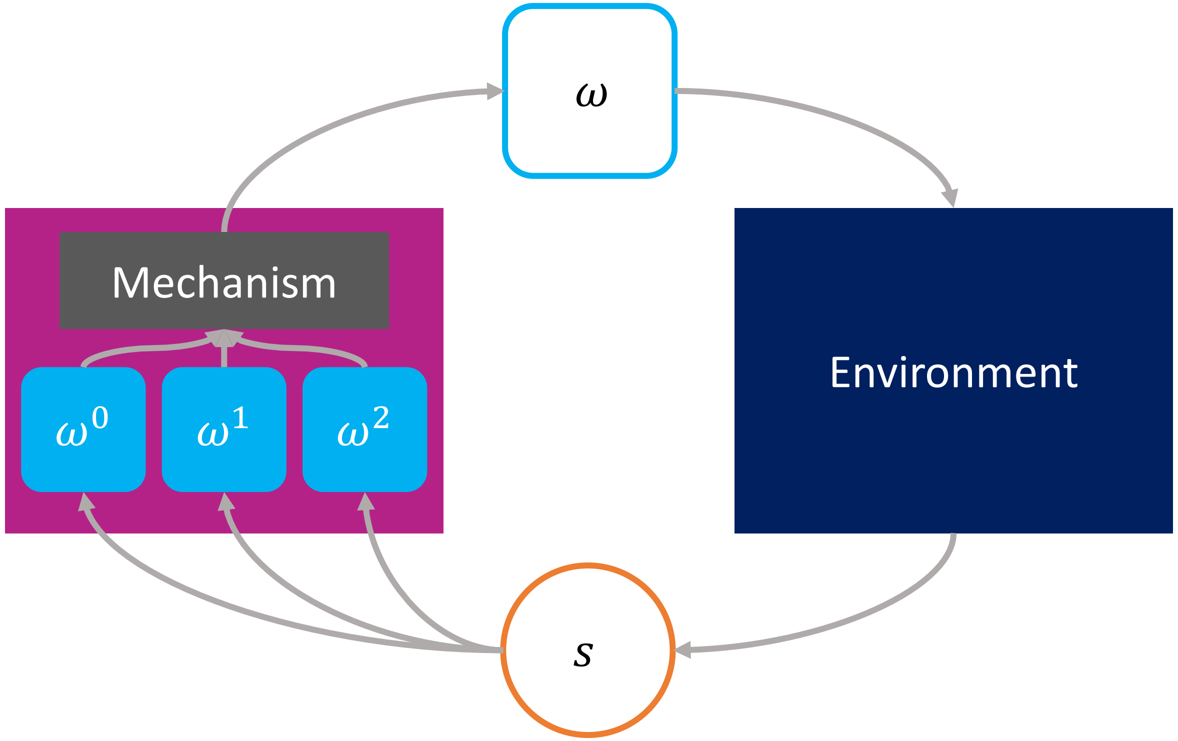
This Paper: Contributions

Assumptions

- Sequential decision making setting
- Each agent produces a specialized transformation to the state (e.g. a literal action)
- Only one agent activates at each time step

Main Contribution

We show that the Vickrey Auction can be adapted to MDPs such that the solution of the global societal objective emerges as a Nash equilibrium strategy profile of the local agents

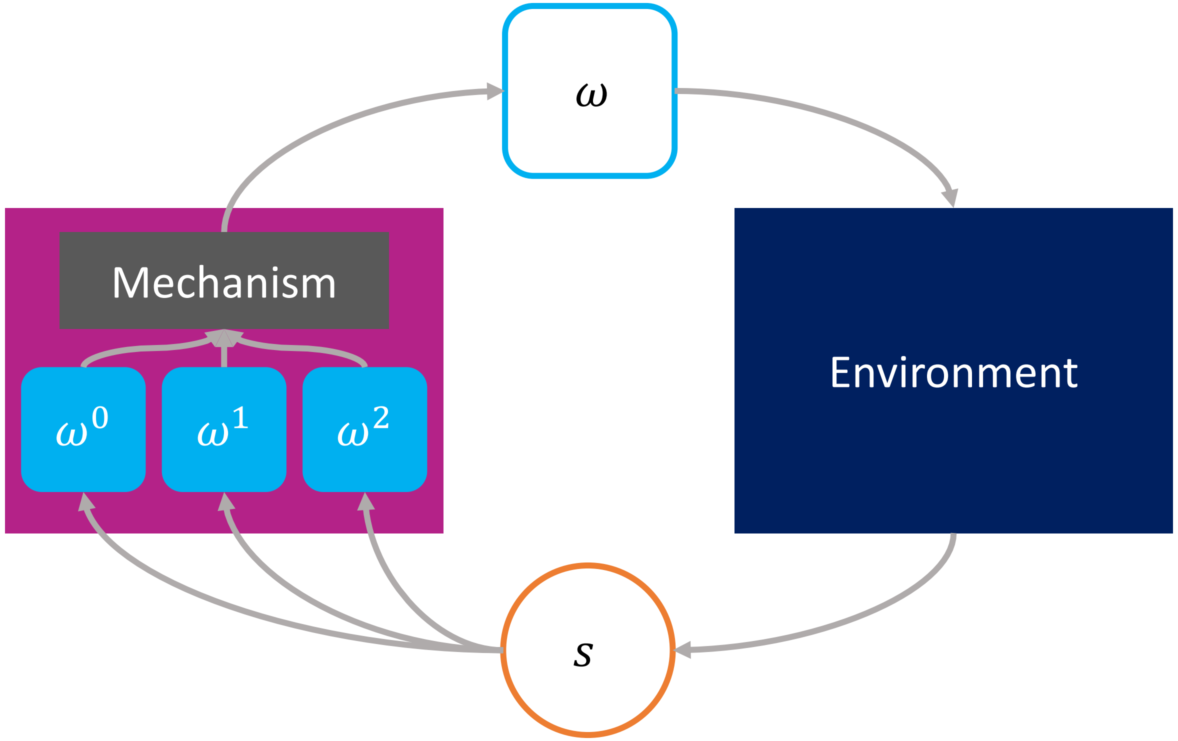


This Paper: Contributions

Assumptions
- Sequential decision making setting
- Each agent produces a specialized transformation to the state (e.g. a literal action)
- Only one agent activates at each time step

Main Contribution
We show that the Vickrey Auction can be adapted to MDPs such that the solution of the global societal objective emerges as a Nash equilibrium strategy profile of the local agents

Implication: Bridging Two Levels of Abstraction
- A recipe for translating a global objective of a society into local learning problems for the agents



This Paper: Contributions

Assumptions

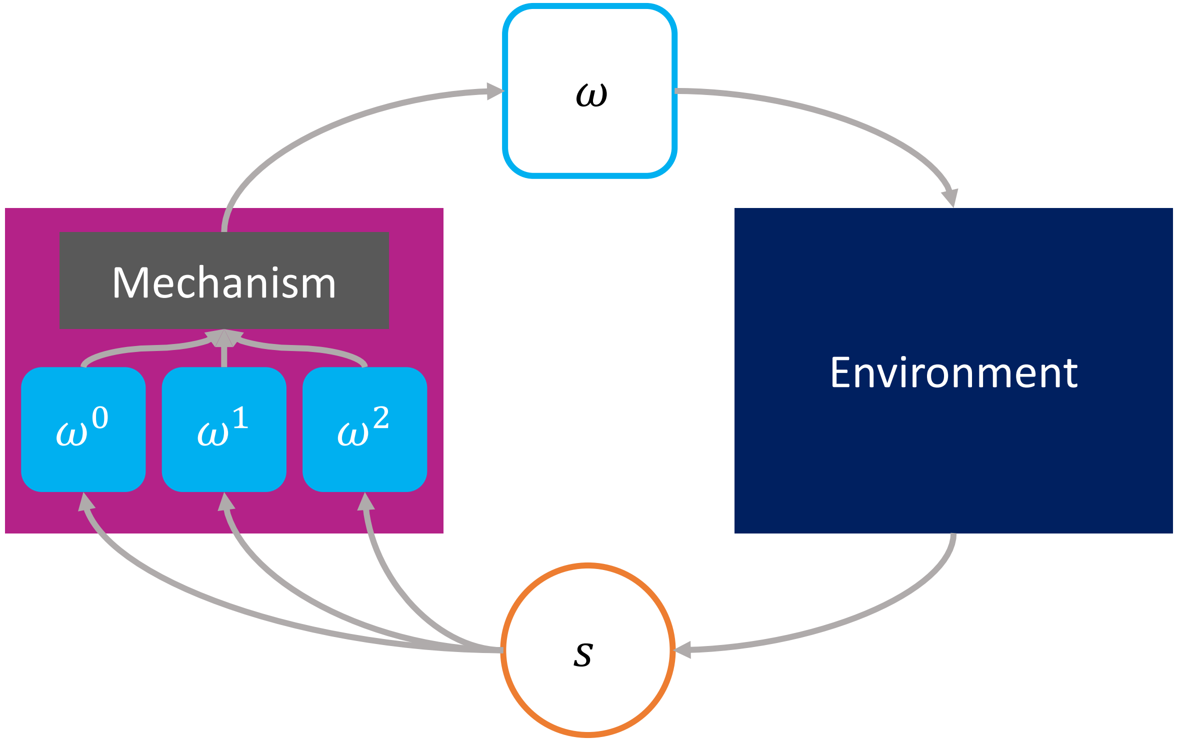
- Sequential decision making setting
- Each agent produces a specialized transformation to the state (e.g. a literal action)
- Only one agent activates at each time step

Main Contribution

We show that the Vickrey Auction can be adapted to MDPs such that the solution of the global societal objective emerges as a Nash equilibrium strategy profile of the local agents

Implication: Bridging Two Levels of Abstraction

- A recipe for translating a global objective of a society into local learning problems for the agents
- A decentralized reinforcement learning algorithm with credit assignment local in space and time



Roadmap

Question	Key Idea

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	
For what auction mechanism would these optimal bids be an equilibrium strategy?	

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	

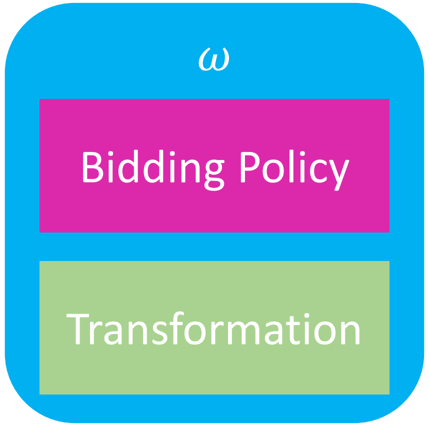
Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

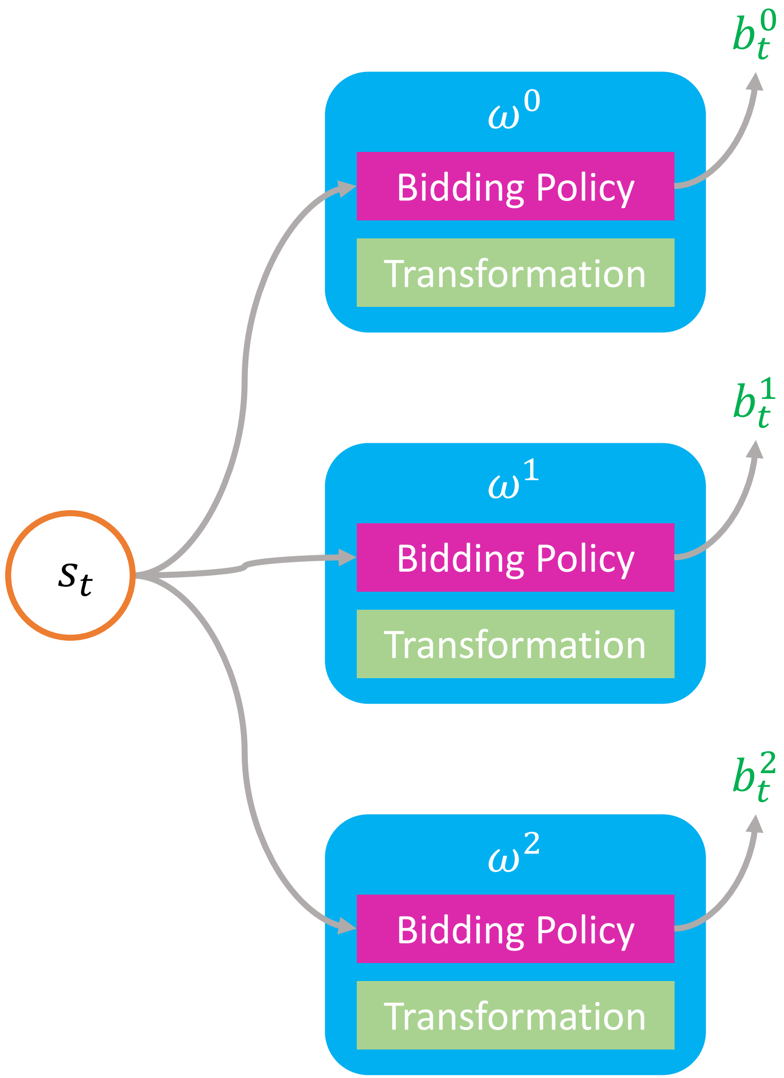
Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

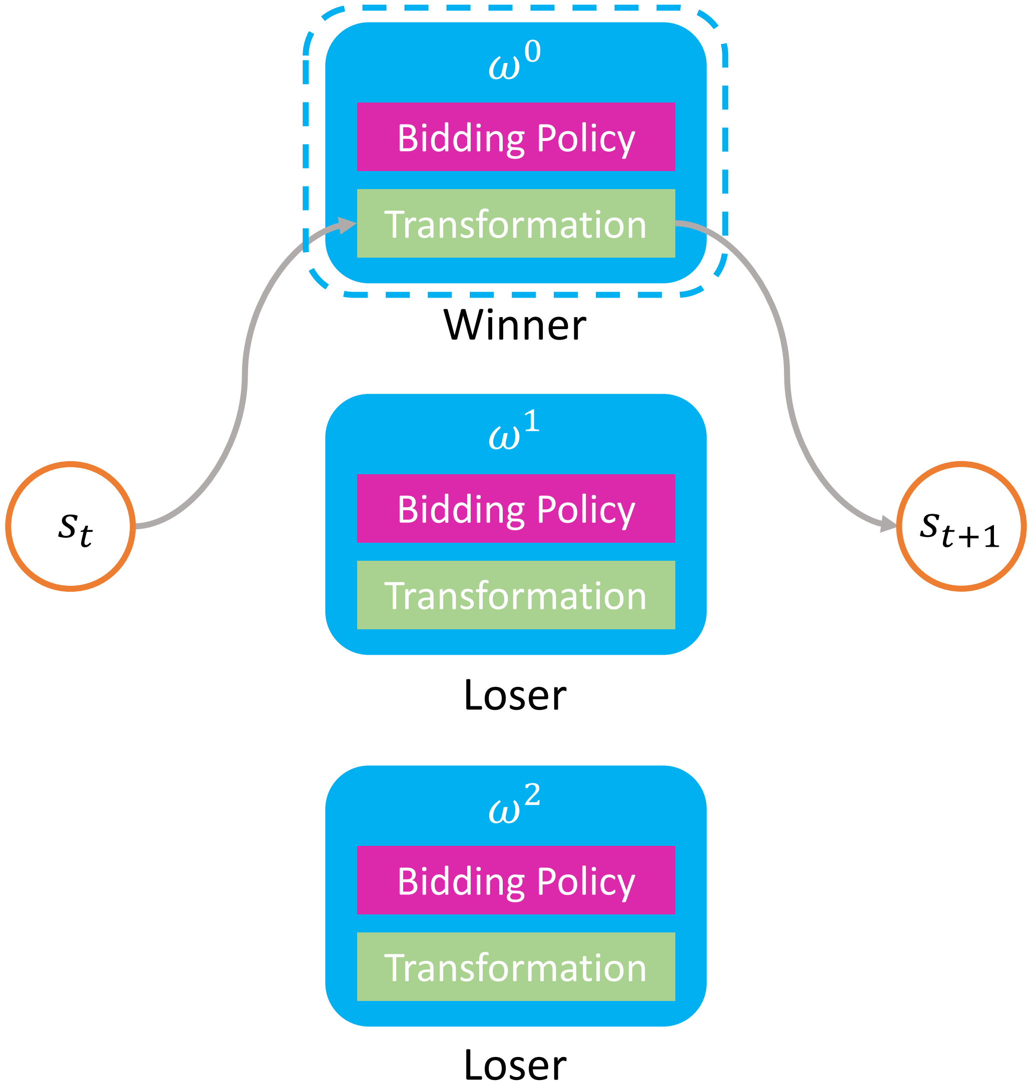
Architecture of an Agent



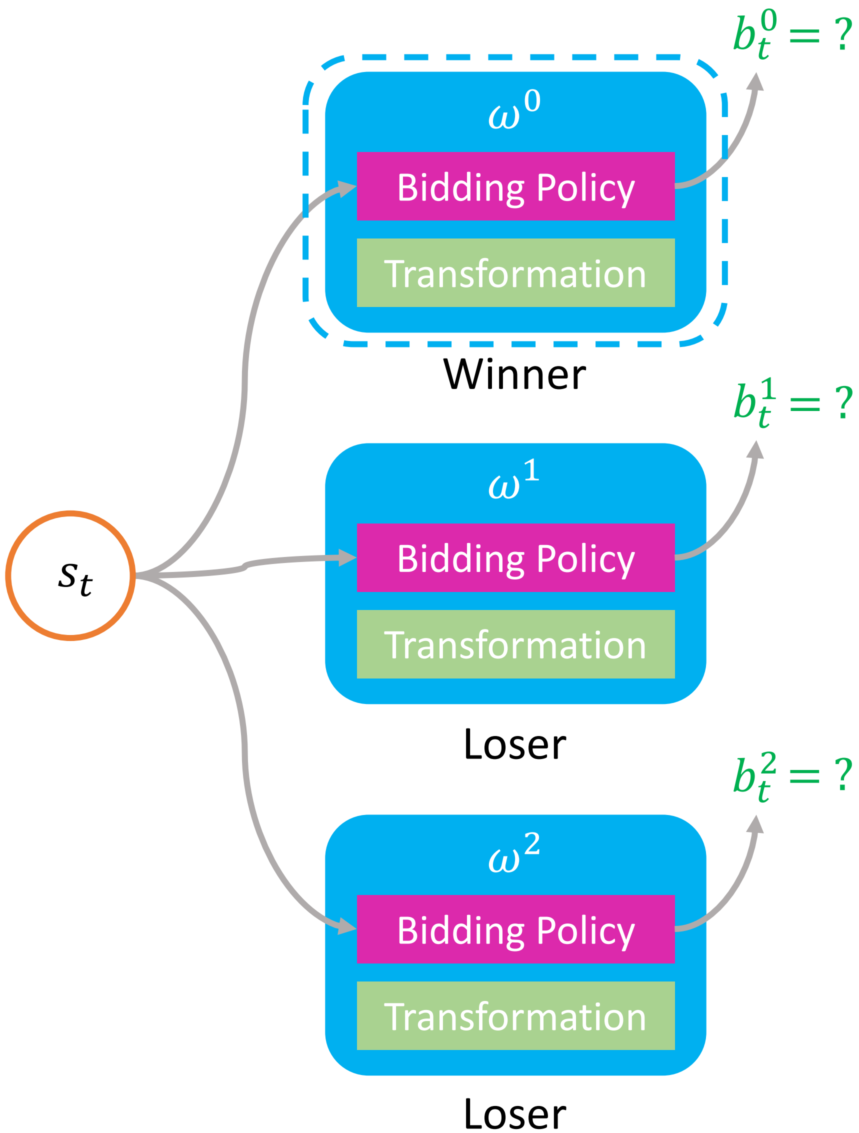
Activating Agents via Auction

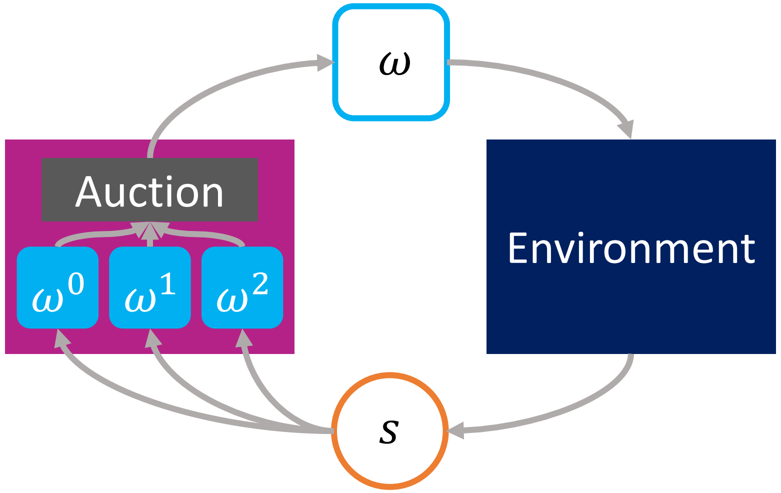


Transforming the State

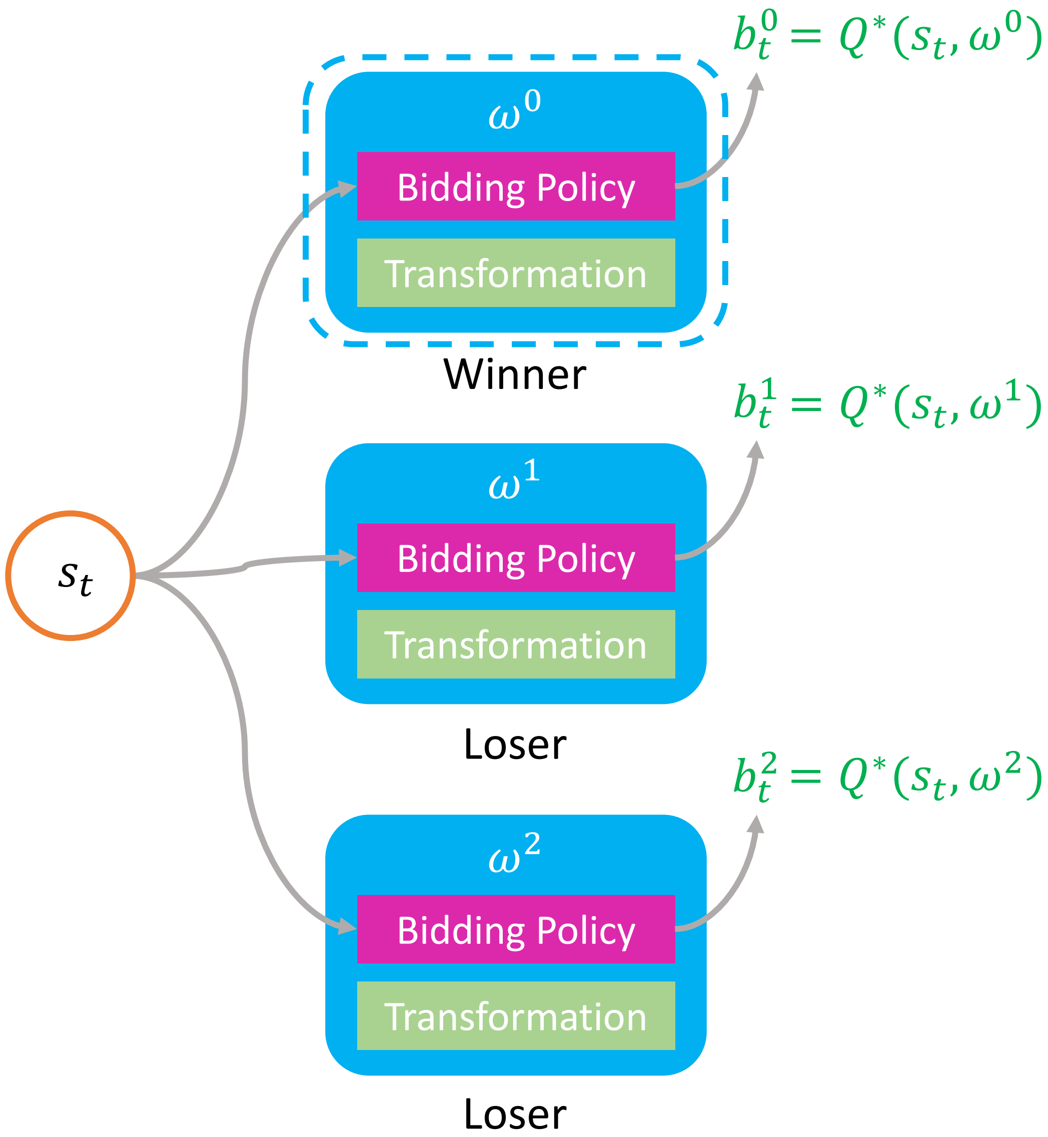


What should the optimal bids be?

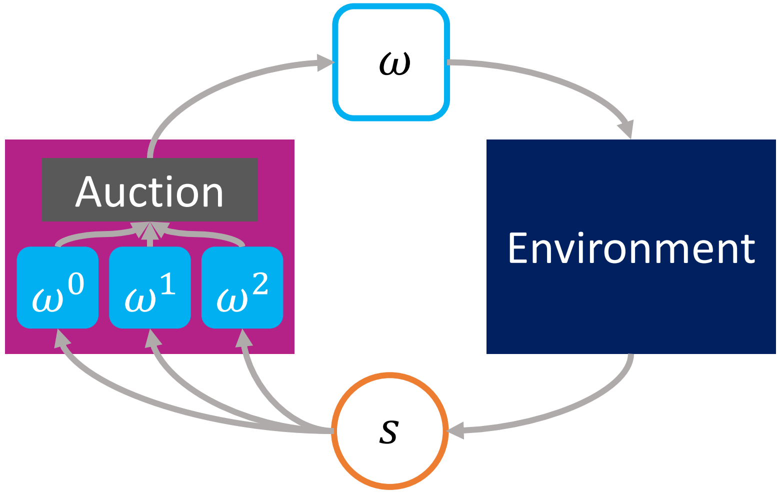




Key Idea: the optimal bid is your optimal Q value



$$\omega^* = \underset{k}{\operatorname{argmax}}\, Q^*(s_t, \omega^k)$$



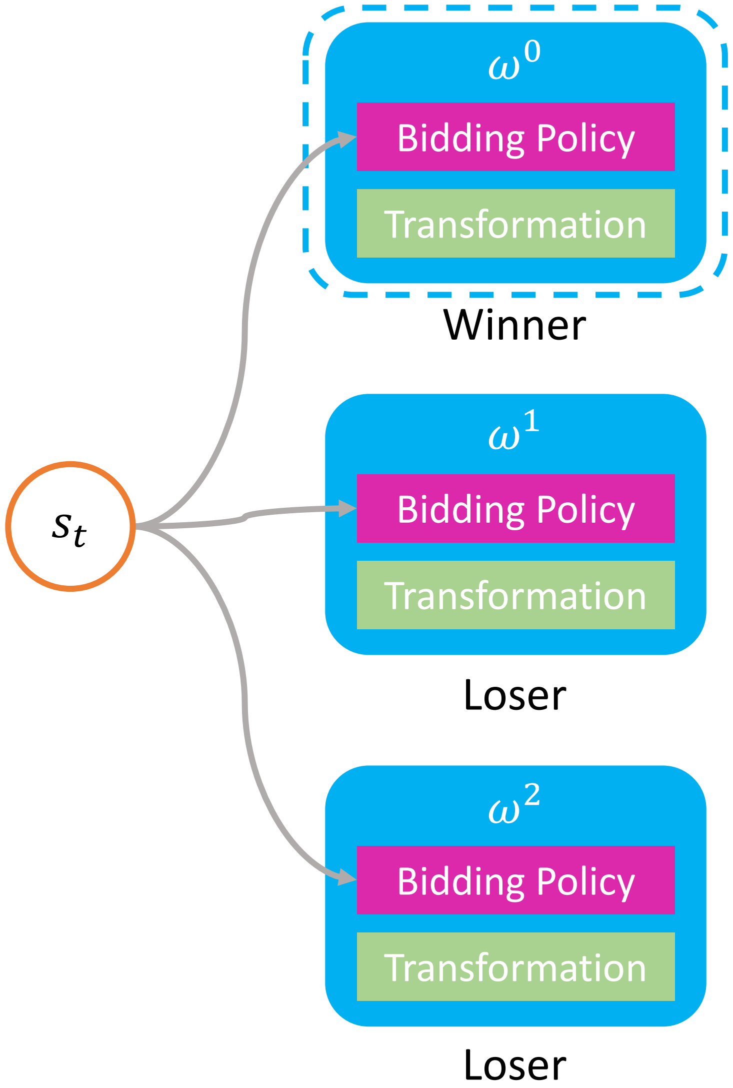
Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

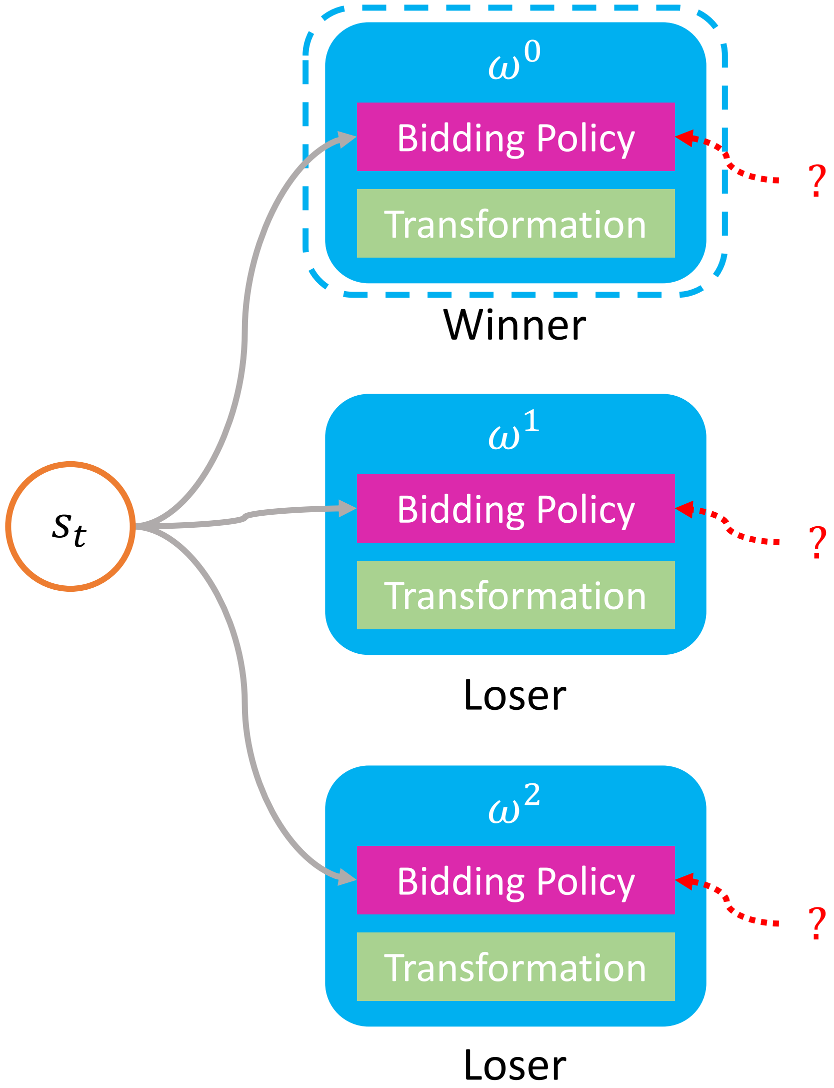
What Should the Auction Mechanism be?



Assume

Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

What Should the Auction Mechanism be?



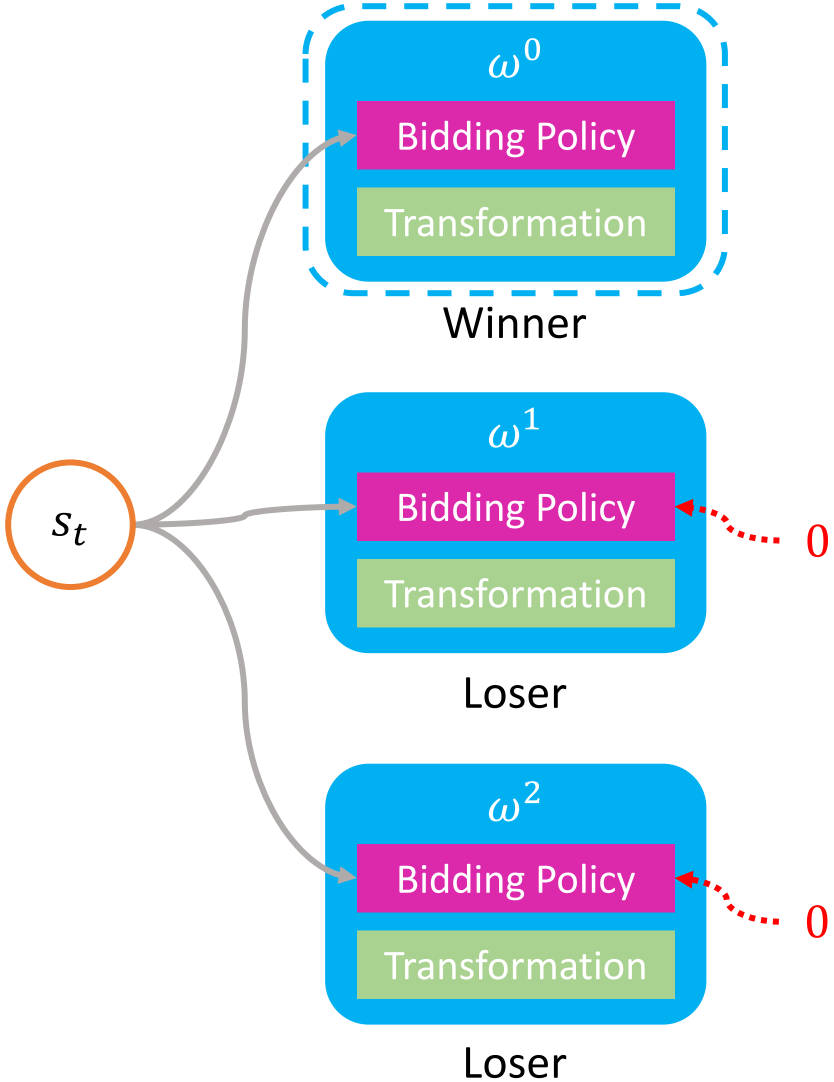
Assume

Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question

What should the agents’ utilities be?

What Should the Auction Mechanism be?

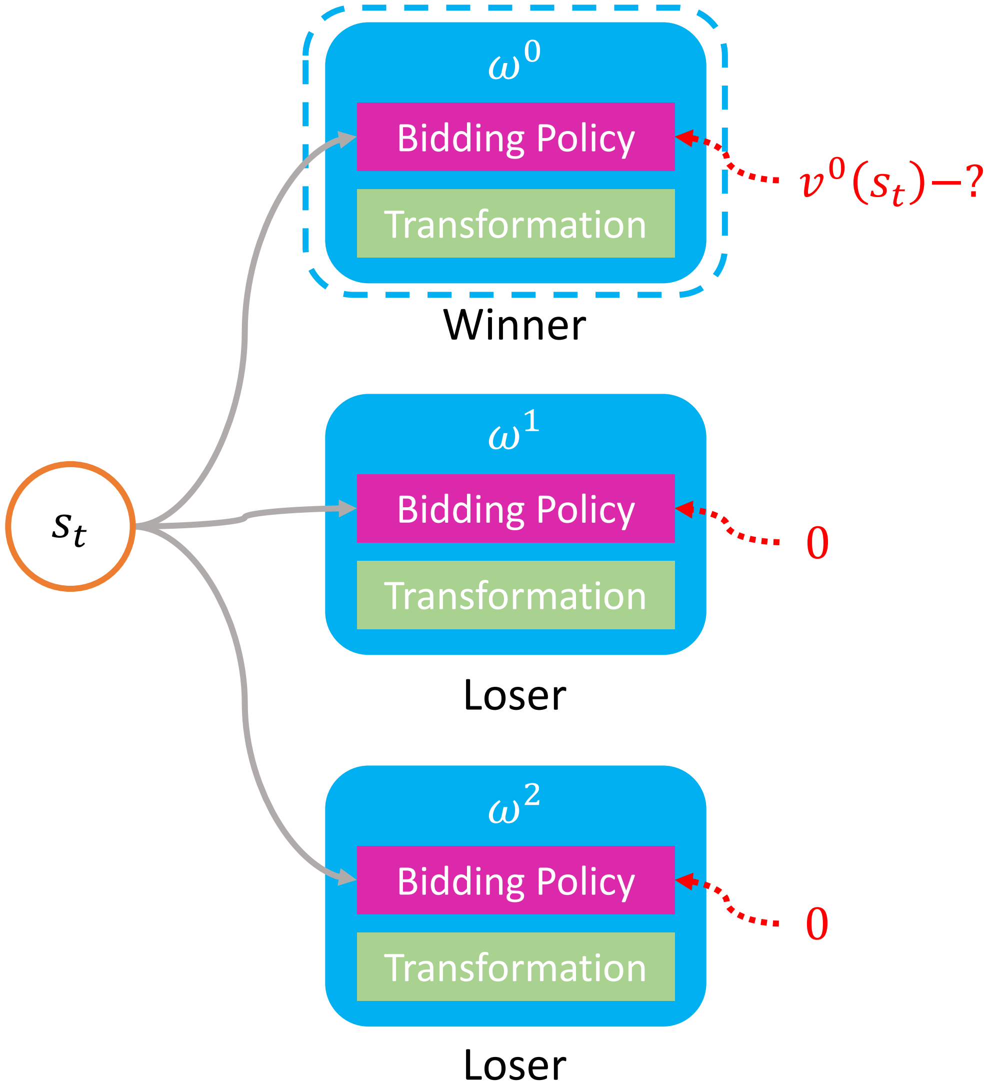


Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question
What should the agents' utilities be?

Utilities?
Losers: $u^i(b) = 0$

What Should the Auction Mechanism be?



Assume

Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question

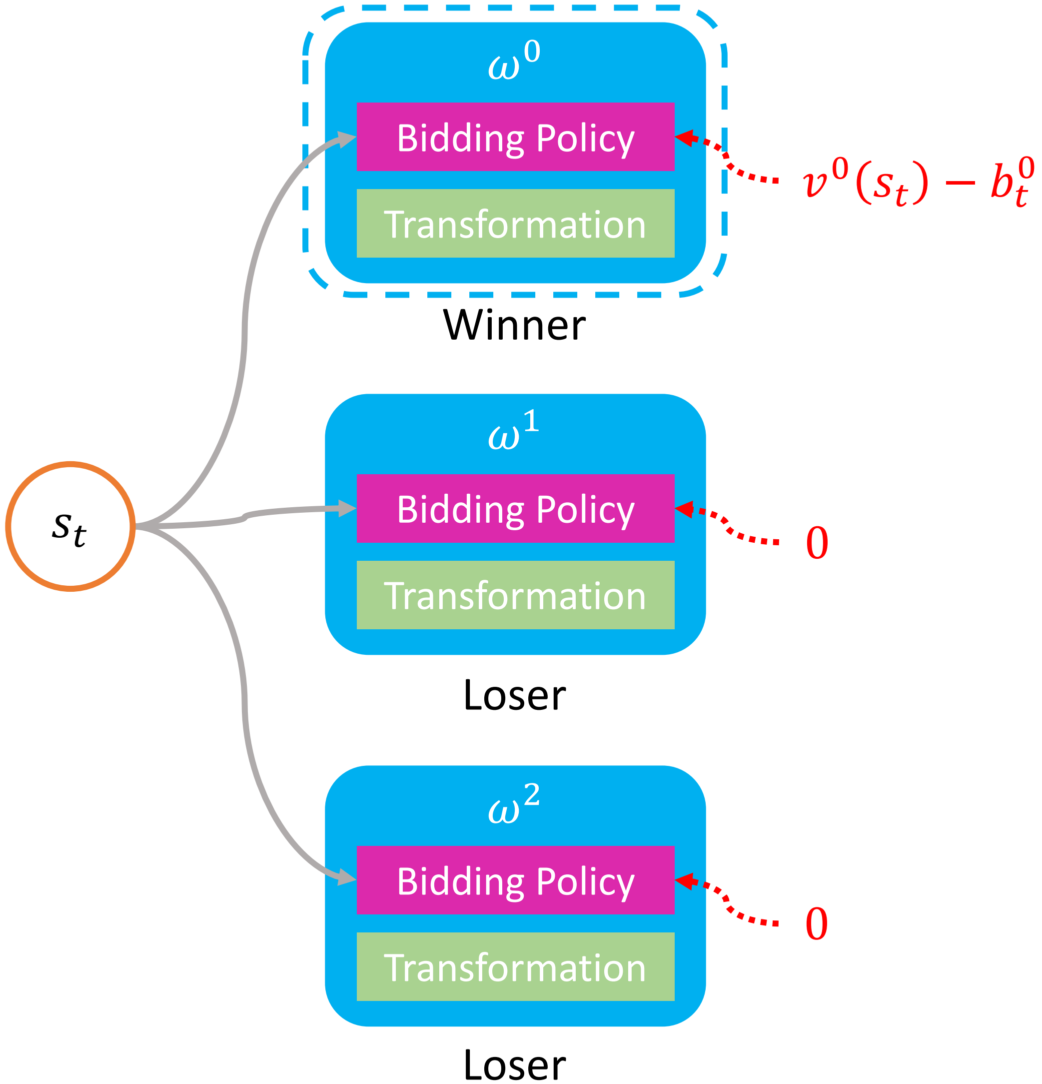
What should the agents' utilities be?

Utilities?

Losers: $u^i(b) = 0$

Winner: $u^i(b) = v^i - ?$

What Should the Auction Mechanism be?

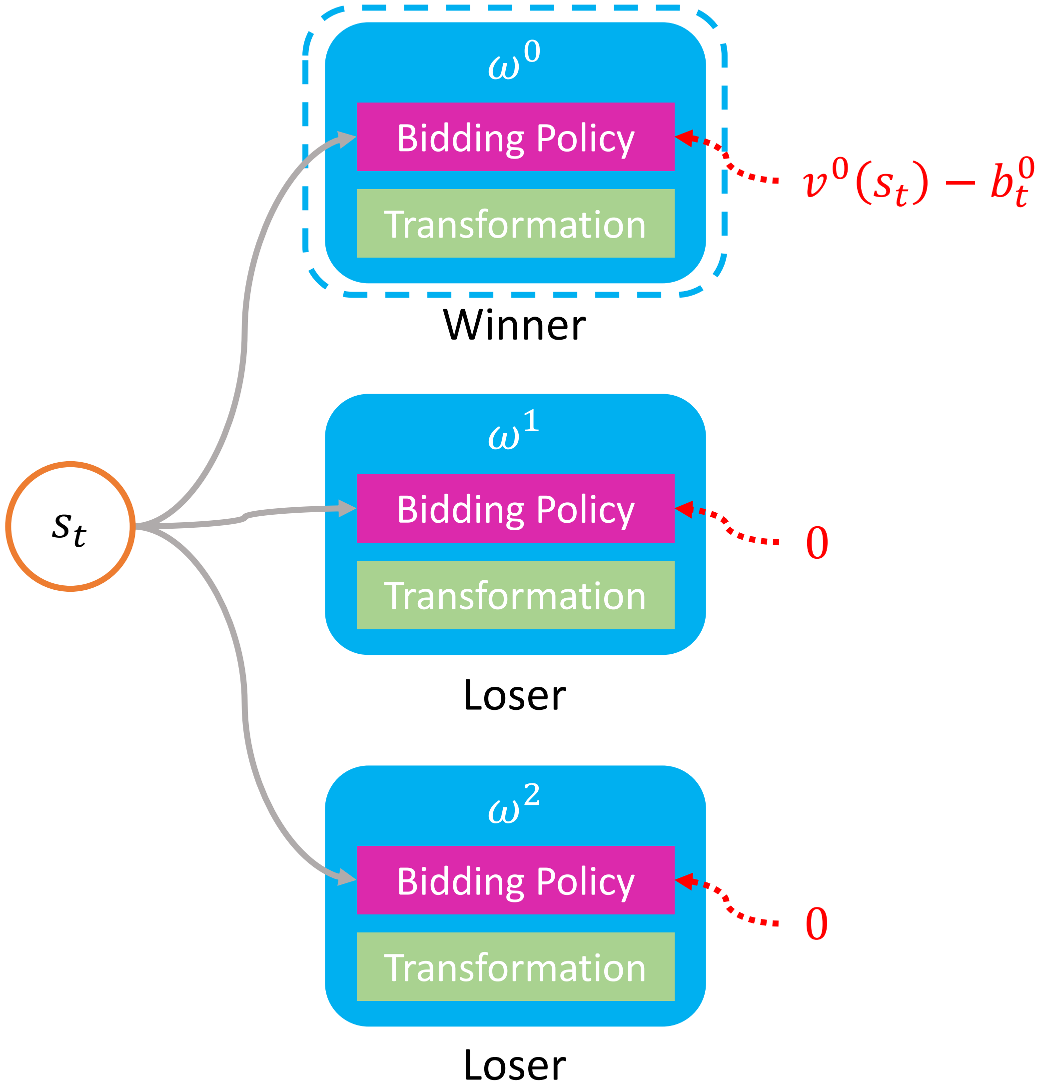


Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question
What should the agents' utilities be?

First Price Sealed-Bid Auction Utilities?
Losers: $u^i(b) = 0$
Winner: $u^i(b) = v^i - b$

What Should the Auction Mechanism be?



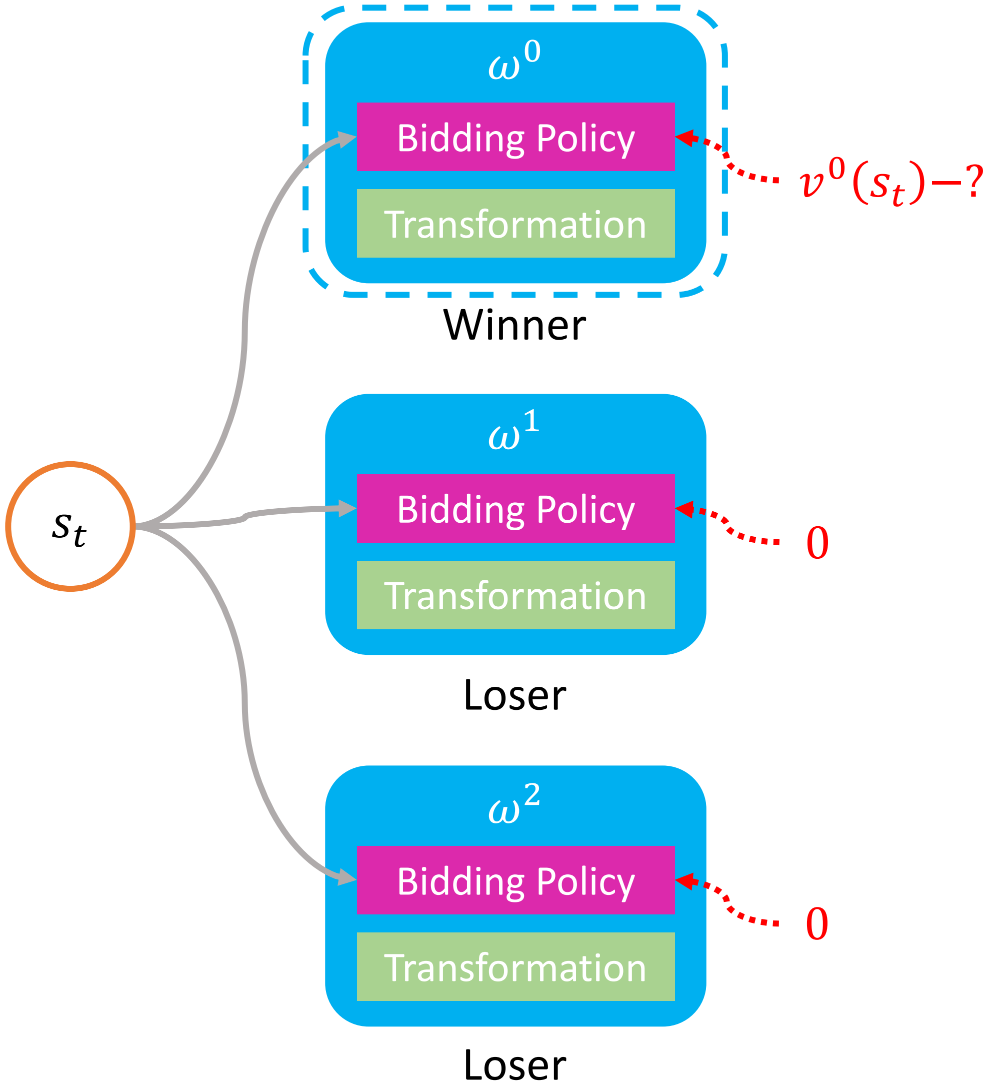
Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question
What should the agents' utilities be?

First Price Sealed-Bid Auction Utilities?
Losers: $u^i(b) = 0$
Winner: $u^i(b) = v^i - b$

Problem with First Price Sealed-Bid Auctions
There is no dominant strategy – the bid that optimizes an agent's utility depends on what other agents bid

What Should the Auction Mechanism be?



Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

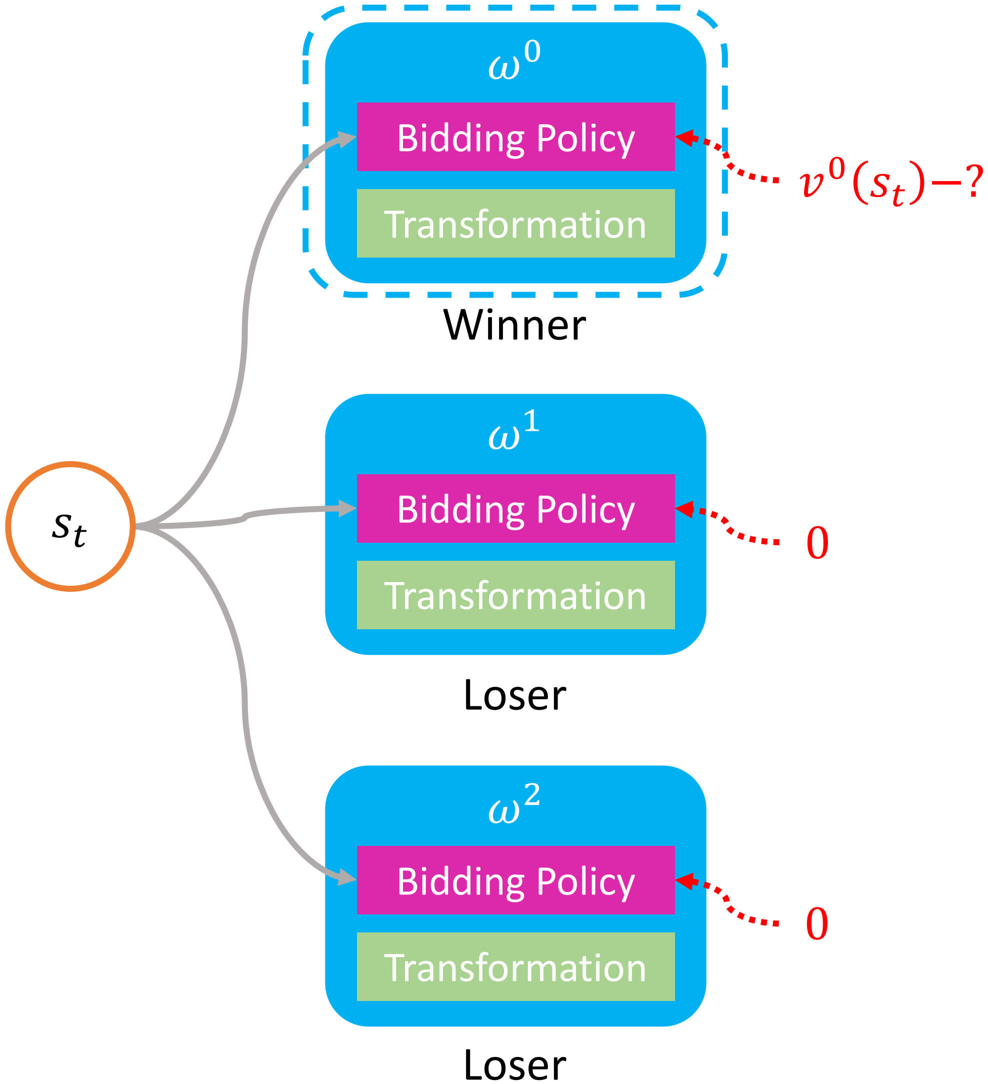
Question
What should the agents' utilities be?

Utilities?
Losers: $u^i(b) = 0$
Winner: $u^i(b) = v^i - ?$

Want: Dominant Strategy Incentive Compatibility
The optimal strategy is to truthfully bid its own valuation:

$$b^i \leftarrow v^i$$

What Should the Auction Mechanism be?



Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question
What should the agents' utilities be?

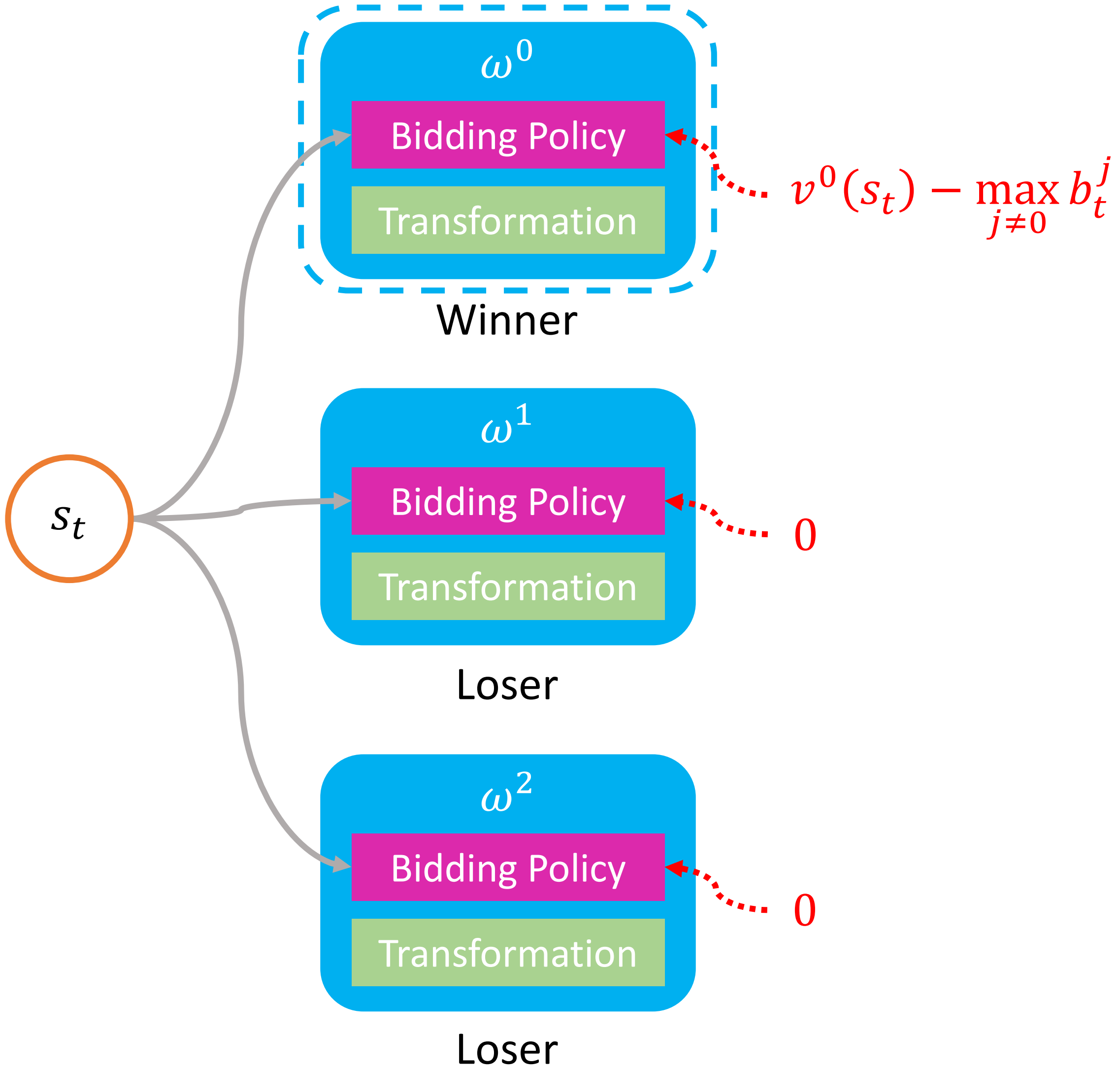
Utilities?
Losers: $u^i(b) = 0$
Winner: $u^i(b) = v^i - ?$

Want: Dominant Strategy Incentive Compatibility
The optimal strategy is to truthfully bid its own valuation:

$$b^i \leftarrow v^i$$

Implication: Set $v^k(s_t) = Q^*(s_t, \omega^k)$!

Vickrey Auction



Assume
Each agent ω^k has a valuation $v^k(s_t)$ for state s_t

Question
What should the agents' utilities be?

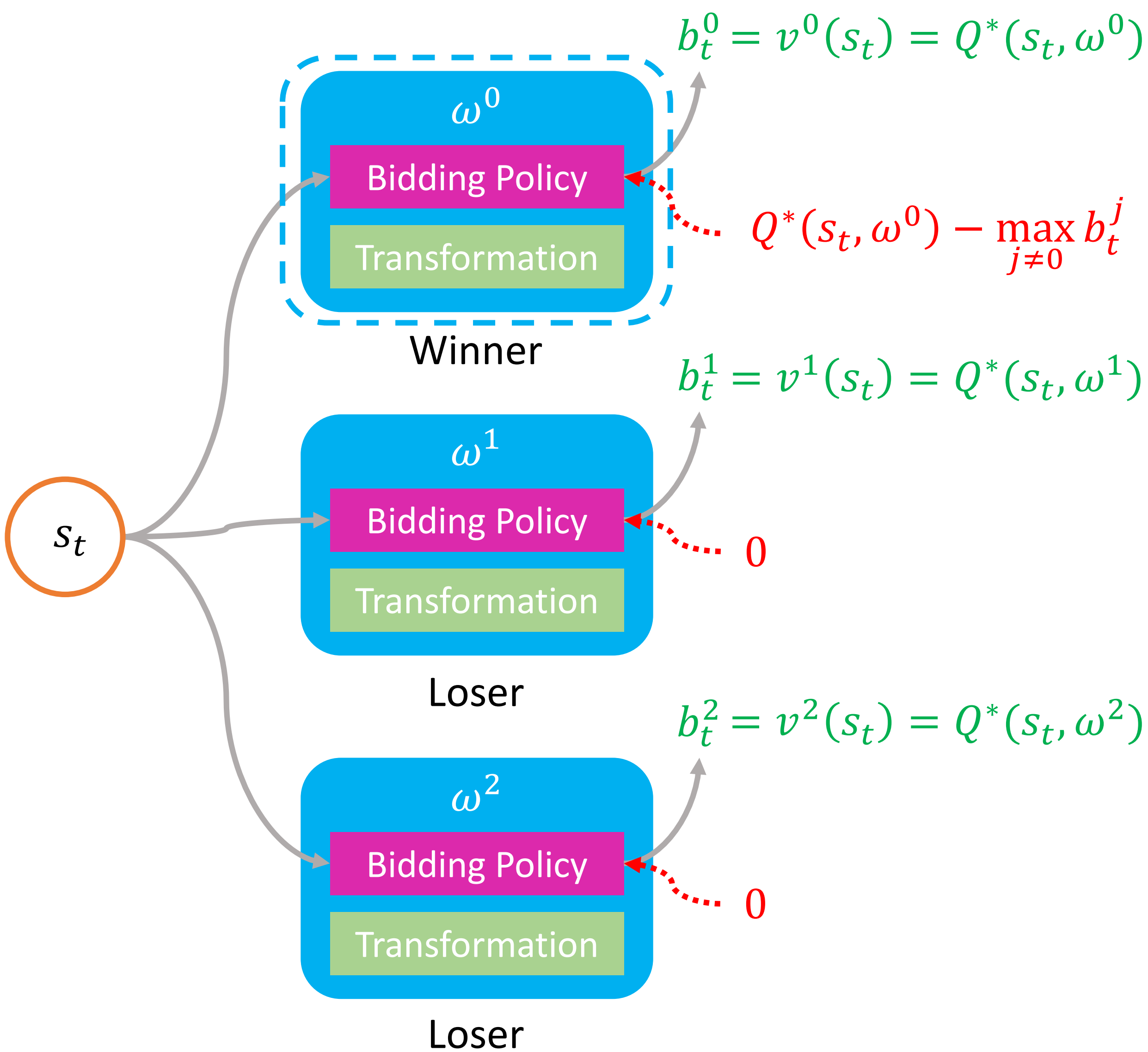
Vickrey Auction Utilities!
Losers: $u^i(b) = 0$
Winner: $u^i(b) = v^i - \max_{j \neq i} b^j$

Want: Dominant Strategy Incentive Compatibility
The optimal strategy is to truthfully bid its own valuation:

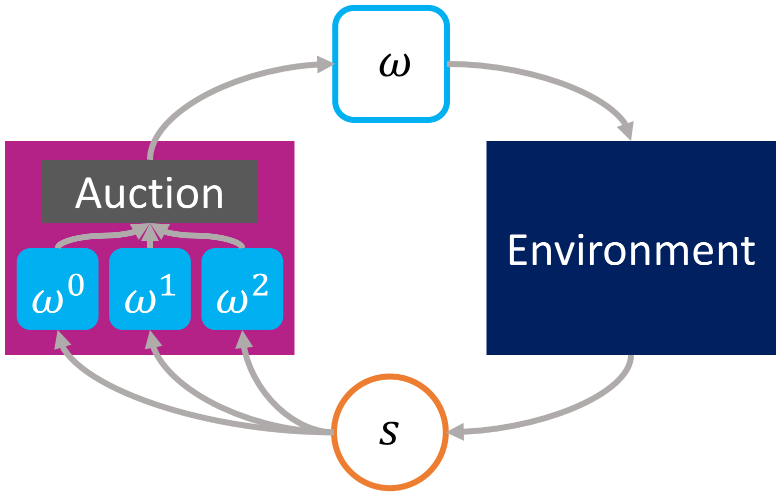
$$b^i \leftarrow v^i$$

Implication: Set $v^k(s_t) = Q^*\left(s_t, \omega^k\right)$!

A Recipe for Relating Local and Global Objectives



$$\omega^* = \operatorname*{argmax}_k Q^*\left(s_t, \omega^k\right)$$



Implication: Set $v^k(s_t) = Q^*\left(s_t, \omega^k\right)$!

Roadmap

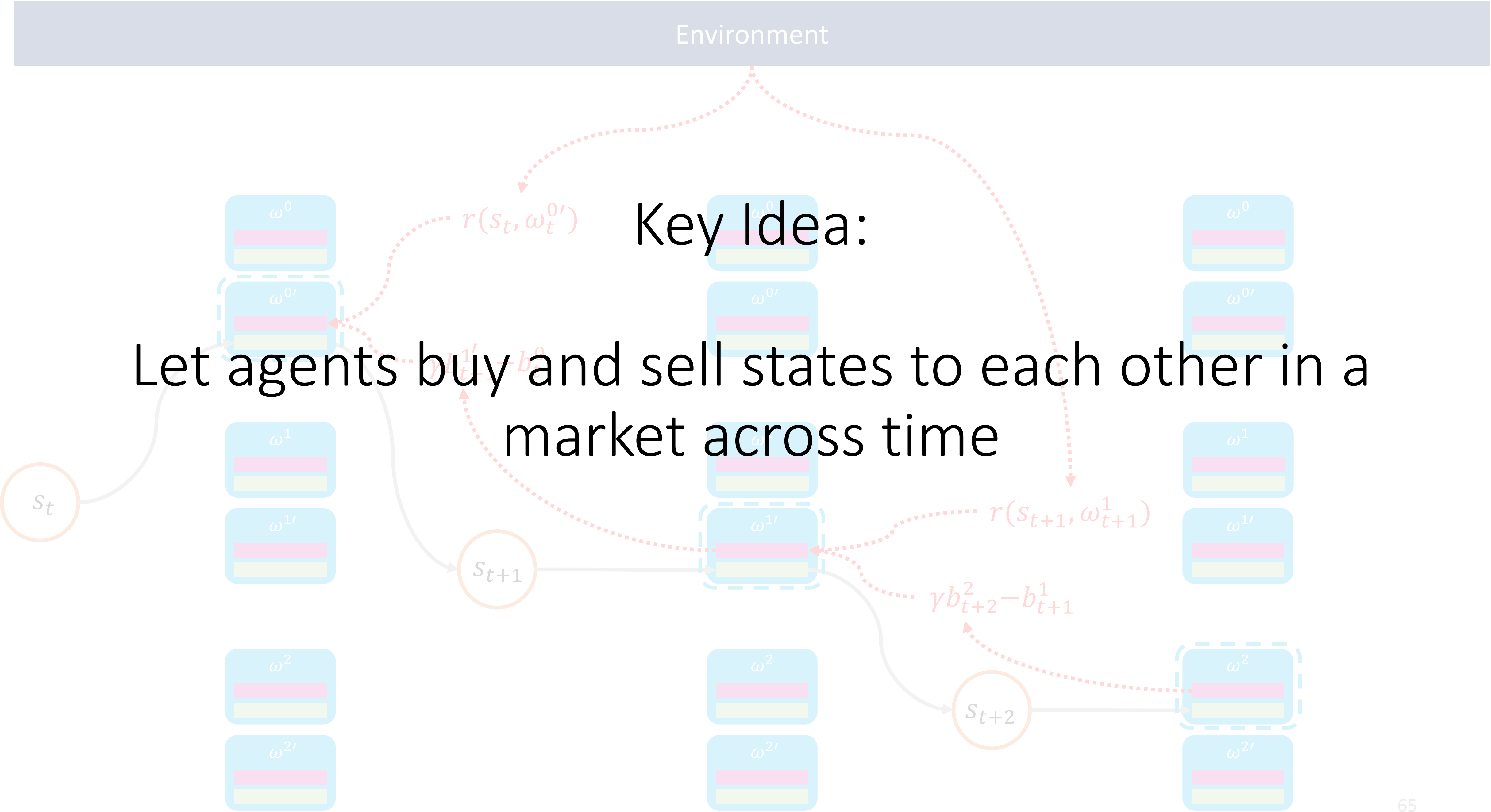
Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q}^{*}(\boldsymbol{s_t}, \boldsymbol{\omega^i})$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^{*}(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^{*}(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

But wait…

Optimal Q values are usually unknown!

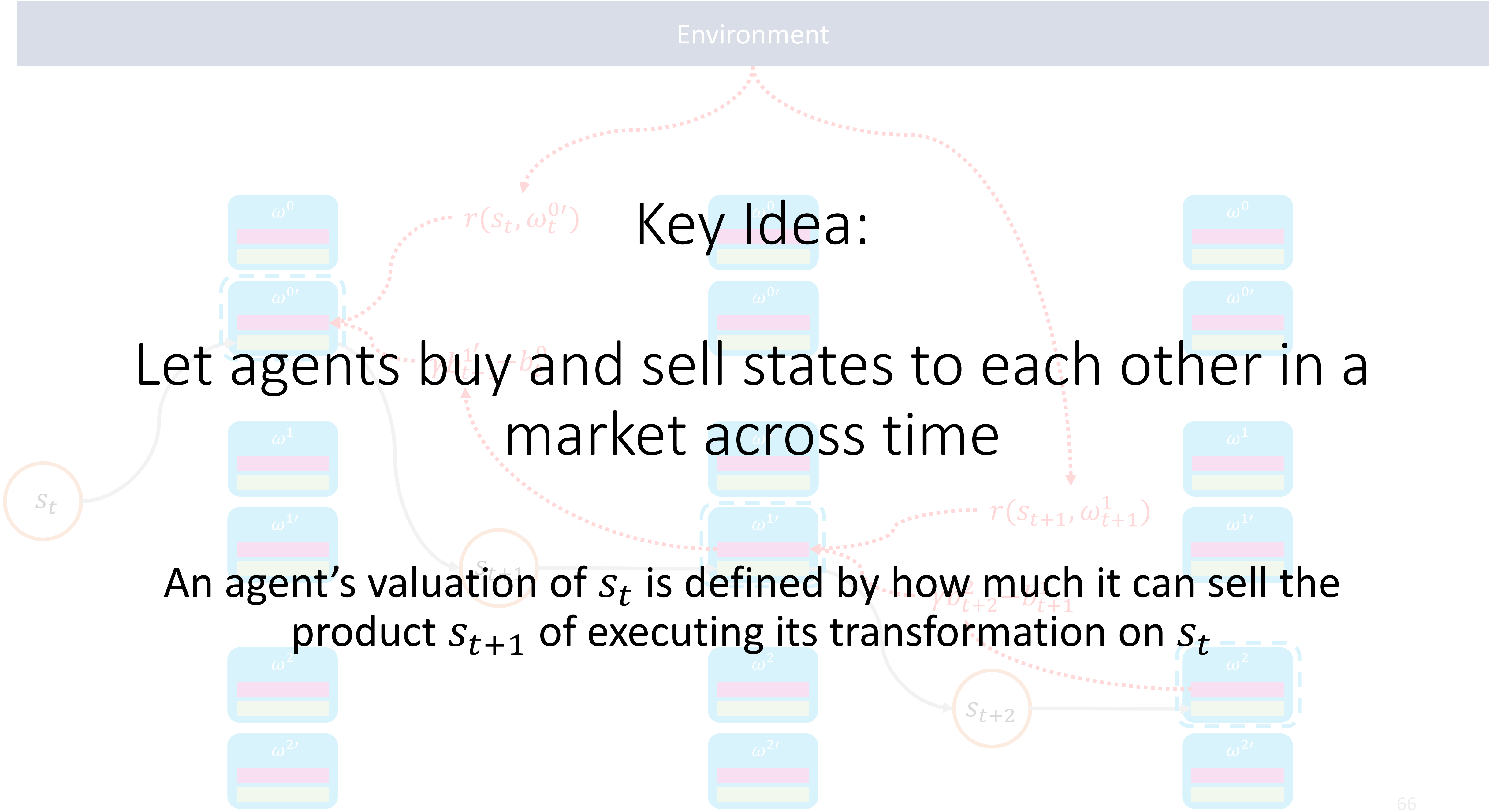
Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$** for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	



Key Idea:

Let agents buy and sell states to each other in a market across time

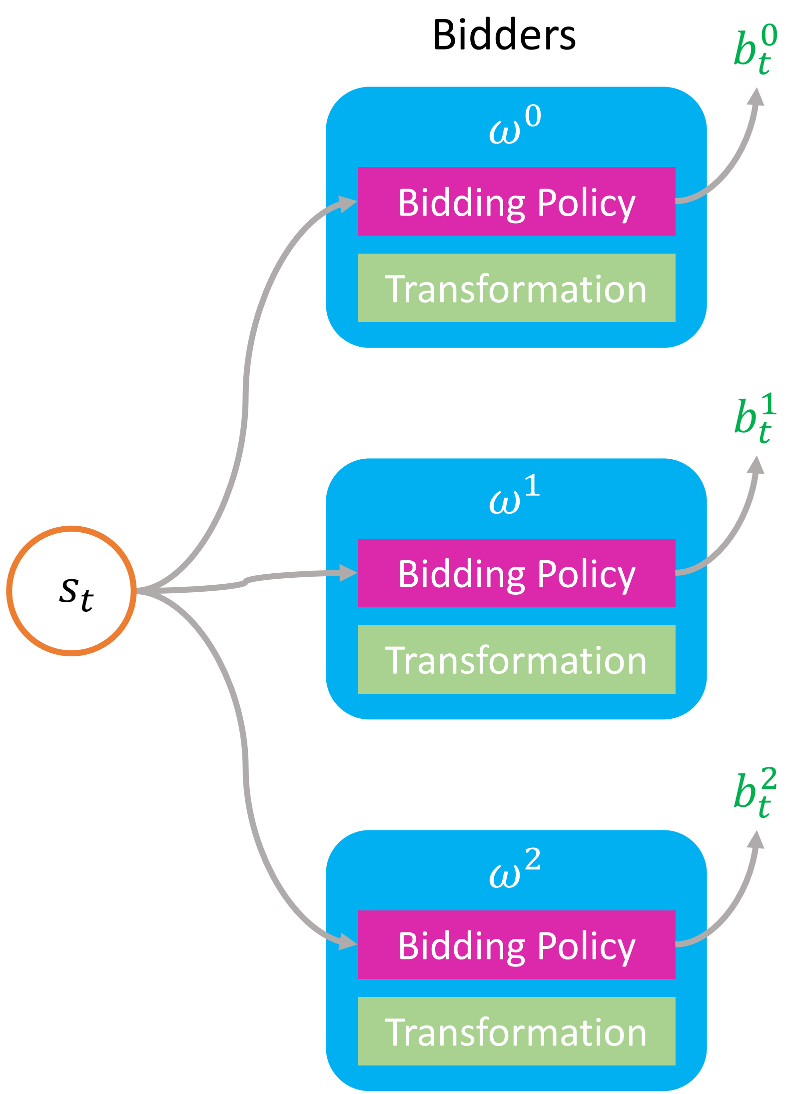


Key Idea:

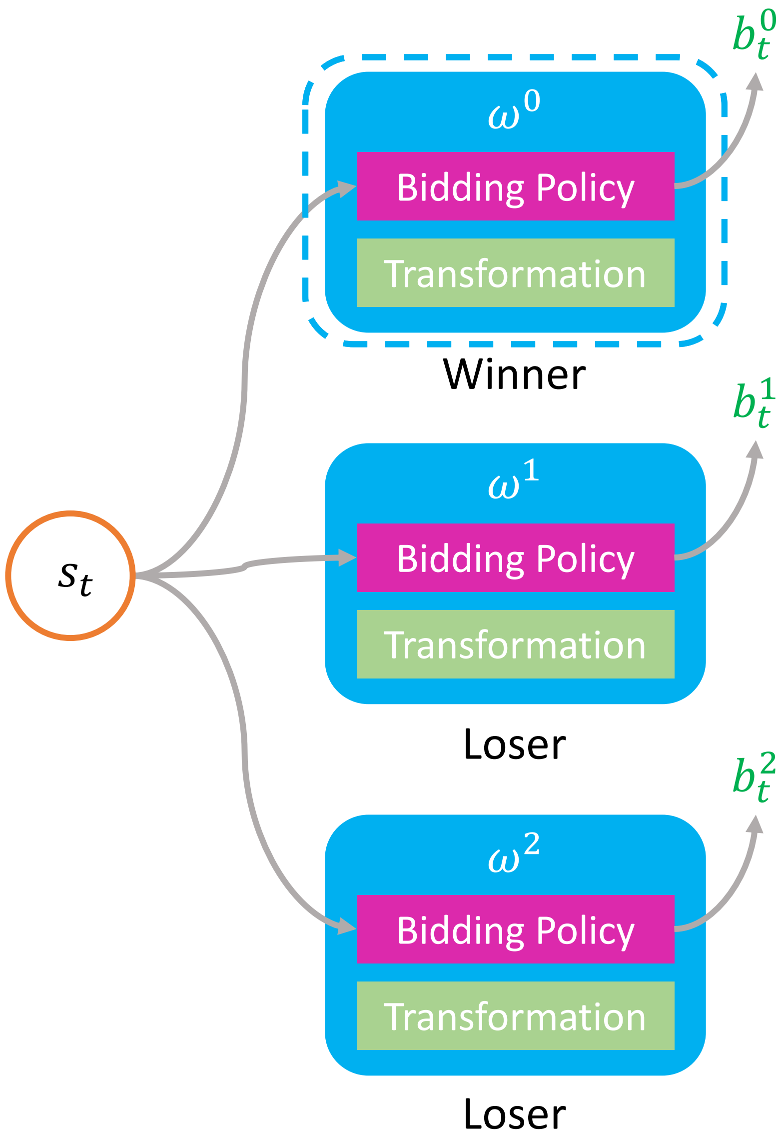
Let agents buy and sell states to each other in a market across time

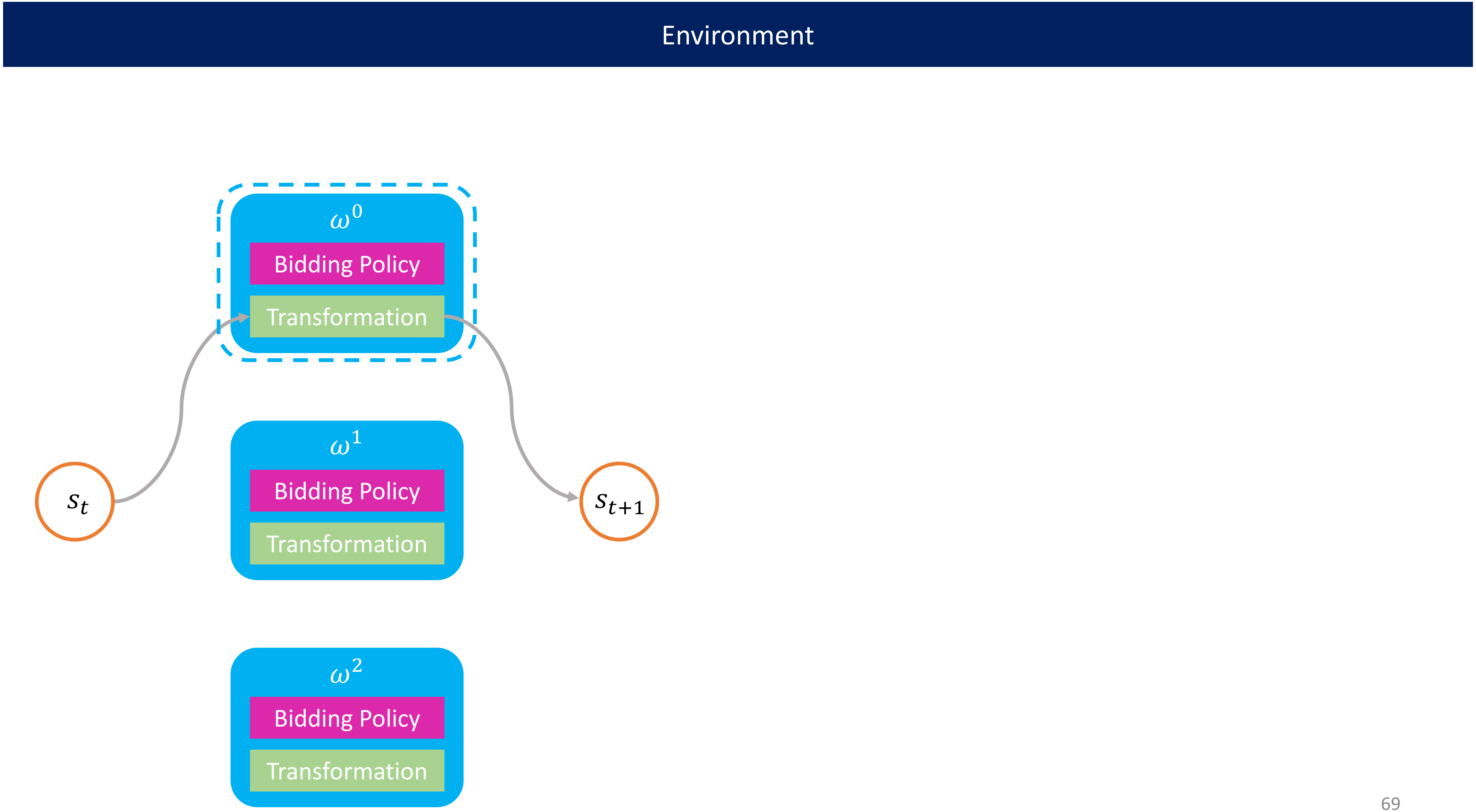
An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t

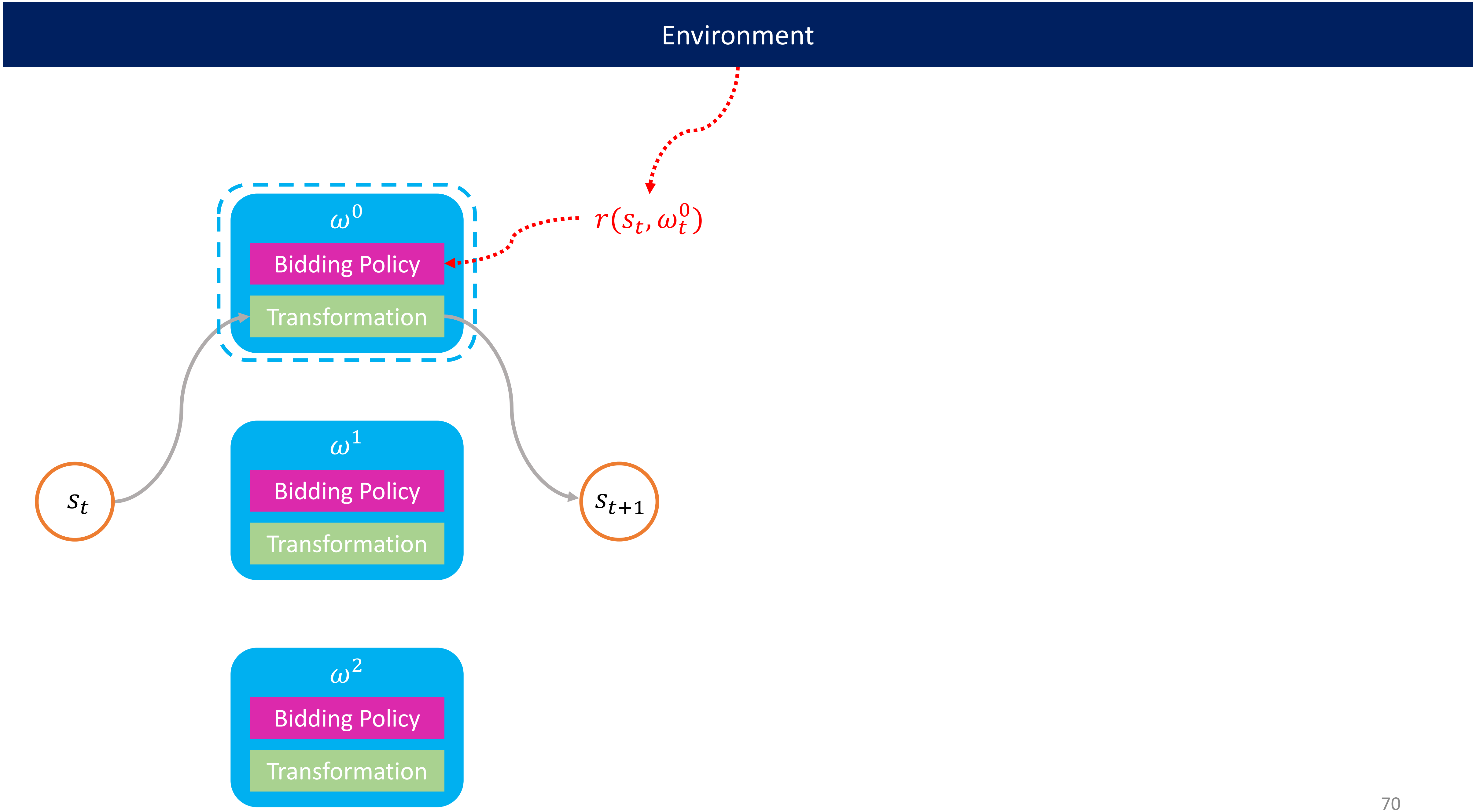
Environment

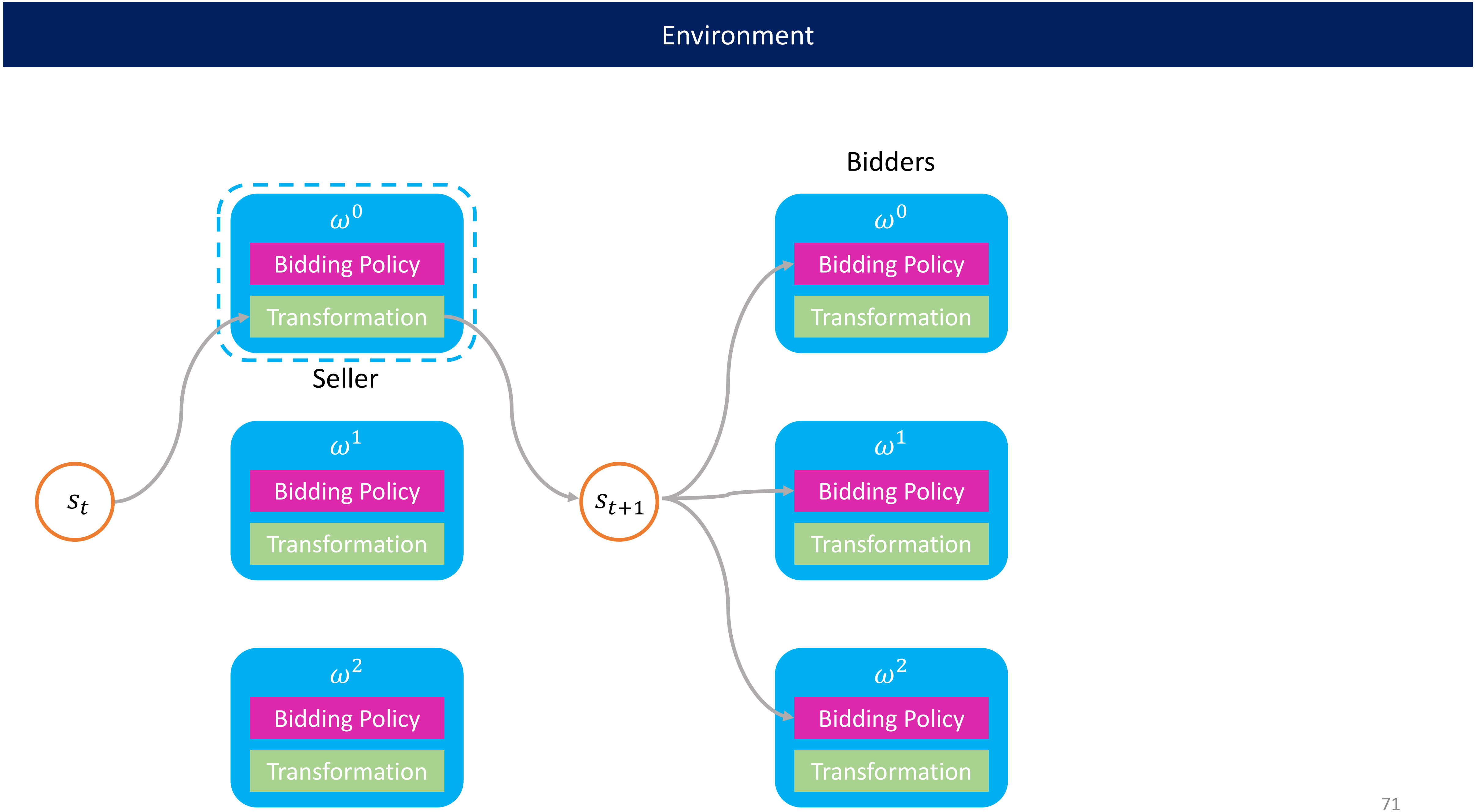


Environment



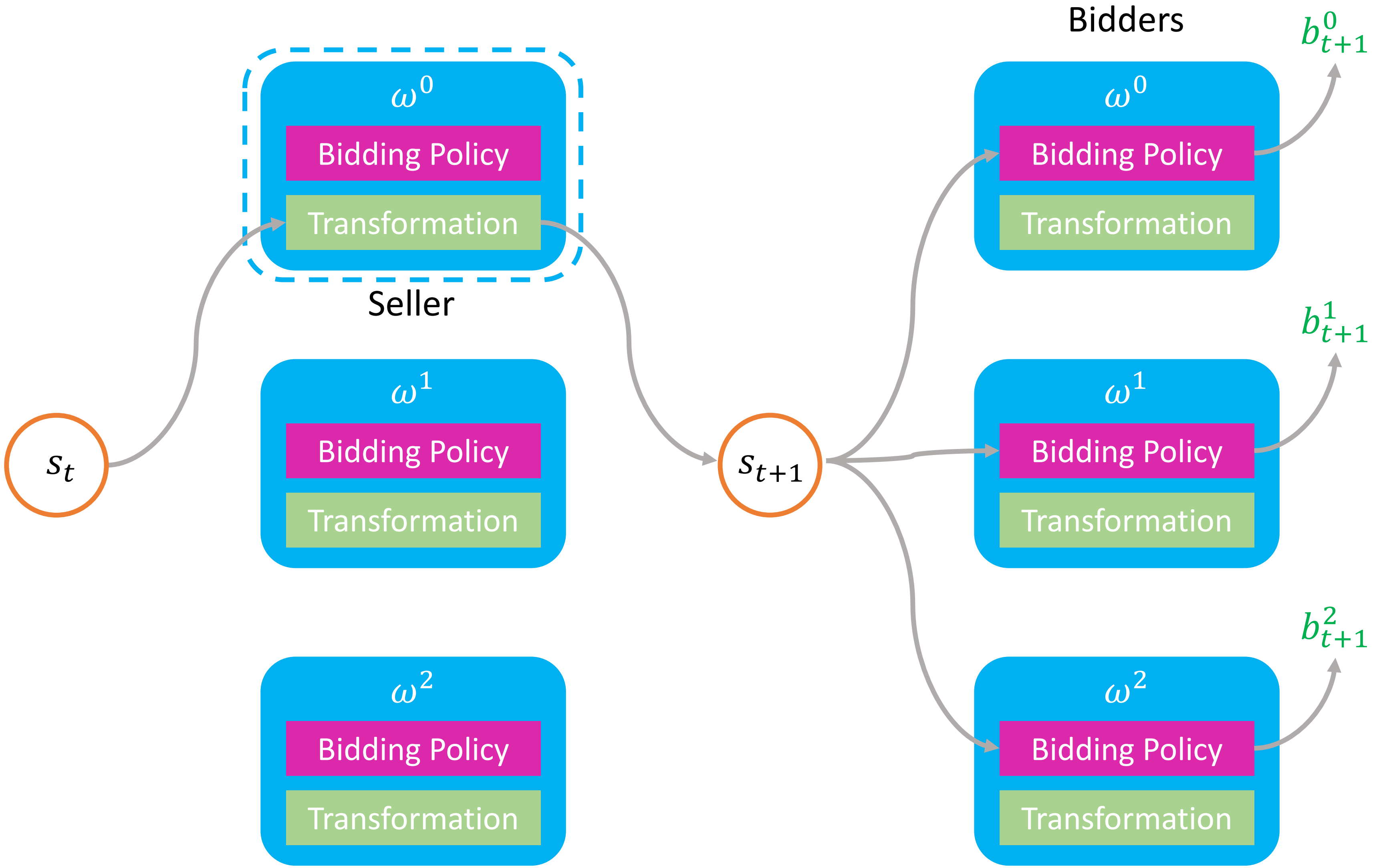

Environment
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation
s_t
s_{t+1}


Environment
$r(s_t, \omega_t^0)$
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation
s_t
s_{t+1}

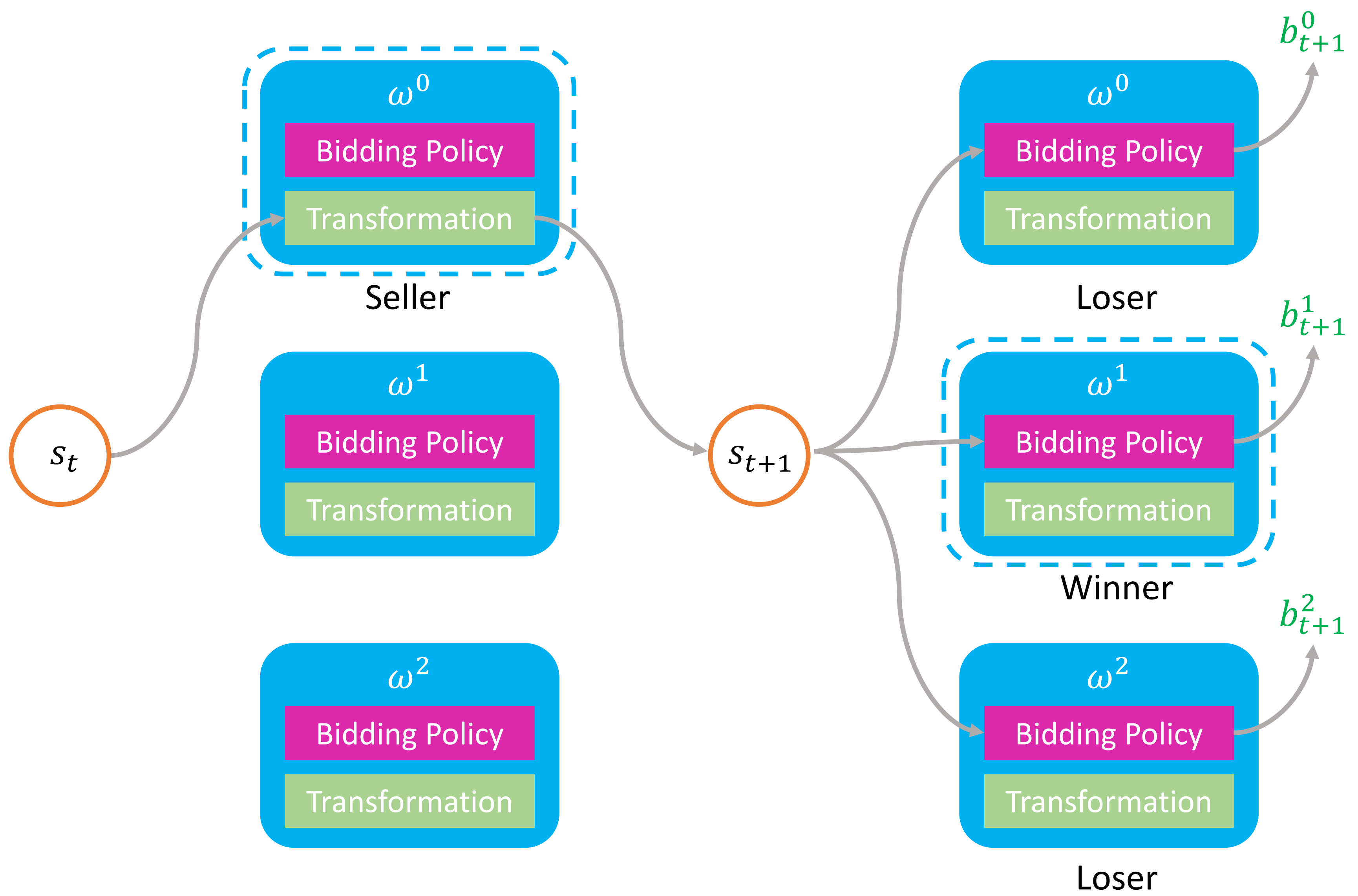


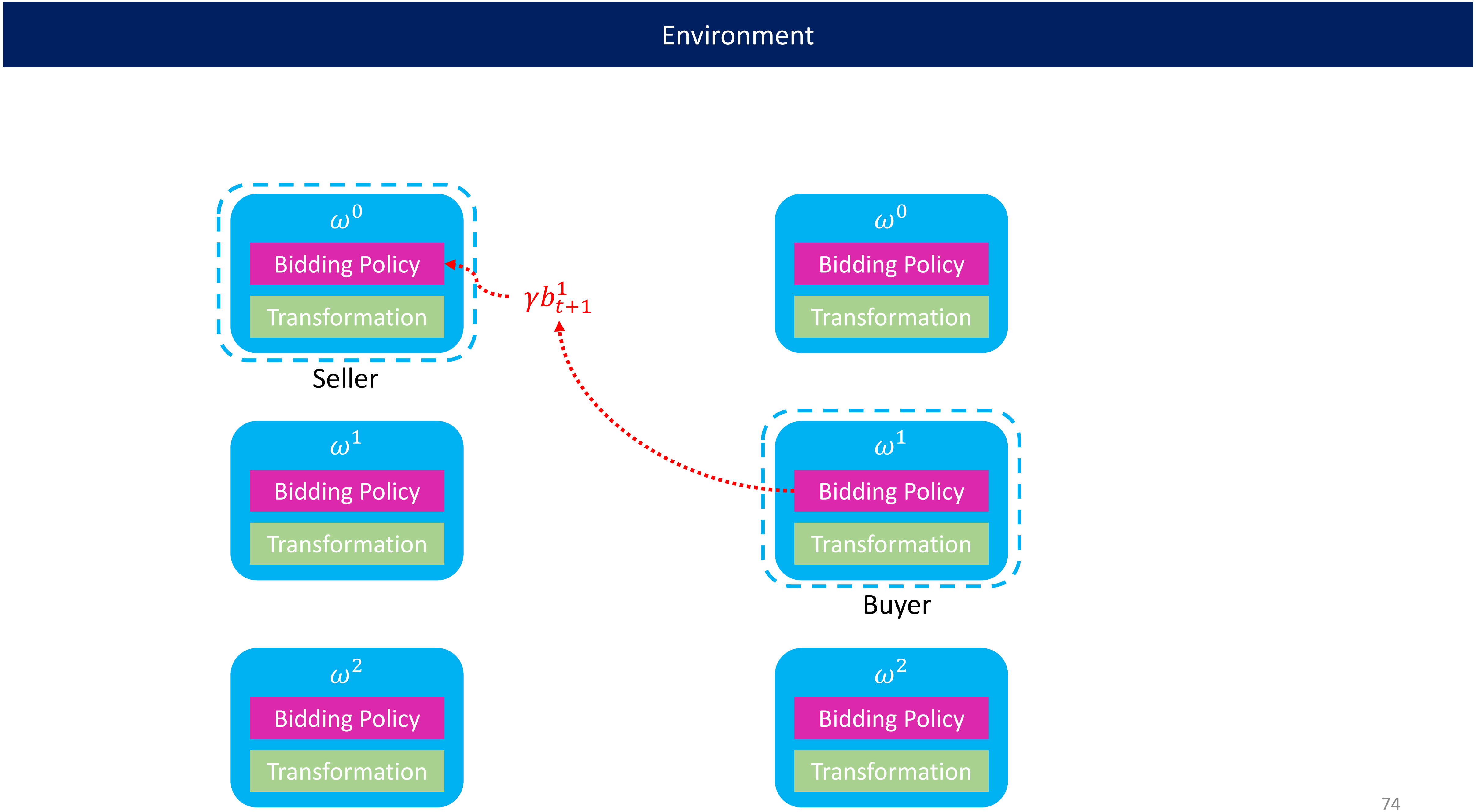
Environment
Bidders
ω^0
Bidding Policy
Transformation
Seller
ω^1
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation
s_t
s_{t+1}
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation

Environment

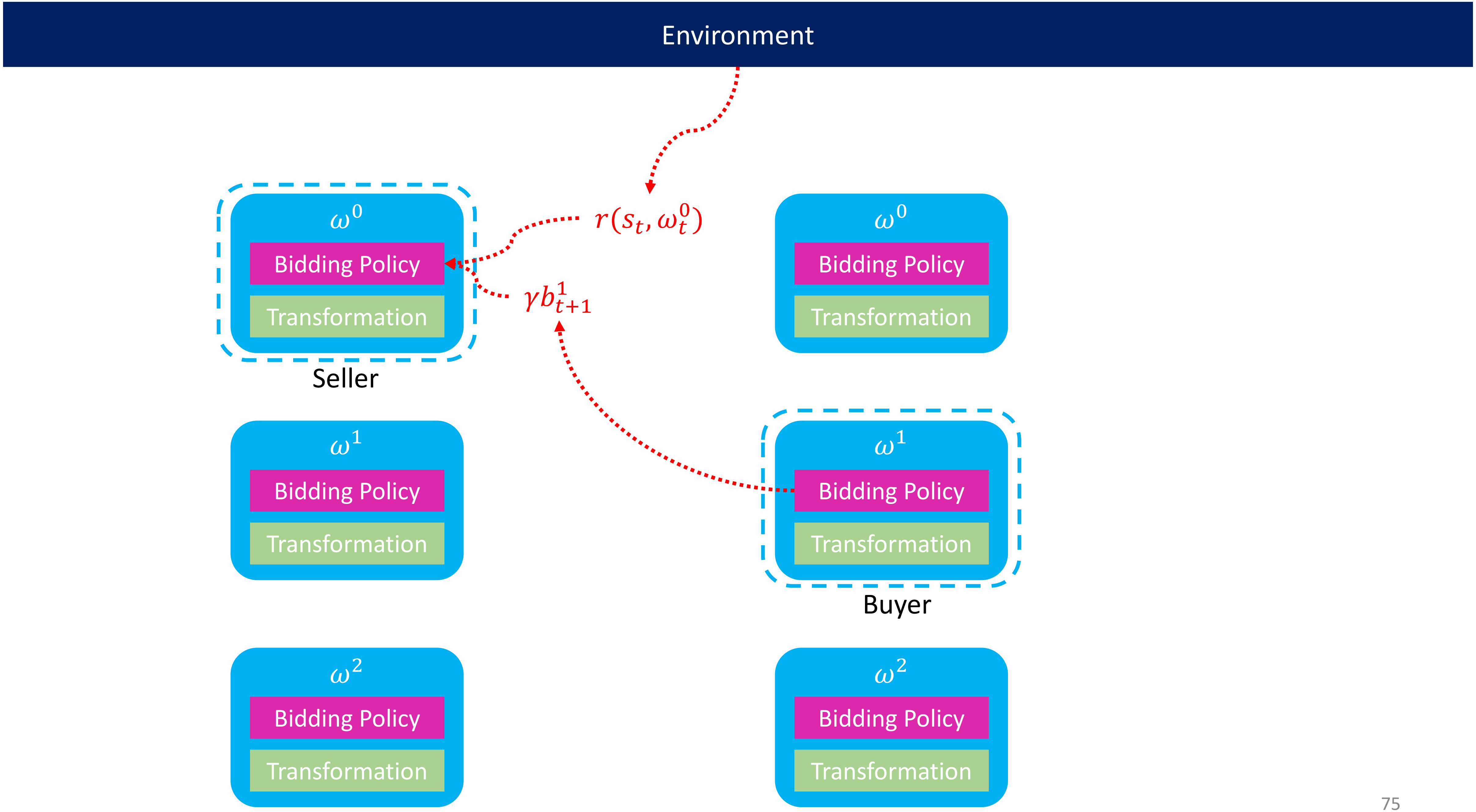


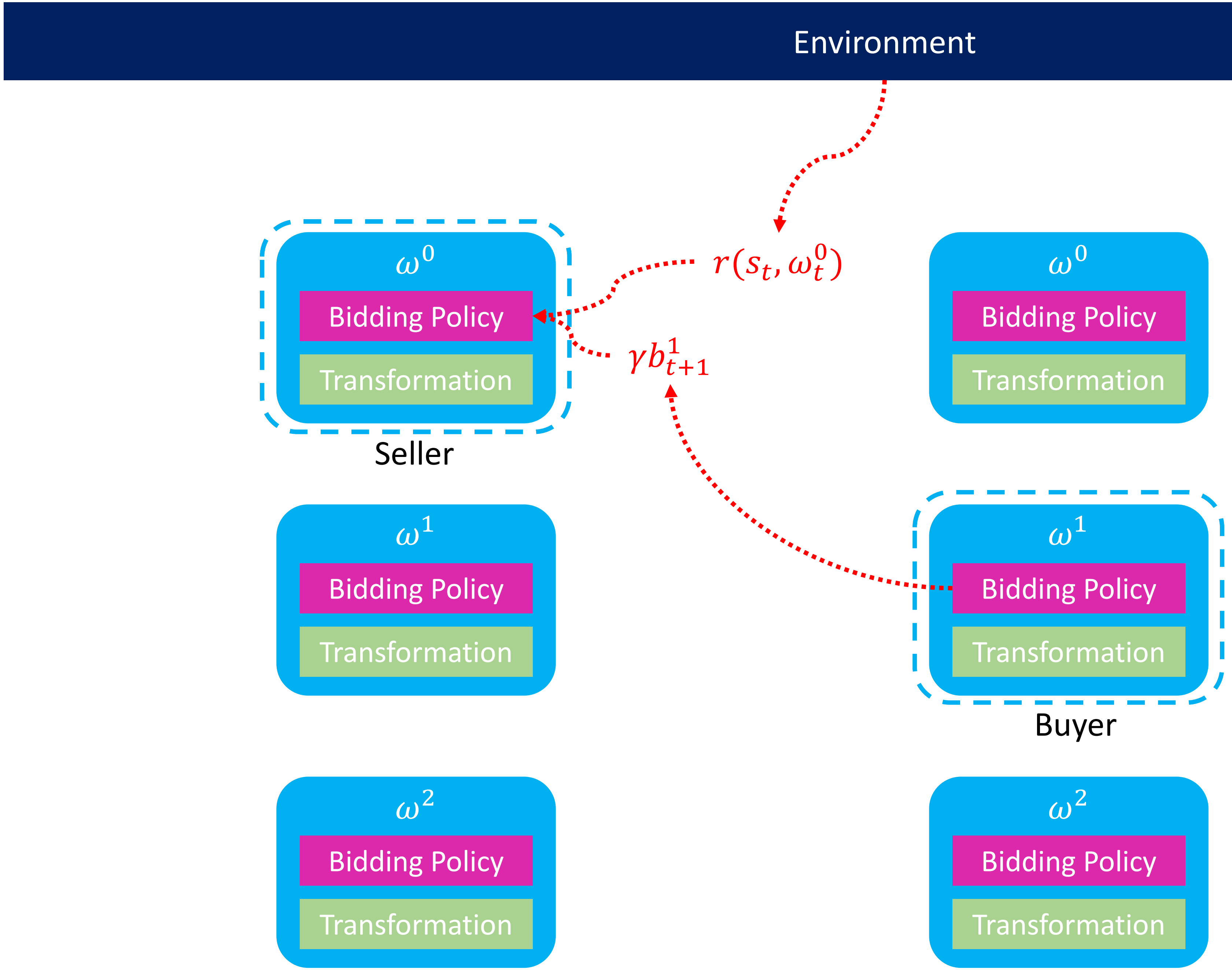
Environment





Environment
ω^0
Bidding Policy
Transformation
Seller
γb_{t+1}^1
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
Buyer
ω^2
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation


Environment
$r(s_t, \omega_t^0)$
γb_{t+1}^1
ω^0
Bidding Policy
Transformation
Seller
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
Buyer
ω^2
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation



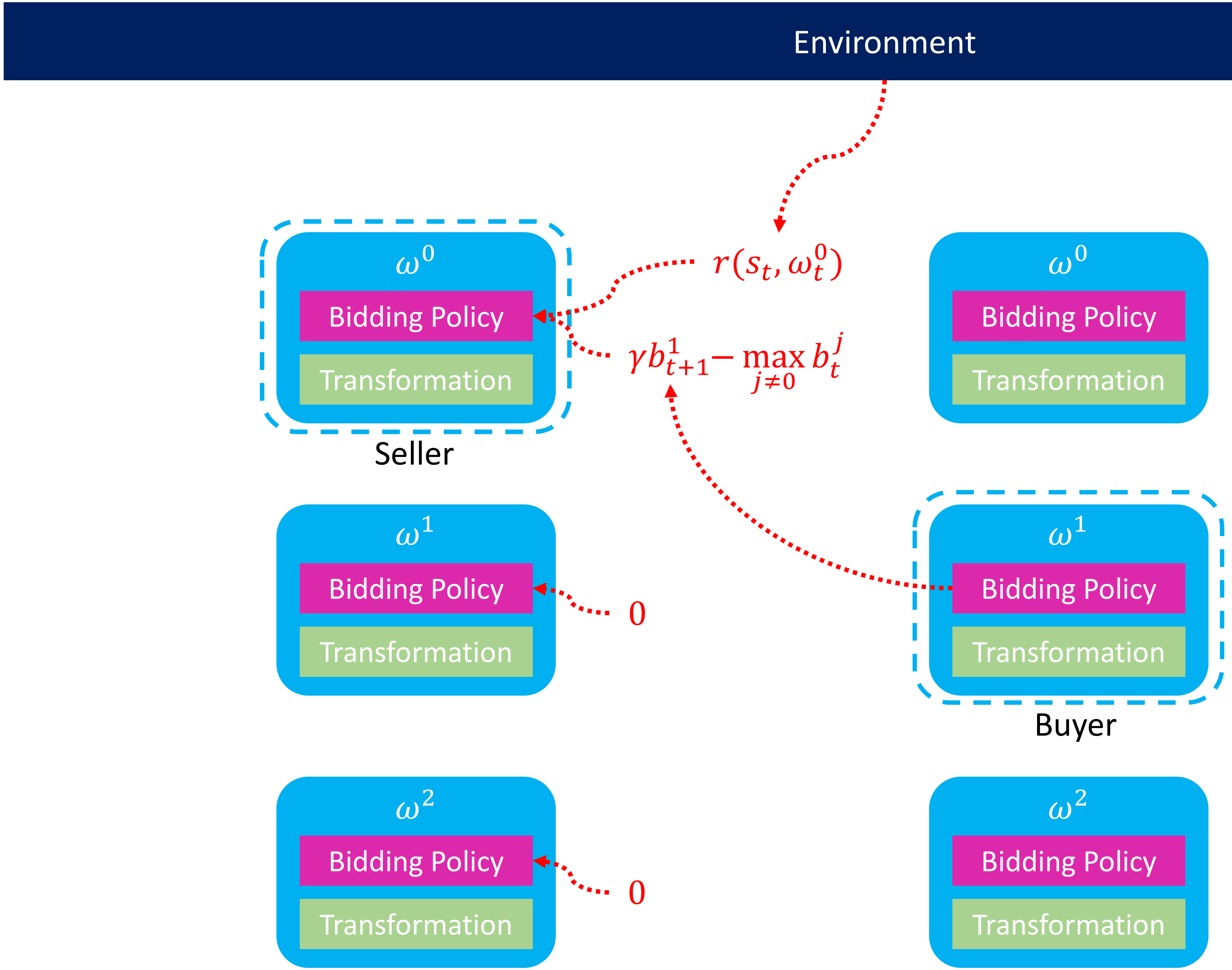
Valuations

Before:

$$v^i(s_t) = Q^*(s_t \omega_t^i)$$

Now:

$$v^i(s_t) = r(s_t, \omega_t^i) + \gamma \max_k b_{t+1}^k$$



Valuations

Before:

$$v^i(s_t) = Q^*(s_t\omega_t^i)$$

Now:

$$v^i(s_t) = r(s_t, \omega_t^i) + \gamma \max_k b_{t+1}^k$$

Utilities

Winner's utility

$$u^i(b) = v^i - \max_{j\neq i} b^j$$

Loser's utility

$$u^i(b) = 0$$

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

Proposition: If the utilities are defined as below, it is a Nash equilibrium for every primitive to bid their optimal Q value in the Global MDP.

Valuations

Before:

$$v^i(s_t) = Q^*\left(s_t, \omega_t^i\right)$$

Now:

$$v^i(s_t) = r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k$$

Utilities

Winners:

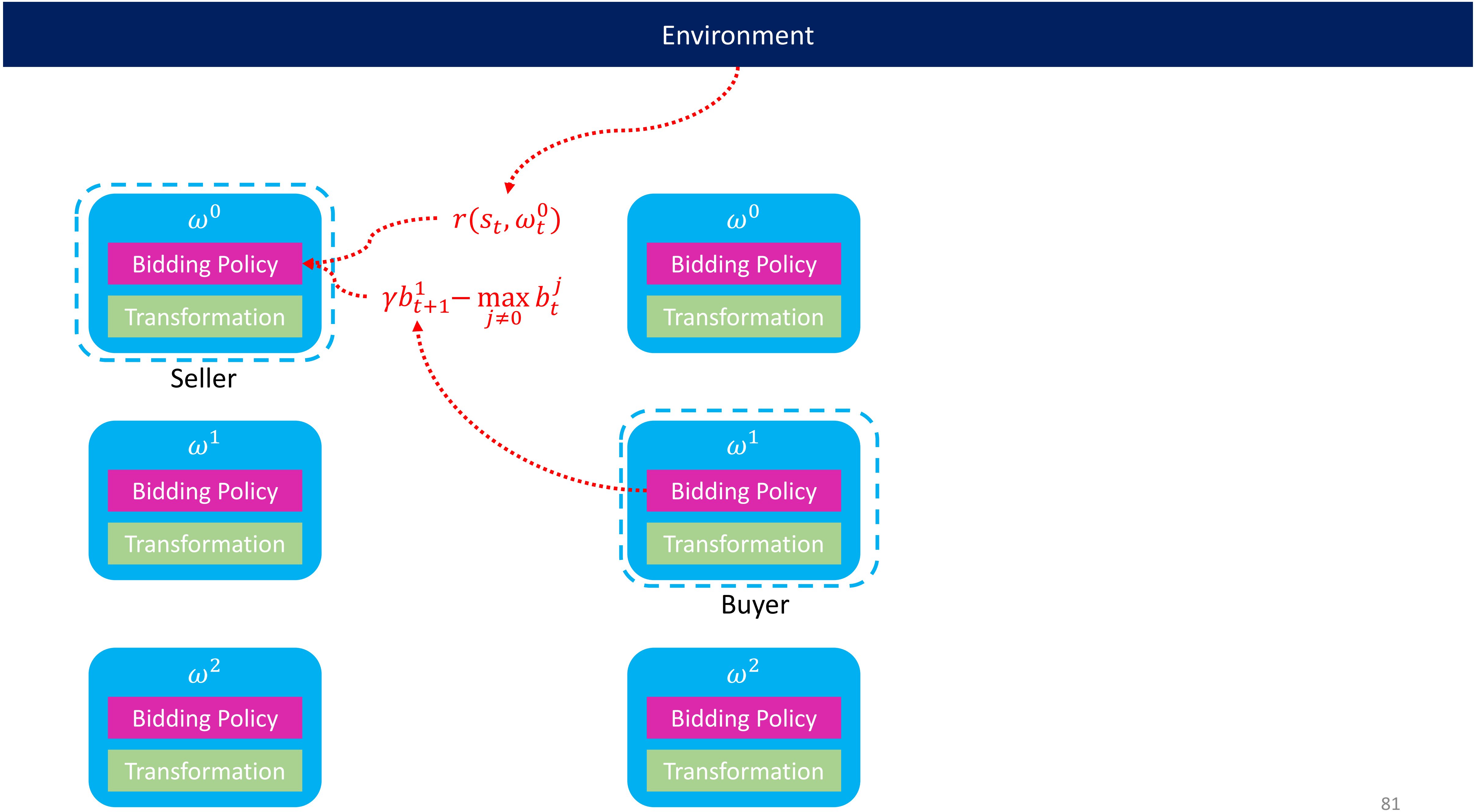
$$u^i(b) = \left[r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k\right] - \max_{j \neq i} b^j$$

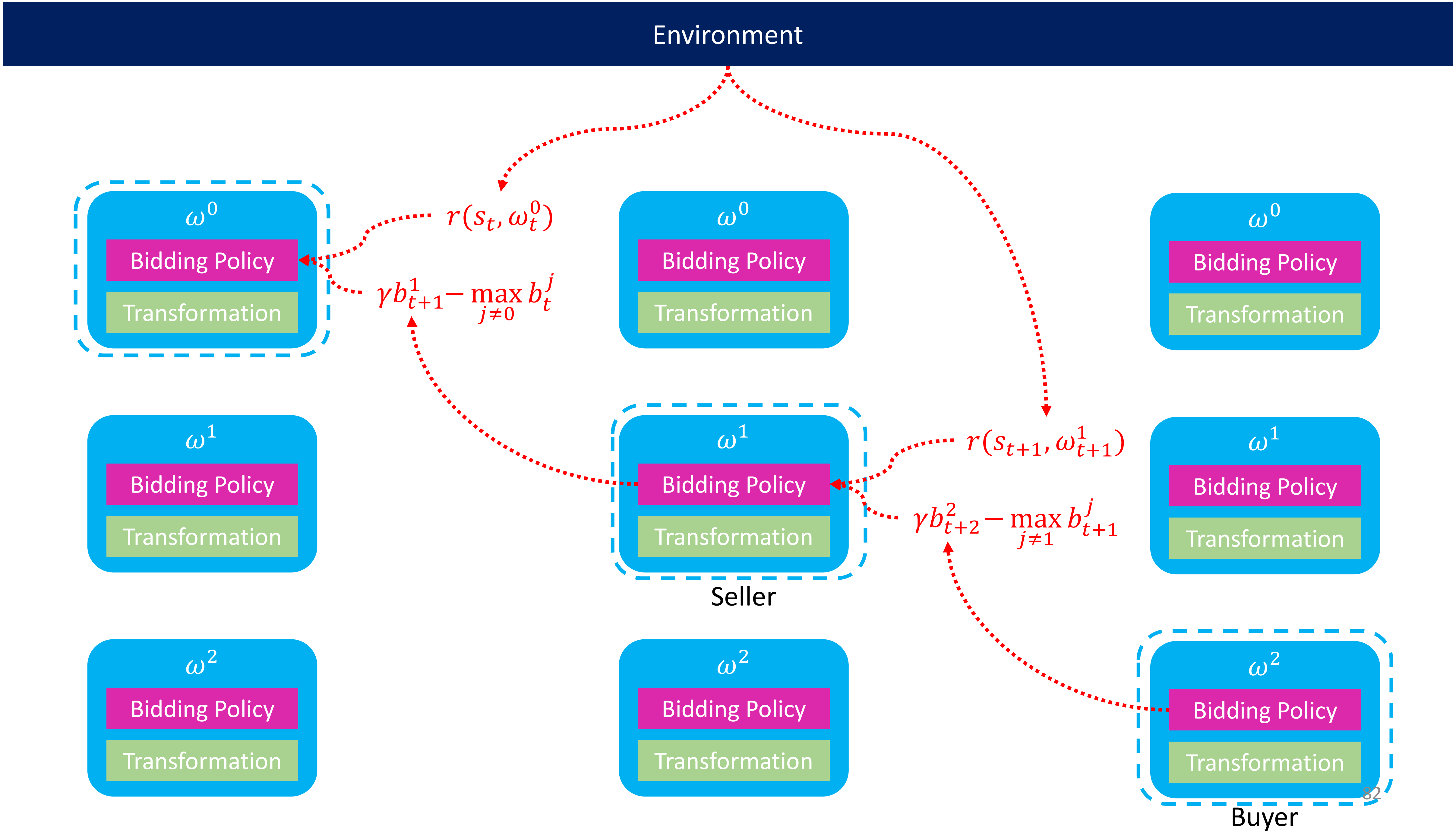
Losers:

$$u^i(b) = 0$$

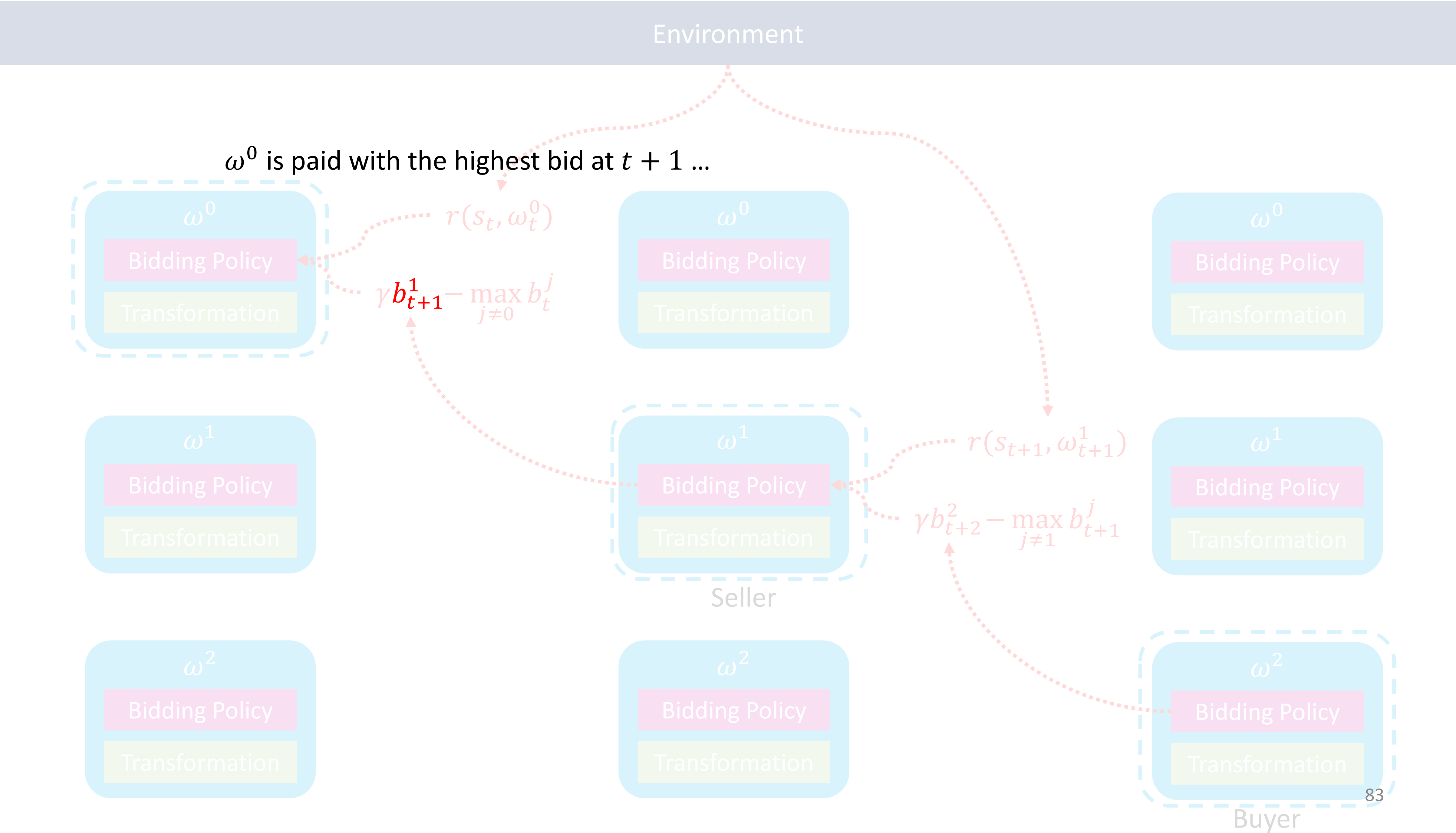
But wait…

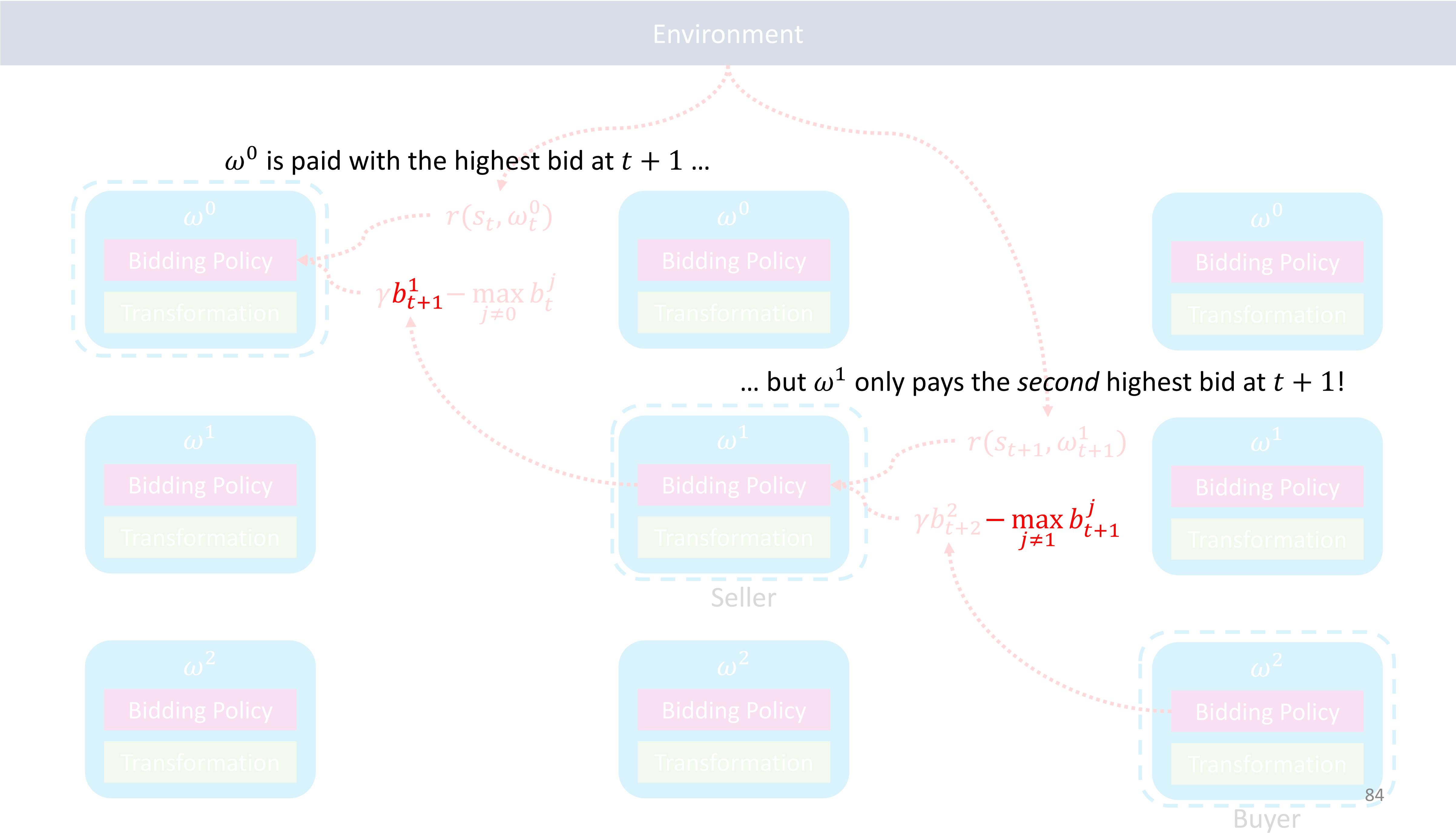
Utility is not conserved!


Environment
ω^0
Bidding Policy
Transformation
Seller
$r(s_t, \omega_t^0)$
$\gamma b_{t+1}^1 - \max_{j \neq 0} b_t^j$
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
Buyer
ω^2
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation


Environment
ω^0
Bidding Policy
Transformation
$r(s_t, \omega_t^0)$
$\gamma b_{t+1}^1 - \max_{j \neq 0} b_t^j$
ω^1
ω^2
$r(s_{t+1}, \omega_{t+1}^1)$
$\gamma b_{t+2}^2 - \max_{j \neq 1} b_{t+1}^j$
Seller
Buyer

ω^0 is paid with the highest bid at $t + 1$ …

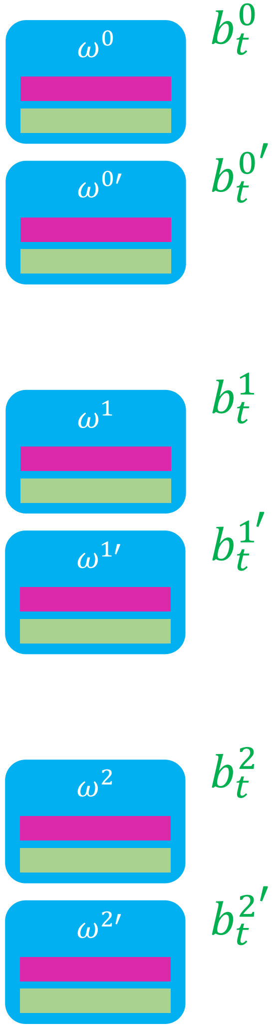



Environment
ω^0 is paid with the highest bid at $t + 1$ …
$r(s_t, \omega_t^0)$
$\gamma b_{t+1}^1 - \max_{j \neq 0} b_t^j$
… but ω^1 only pays the *second* highest bid at $t + 1$!
$r(s_{t+1}, \omega_{t+1}^1)$
$\gamma b_{t+2}^2 - \max_{j \neq 1} b_{t+1}^j$
ω^0
Bidding Policy
Transformation
ω^1
Bidding Policy
Transformation
ω^2
Bidding Policy
Transformation
Seller
Buyer

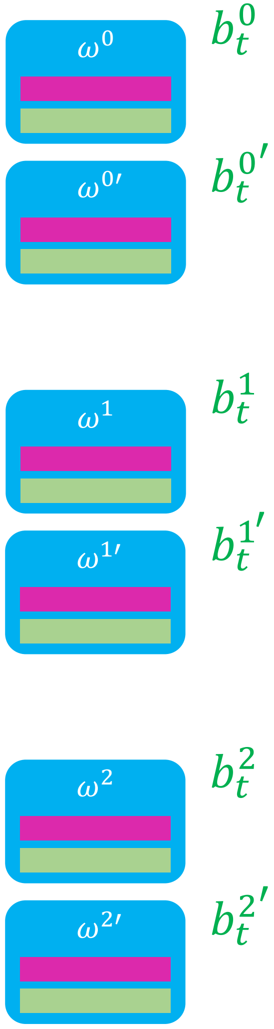
Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

Environment



Environment

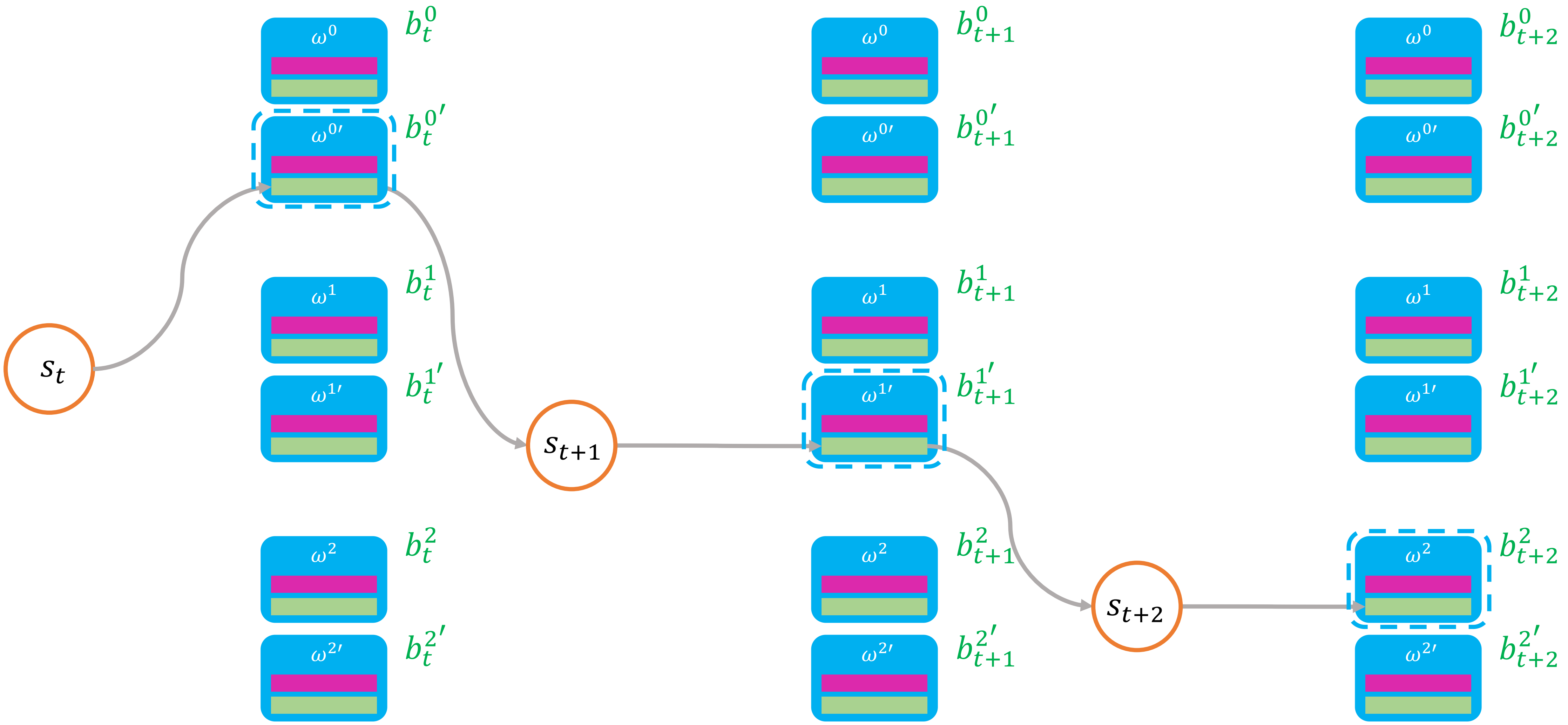


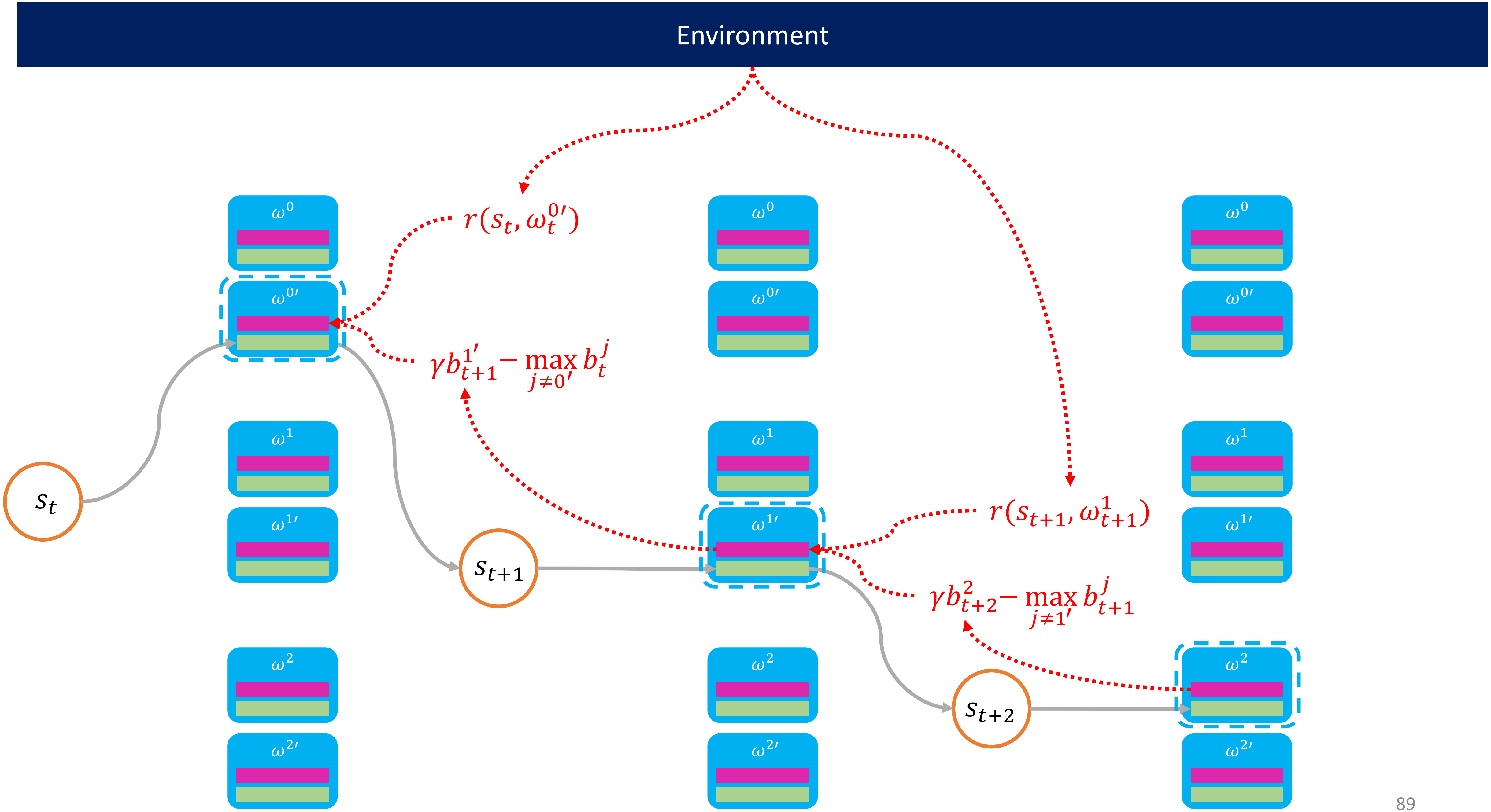
$$b_t^0 = b_t^{0\prime}$$

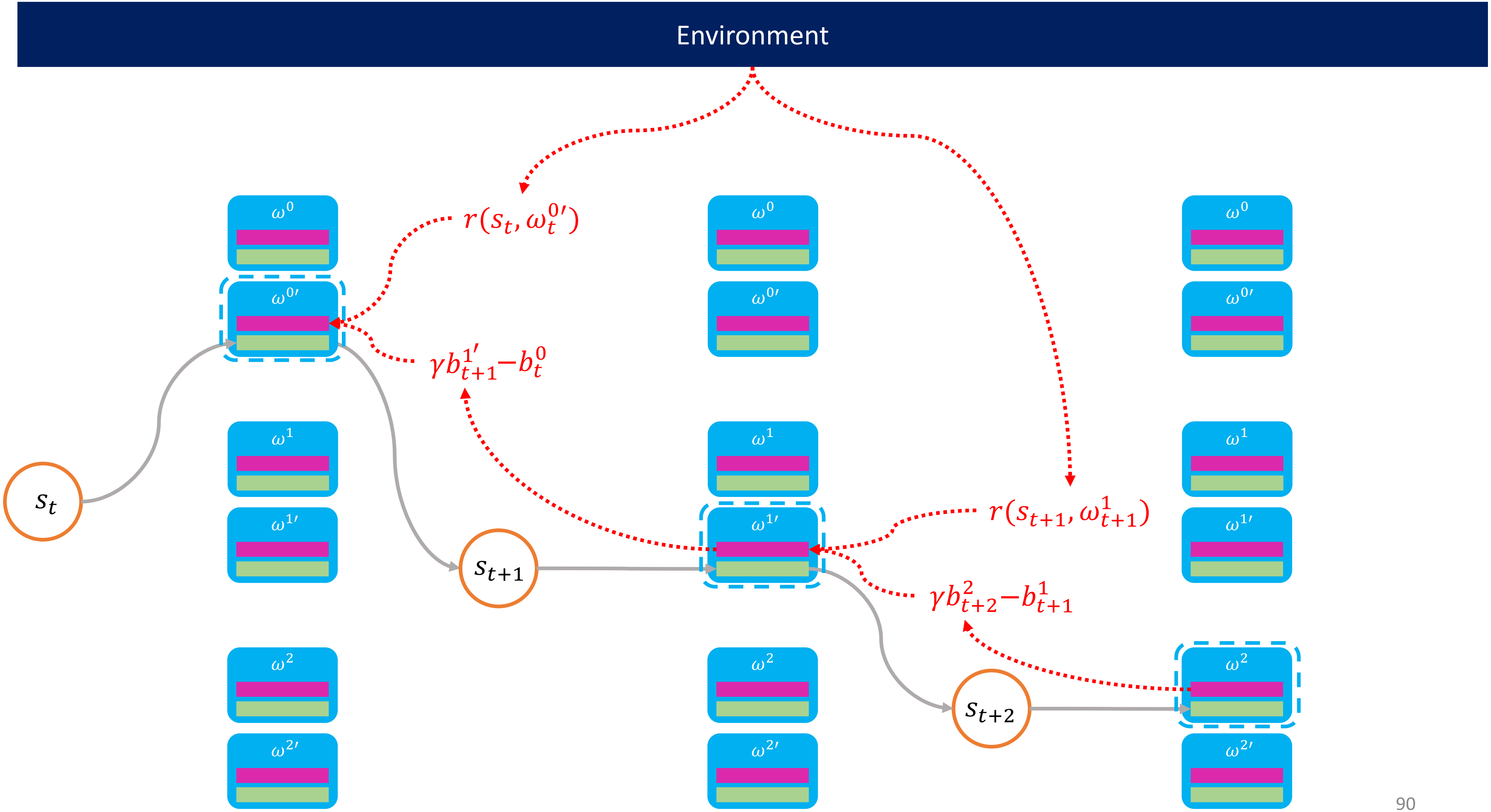
$$b_t^1 = b_t^{1\prime}$$

$$b_t^2 = b_t^{2\prime}$$

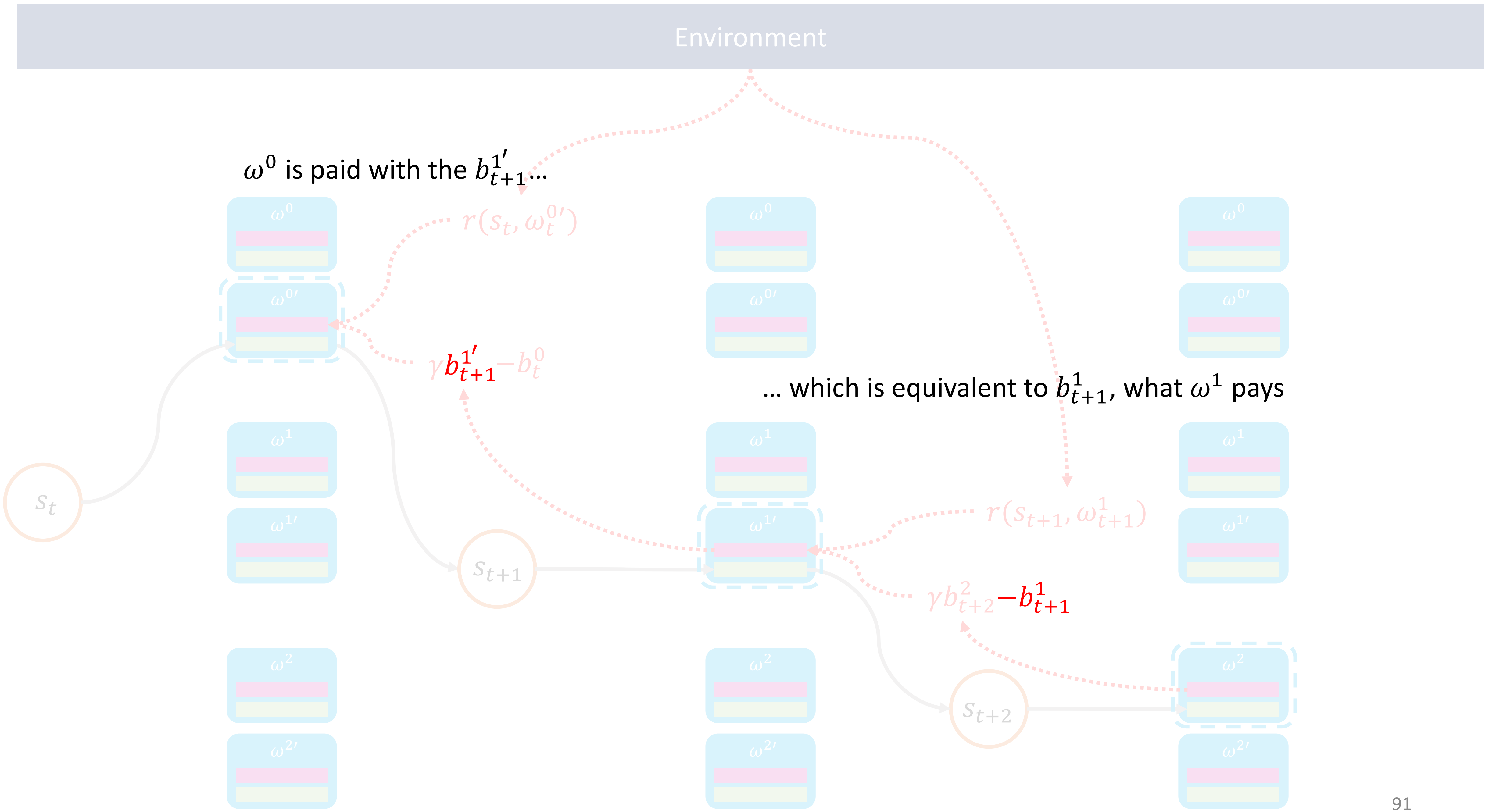
Environment




Environment
$r(s_t, \omega_t^{0'})$
$\gamma b_{t+1}^{1'} - \max_{j \neq 0'} b_t^j$
$r(s_{t+1}, \omega_{t+1}^{1})$
$\gamma b_{t+2}^{2} - \max_{j \neq 1'} b_{t+1}^j$
ω^0
$\omega^{0'}$
ω^1
$\omega^{1'}$
ω^2
$\omega^{2'}$
s_t
s_{t+1}
s_{t+2}

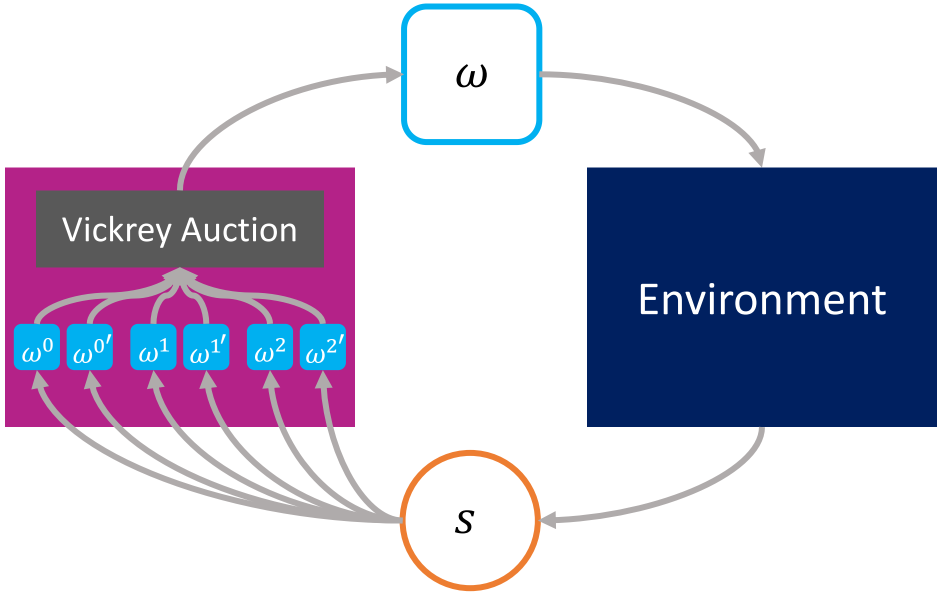


Environment
$r(s_t, \omega_t^{0\prime})$
$\gamma b_{t+1}^{1\prime} - b_t^0$
$r(s_{t+1}, \omega_{t+1}^1)$
$\gamma b_{t+2}^2 - b_{t+1}^1$
ω^0
$\omega^{0\prime}$
ω^1
$\omega^{1\prime}$
ω^2
$\omega^{2\prime}$
s_t
s_{t+1}
s_{t+2}


Environment
ω^0 is paid with the $b_{t+1}^{1\prime}$…
ω^0
$\omega^{0\prime}$
$r(s_t, \omega_t^{0\prime})$
$\gamma b_{t+1}^{1\prime} - b_t^0$
… which is equivalent to b_{t+1}^1, what ω^1 pays
ω^1
$\omega^{1\prime}$
$r(s_{t+1}, \omega_{t+1}^1)$
$\gamma b_{t+2}^2 - b_{t+1}^1$
ω^2
$\omega^{2\prime}$
s_t
s_{t+1}
s_{t+2}

Main Result: Cloned Vickrey Society

Cloned Vickrey Society



Utilities

Winners:

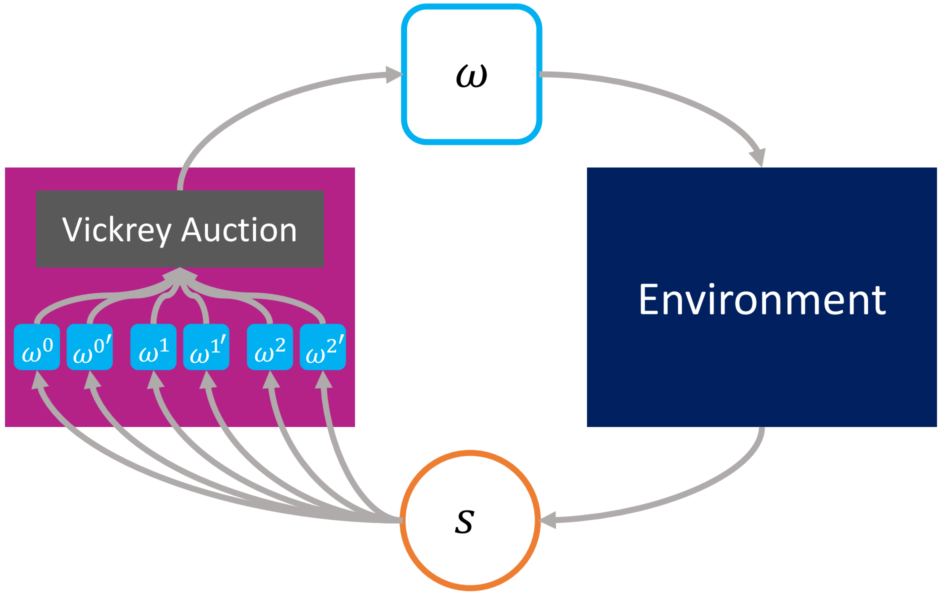
$$u^i(b) = \left[r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k\right] - \max_{j \neq i} b^j$$

Losers:

$$u^i(b) = 0$$

Theorem: In a Cloned Vickrey Society, it is a Nash equilibrium for every primitive to bid their optimal Q value in the Global MDP and utility is conserved.

Cloned Vickrey Society



Utilities

Winners:

$$u^i(b) = \left[r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k\right] - \max_{j \neq i} b^j$$

Losers:

$$u^i(b) = 0$$

Roadmap

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value $\boldsymbol{Q}^*(\boldsymbol{s_t}, \boldsymbol{\omega^i})$** for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	**Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	

From Equilibria to Learning Objectives

Each agent learns a bidding policy by optimizes their utility as reward:

Winners:

$$u^i(b) = \left[r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k\right] - \max_{j \neq i} b^j$$

Losers:

$$u^i(b) = 0$$

Train bidding policies using standard reinforcement learning algorithms

Decentralized Reinforcement Learning

Each agent learns a bidding policy by optimizes their utility as reward:

Winners:

$$u^i(b) = \left[r\left(s_t, \omega_t^i\right) + \gamma \max_k b_{t+1}^k \right] - \max_{j \neq i} b^j$$

Losers:

$$u^i(b) = 0$$

Train bidding policies using standard reinforcement learning algorithms

Society: an emergent solution that is **global** in space and time

Agent: learns via credit assignment **local** in space and time

Contributions

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q^*(s_t, \omega^i)}$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	**Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

Contributions

Assumptions	Key Idea
Assume the agents ω^i know their valuations as $v^i(s_t) = Q^*(s_t, \omega_t^i)$ Dominant strategy equilibrium in auction = solution to Global MDP Pro: provable dominant strategy equilibrium Con: assumes optimal Q-values are known	Define the optimal bid as the **optimal Q value $\boldsymbol{Q^*(s_t, \omega^i)}$** for activating agent ω^i at state s_t. By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
	Temporally couple the agents in a market: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t. **Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

Contributions

Assumptions	Key Idea
Assume the agents ω^i know their valuations as $$v^i(s_t) = Q^*(s_t, \omega_t^i)$$ Dominant strategy equilibrium in auction = solution to Global MDP Pro: provable dominant strategy equilibrium Con: assumes optimal Q-values are known	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q^*(s_t, \omega^i)}$ for activating agent ω^i at state s_t. By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
Assume the agents ω^i know their valuations as $$v^i(s_t) = r(s_t, \omega_t^i) + \gamma \max_k b_{t+1}^k$$ Nash equilibrium in auction = solution to Global MDP Pro: does not assume optimal Q-value is known Con: assumes valuations are known	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t. **Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

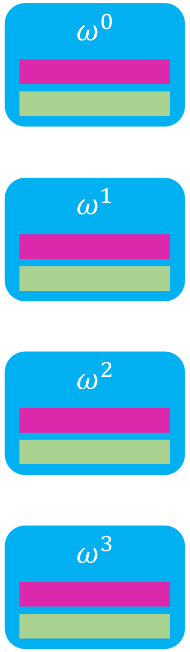
Contributions

Assumptions	Key Idea
Assume the agents ω^i know their valuations as $$v^i(s_t) = Q^*(s_t, \omega_t^i)$$ Dominant strategy equilibrium in auction = solution to Global MDP Pro: provable dominant strategy equilibrium Con: assumes optimal Q-values are known	Define the optimal bid as the **optimal Q value $\boldsymbol{Q^*(s_t, \omega^i)}$** for activating agent ω^i at state s_t. By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
Assume the agents ω^i know their valuations as $$v^i(s_t) = r(s_t, \omega_t^i) + \gamma \max_k b_{t+1}^k$$ Nash equilibrium in auction = solution to Global MDP Pro: does not assume optimal Q-value is known Con: assumes valuations are known	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t. **Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
Assume the agents ω^i learn their valuations through interaction. Nash equilibrium in auction = solution to Global MDP Pro: does not assume valuations are known Con: difficult to prove convergence to equilibrium	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

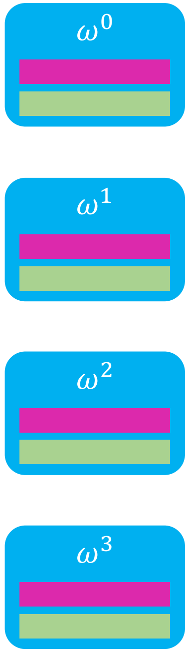
Numerical Simulations

1. How closely do the bids the agents learn match their optimal Q-values?
2. Does the solution to the global objective emerge from the competition among the agents?
3. How does redundancy affect the solutions the agents converge to?
4. Does the modularity of such a decentralized system offer benefit in transferring to new tasks?

Warm-Up: Bandit



Warm-Up: Bandit



Reward $r(\omega^i)$
0.2
0.4
0.6
0.8

Global Objective for the Society
Maximize reward

Local Objectives for the Agents
Maximize utility in the auction

Warm-Up: Bandit

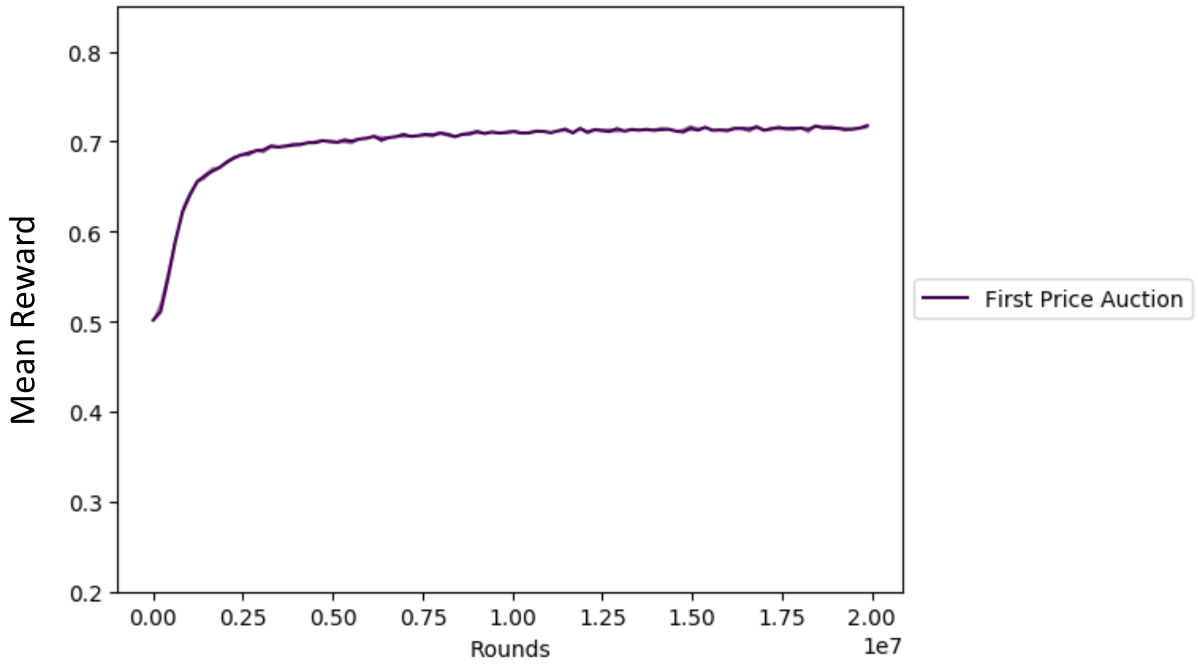
	Reward $r(\omega^i)$	Truthful Bid b^i
ω^0	0.2	0.2
ω^1	0.4	0.4
ω^2	0.6	0.6
ω^3	0.8	0.8

Global Objective for the Society
Maximize reward

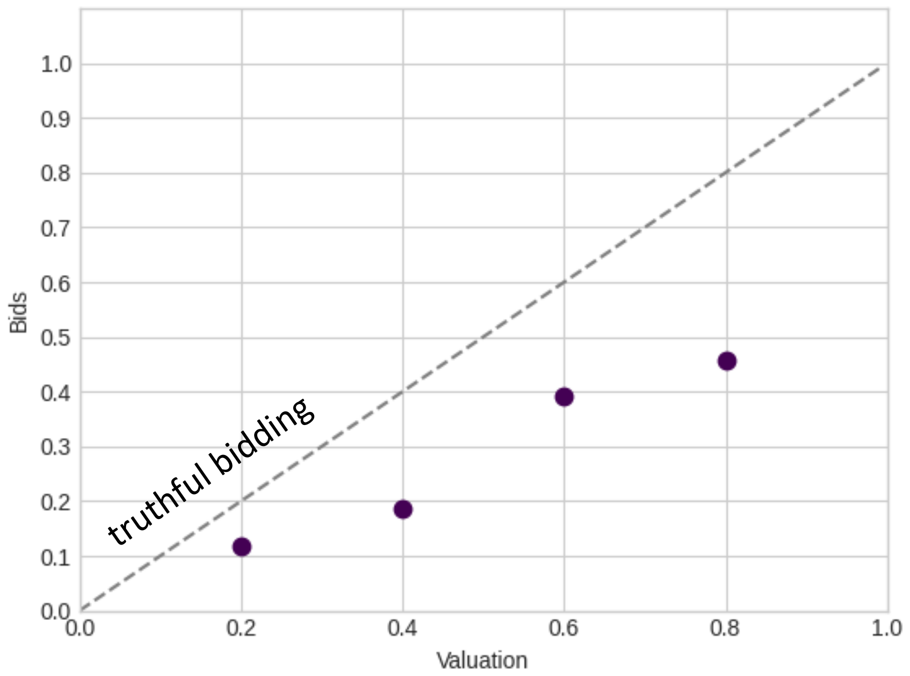
Local Objectives for the Agents
Maximize utility in the auction

Warm-Up: Bandit

Does the solution to the global objective emerge from the competition among the agents?

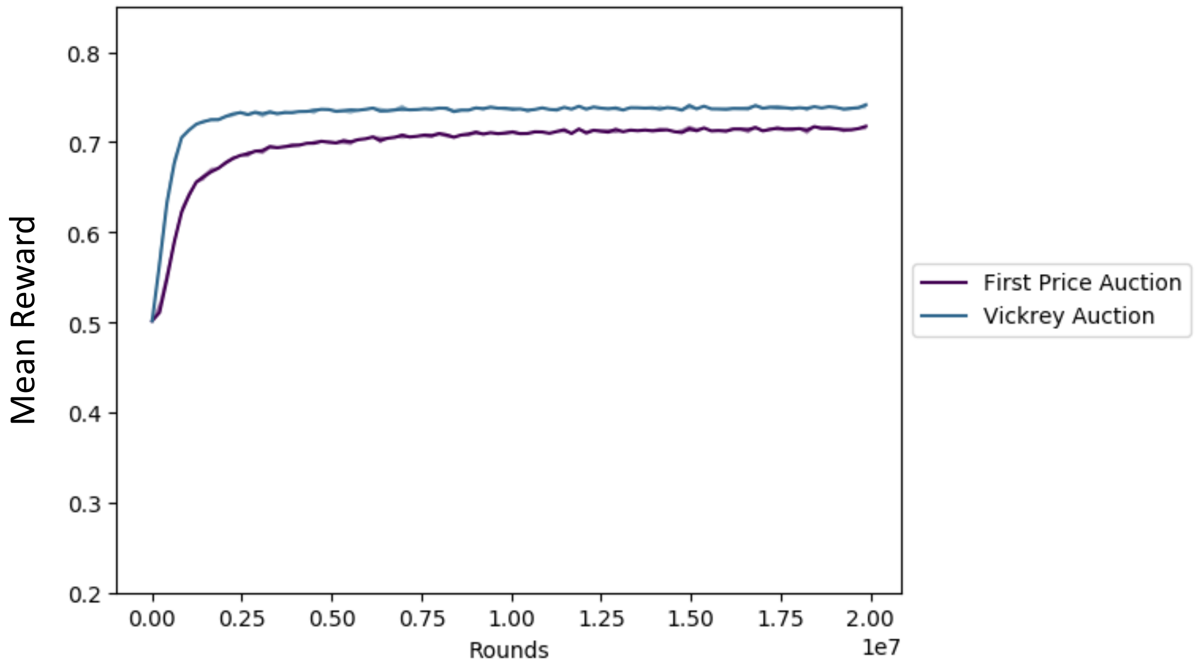


How closely do the bids the agents learn match their optimal Q-values?

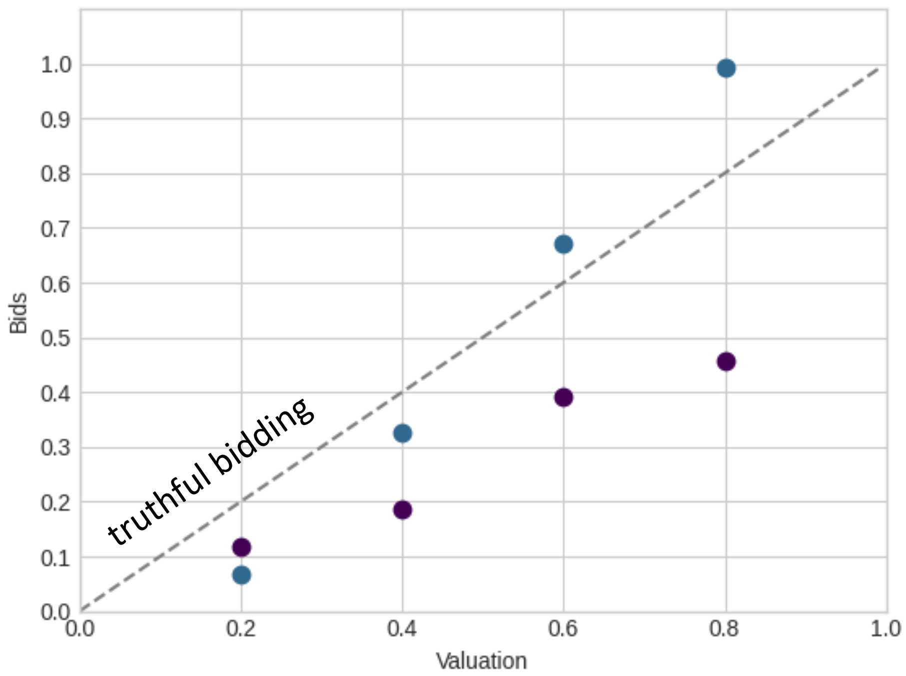


Warm-Up: Bandit

Does the solution to the global objective emerge from the competition among the agents?



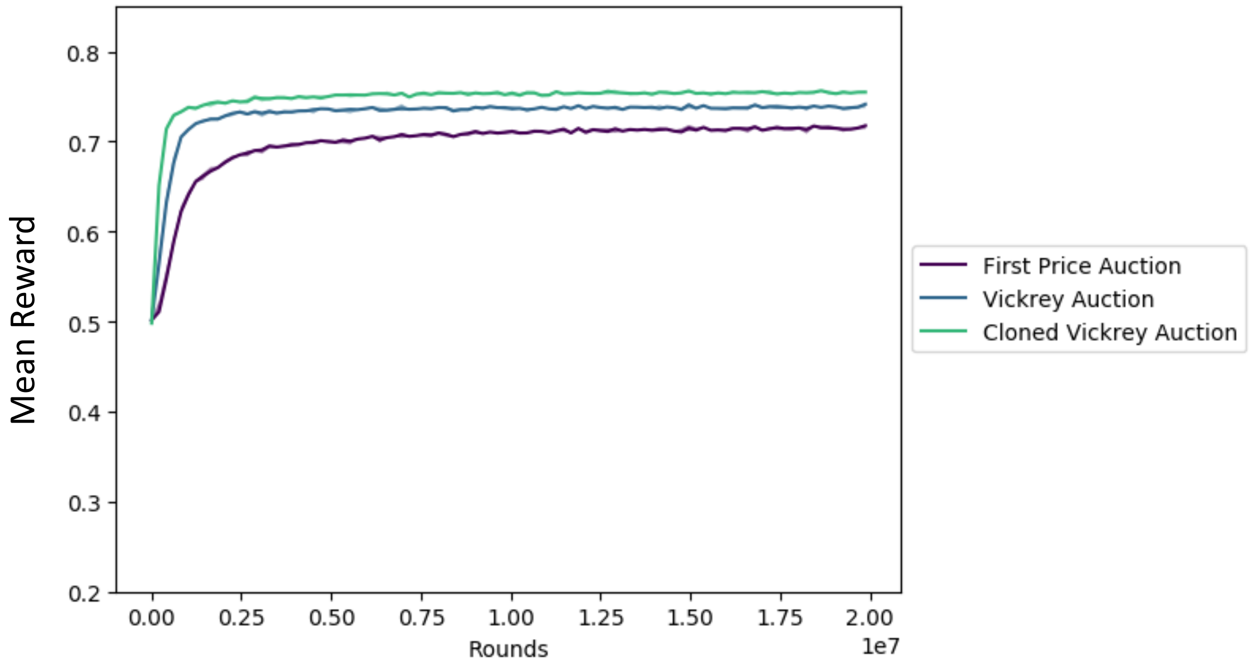
How closely do the bids the agents learn match their optimal Q-values?

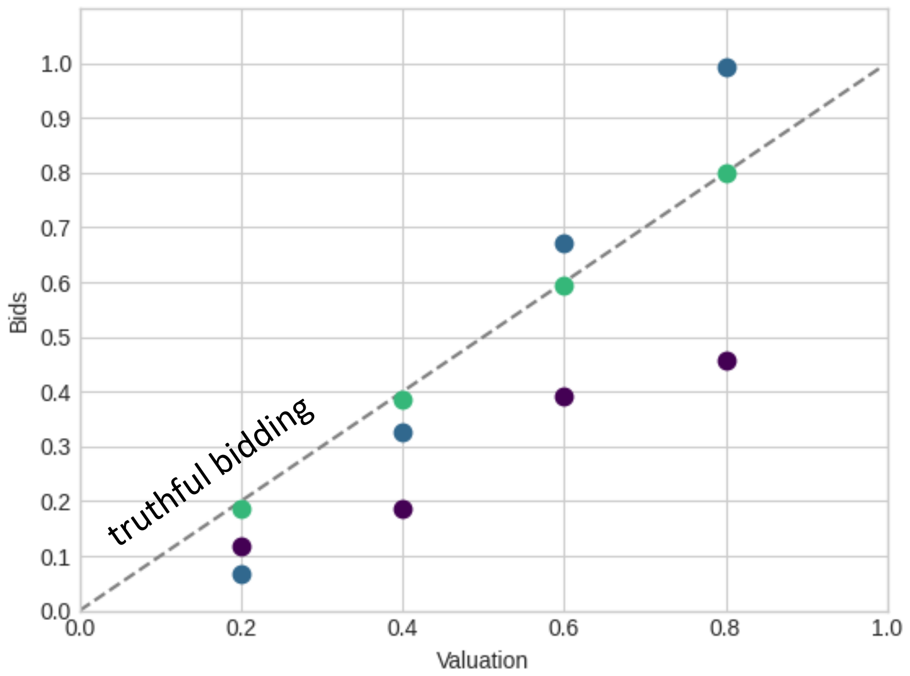


Warm-Up: Bandit

Does the solution to the global objective emerge from the competition among the agents?

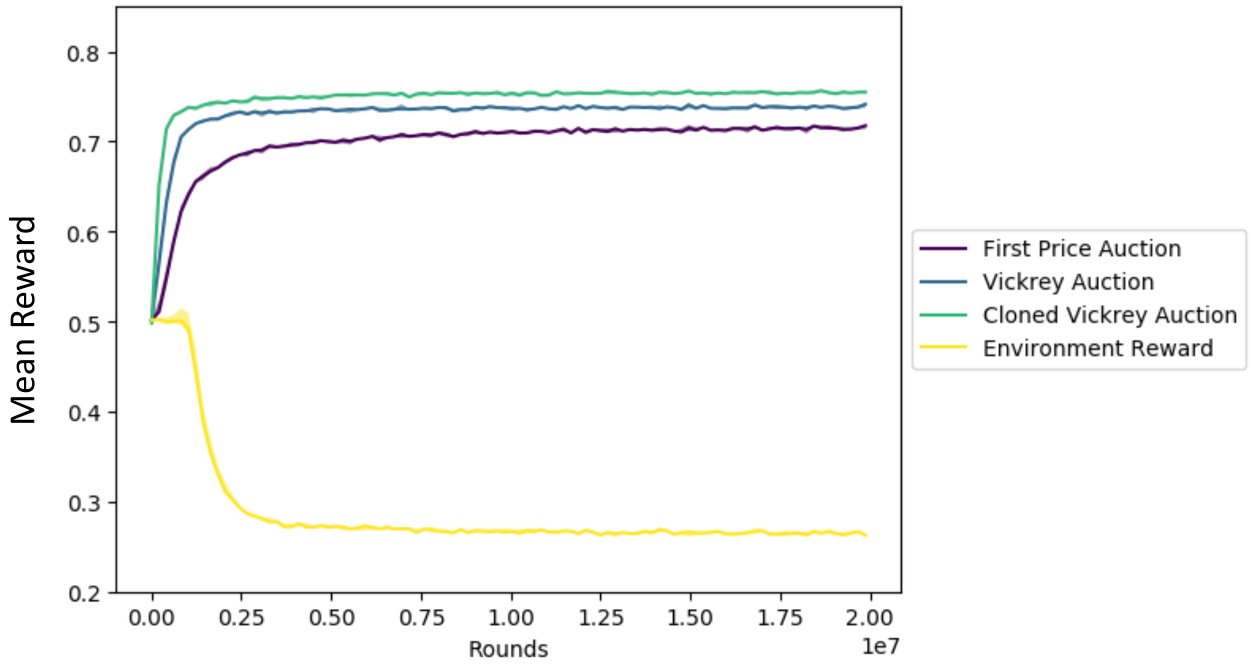
How closely do the bids the agents learn match their optimal Q-values?



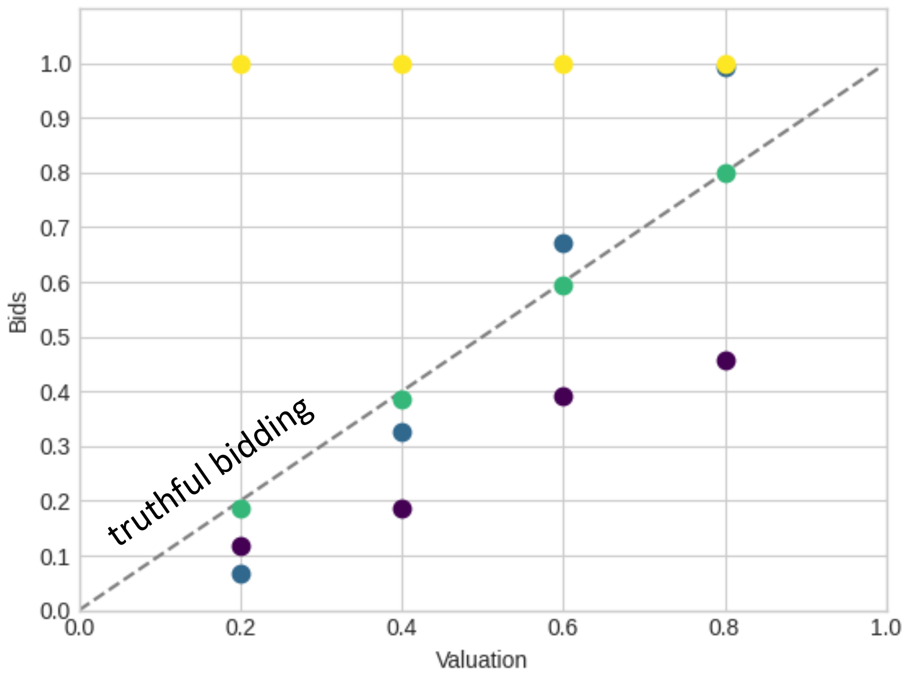


Warm-Up: Bandit

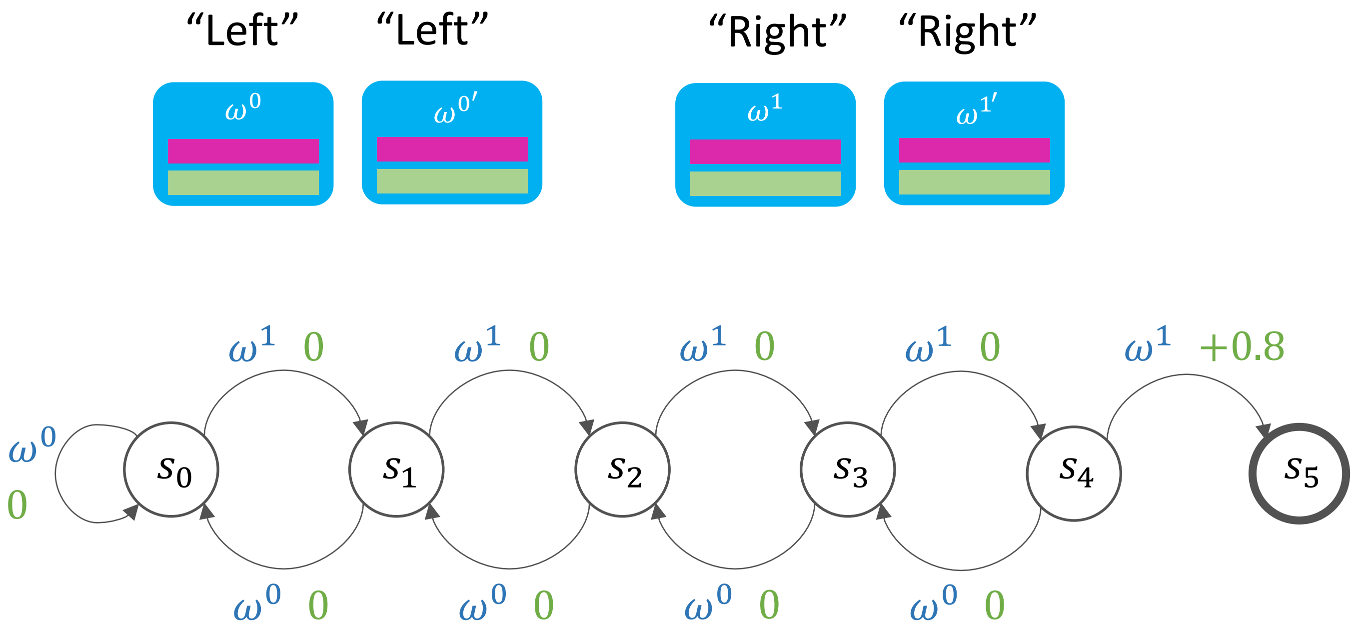
Does the solution to the global objective emerge from the competition among the agents?



How closely do the bids the agents learn match their optimal Q-values?



Multi-Step MDP



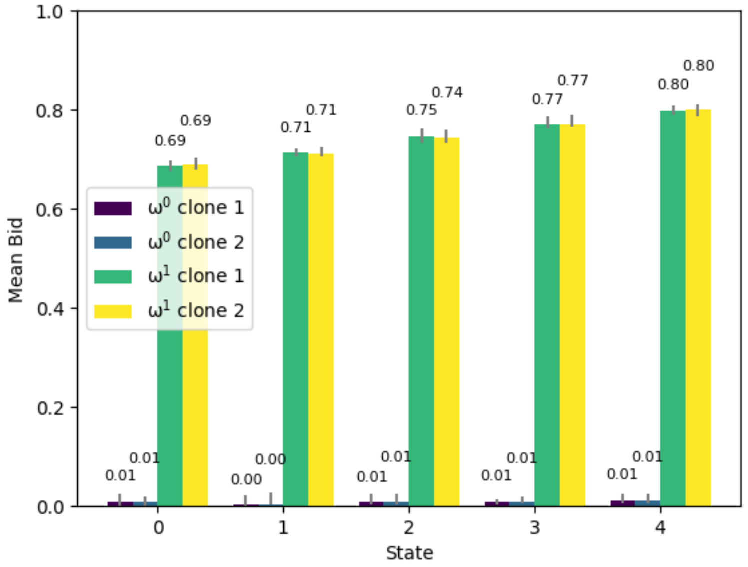
Global Objective for the Society
Maximize return

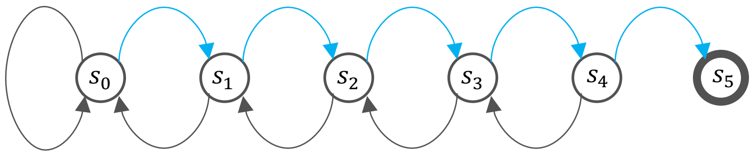
Local Objectives for the Agents
Maximize utility in the auction

Multi-Step MDP

How closely do the bids the agents learn match their optimal Q-values?

Cloned Vickrey Auction

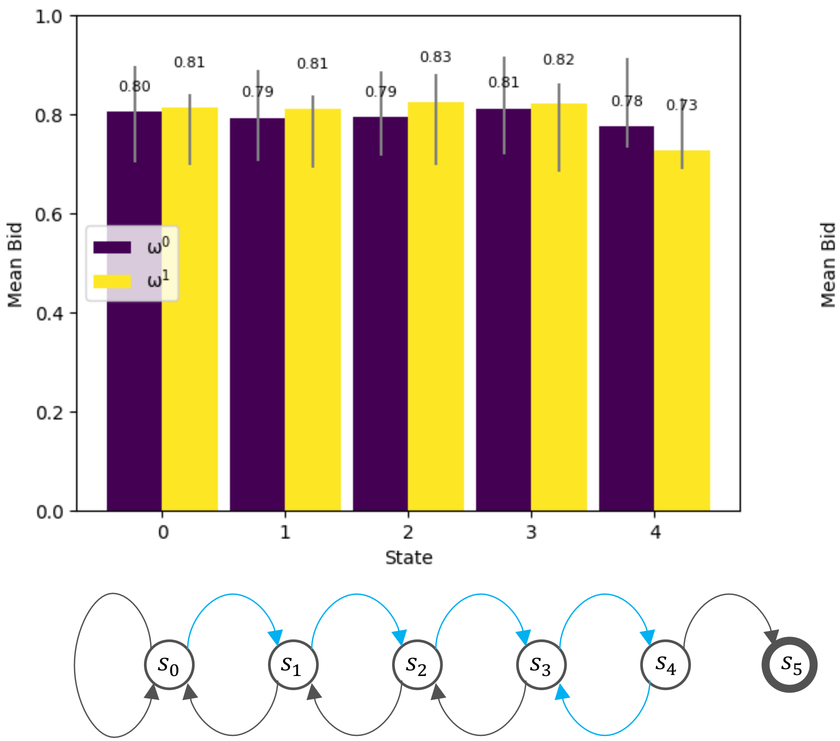




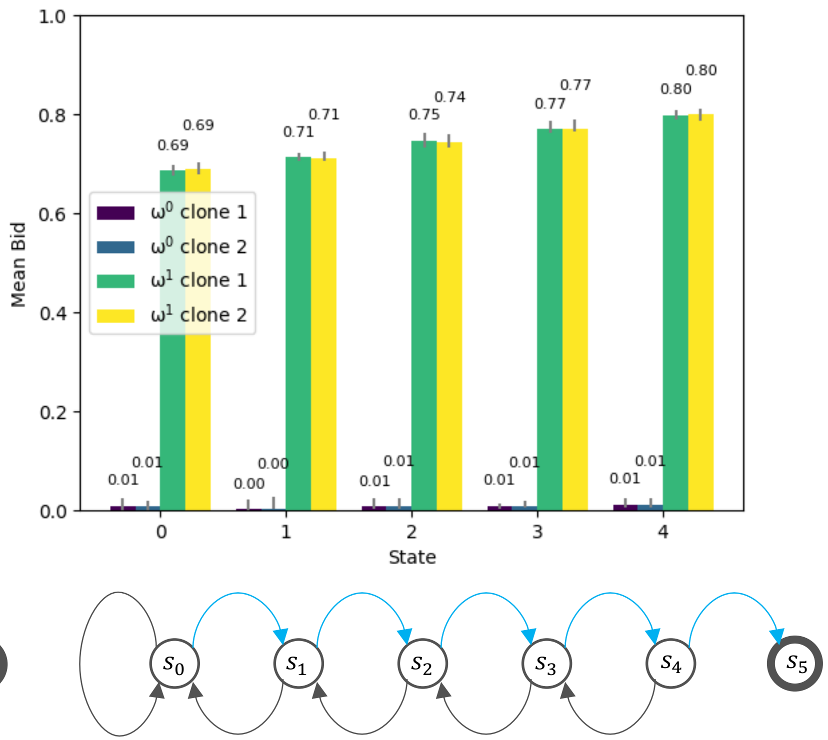
Multi-Step MDP

How closely do the bids the agents learn match their optimal Q-values?

Vickrey Auction



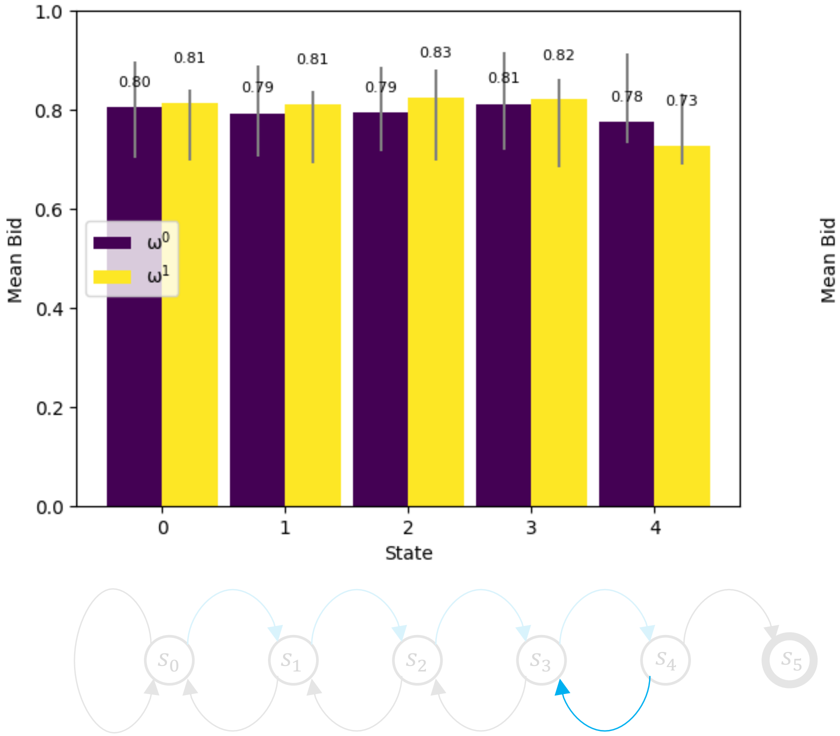
Cloned Vickrey Auction



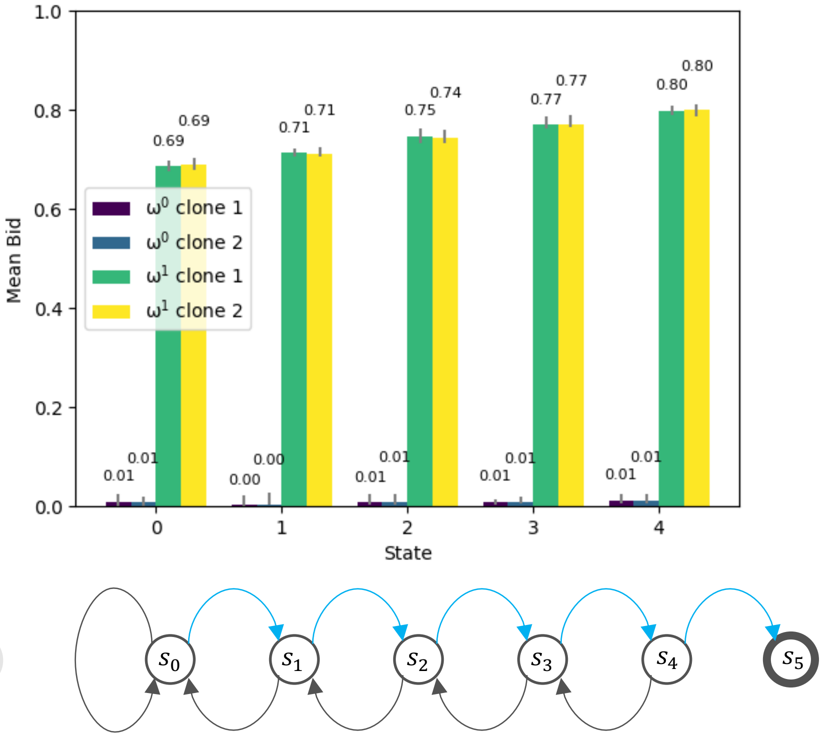
Multi-Step MDP

How closely do the bids the agents learn match their optimal Q-values?

Vickrey Auction



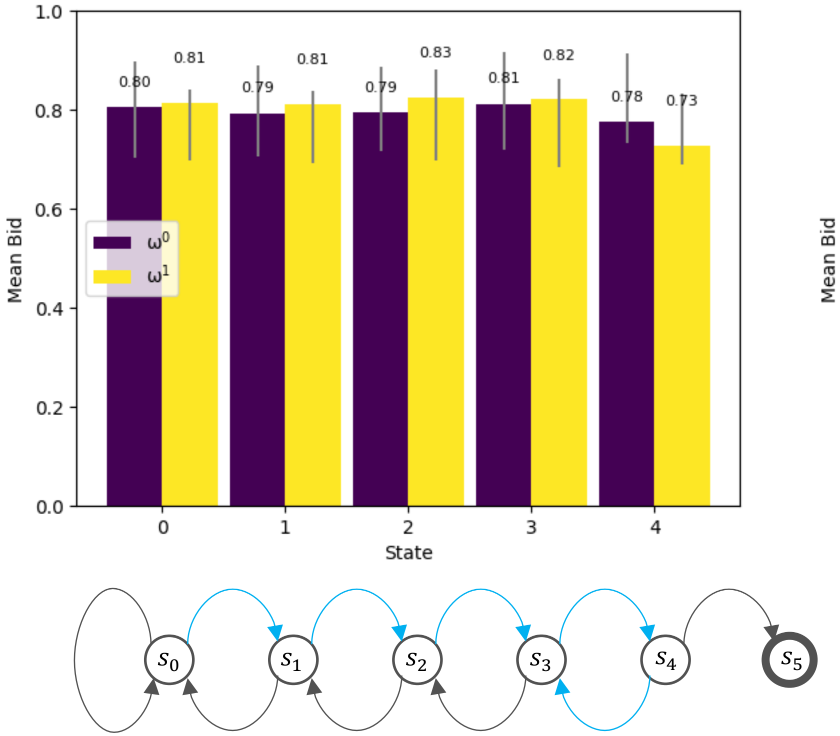
Cloned Vickrey Auction



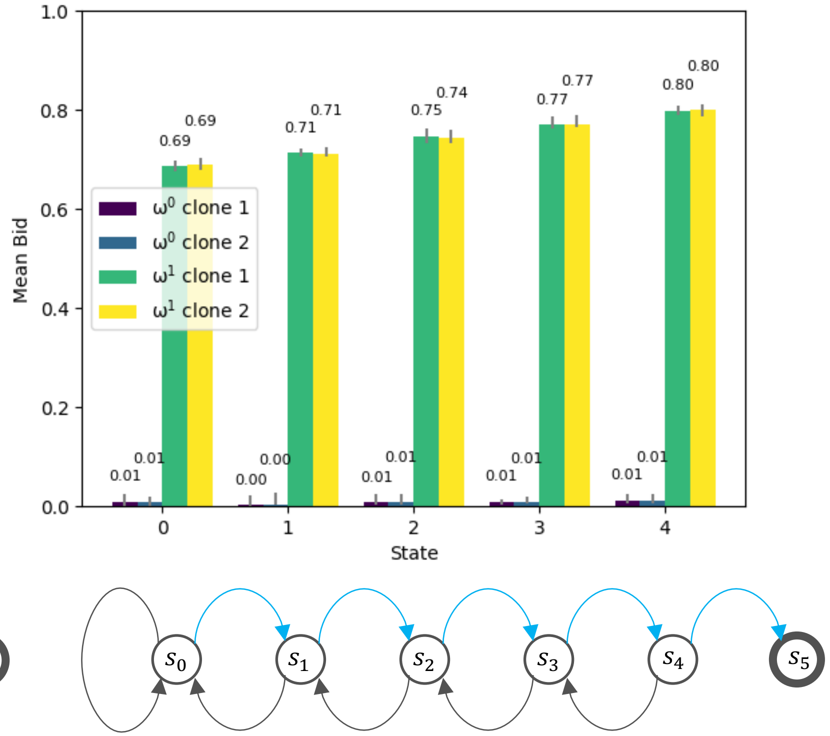
Multi-Step MDP

How closely do the bids the agents learn match their optimal Q-values?

Vickrey Auction



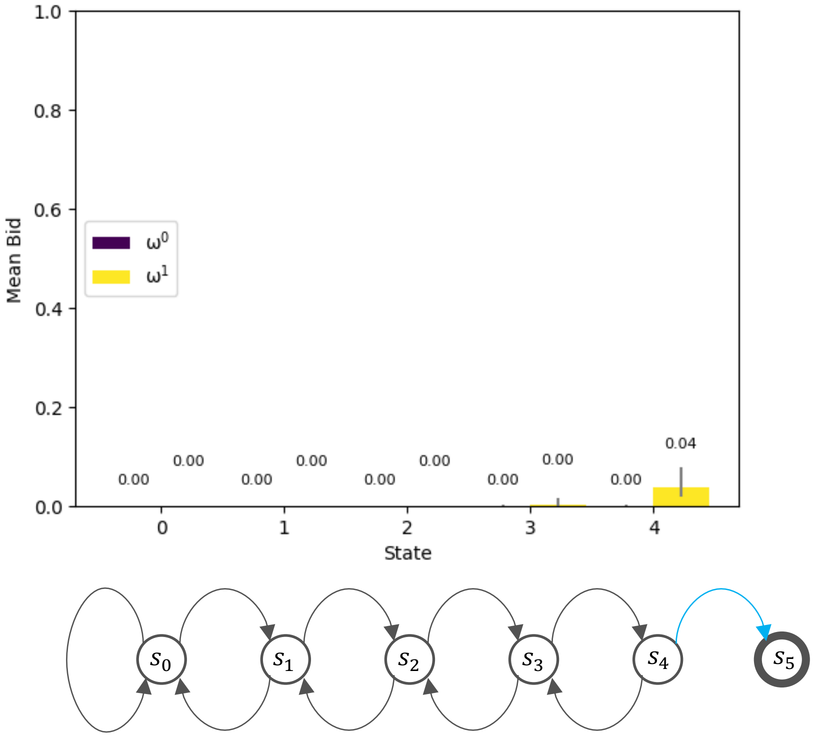
Cloned Vickrey Auction



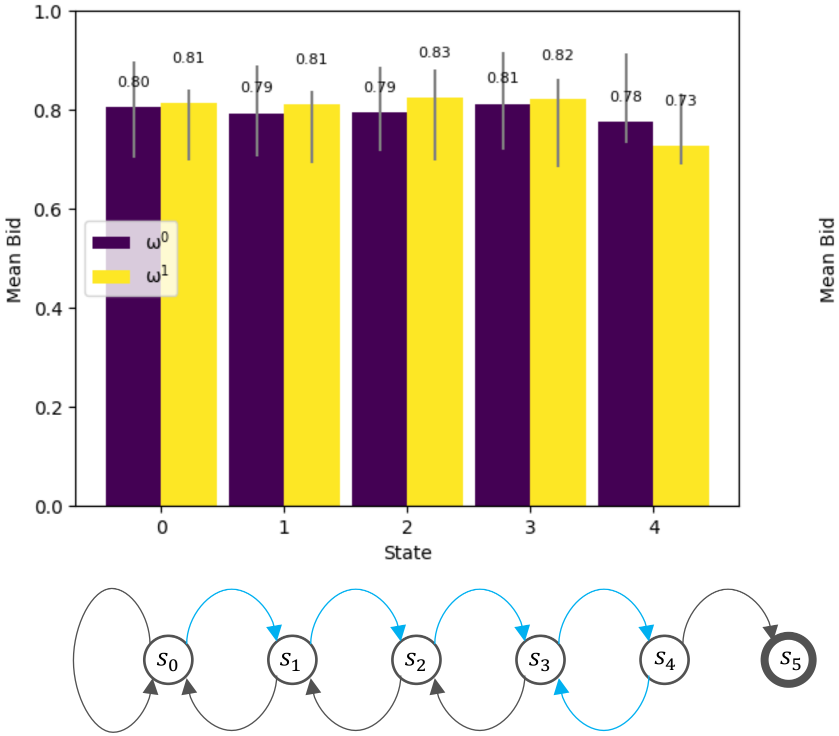
Multi-Step MDP

How closely do the bids the agents learn match their optimal Q-values?

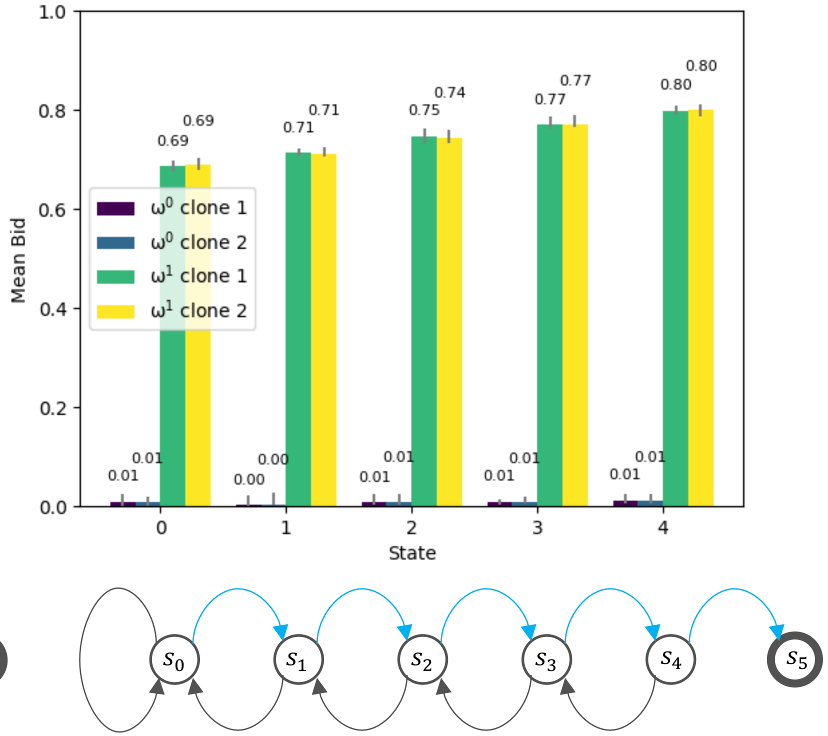
First Price Auction



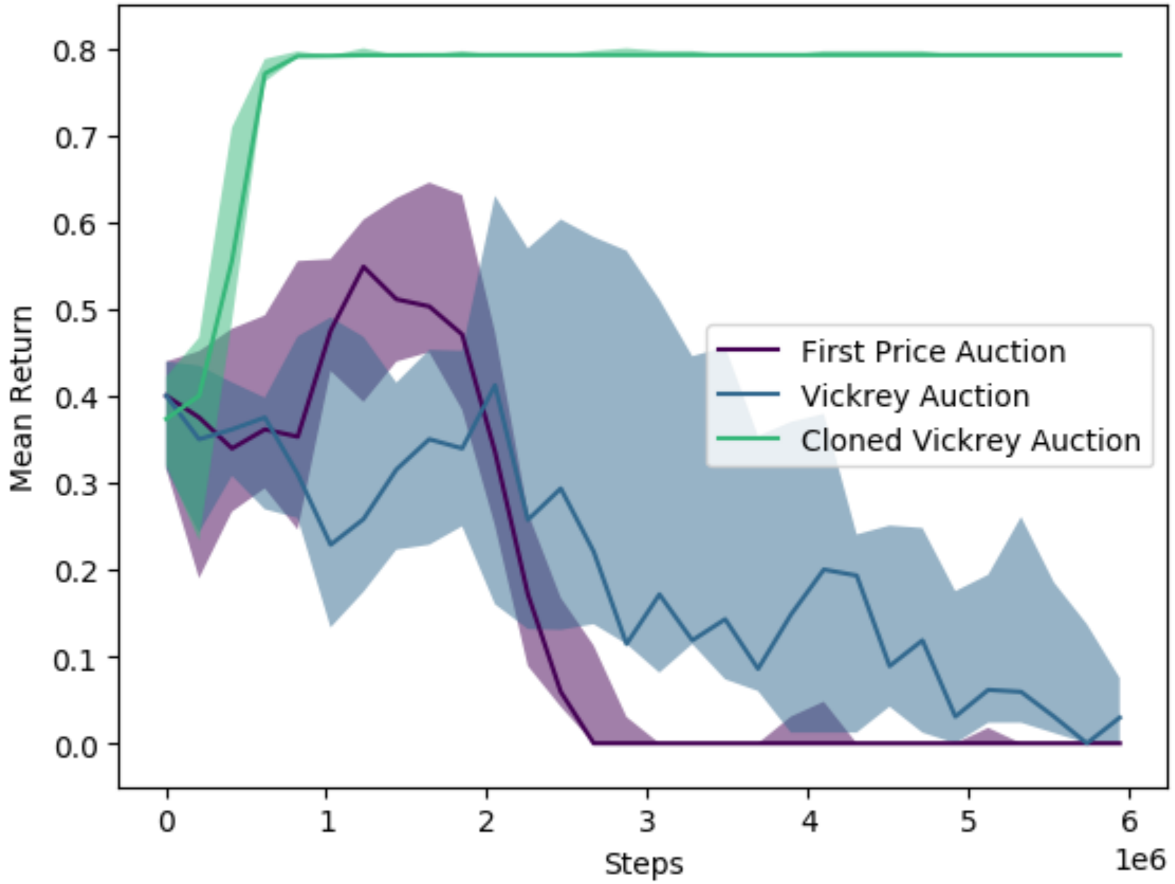
Vickrey Auction



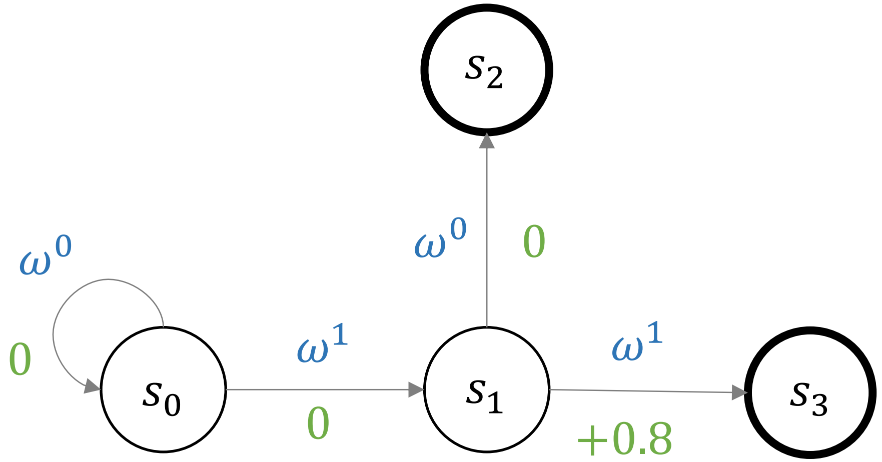
Cloned Vickrey Auction



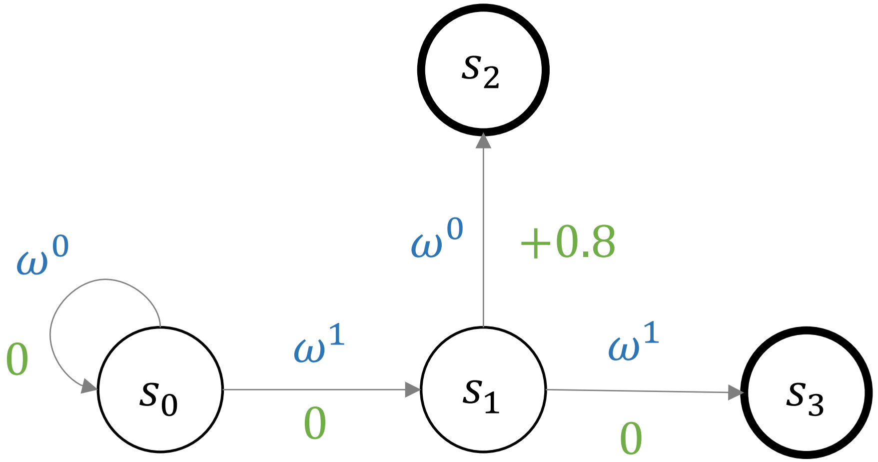
Multi-Step MDP



Transfer



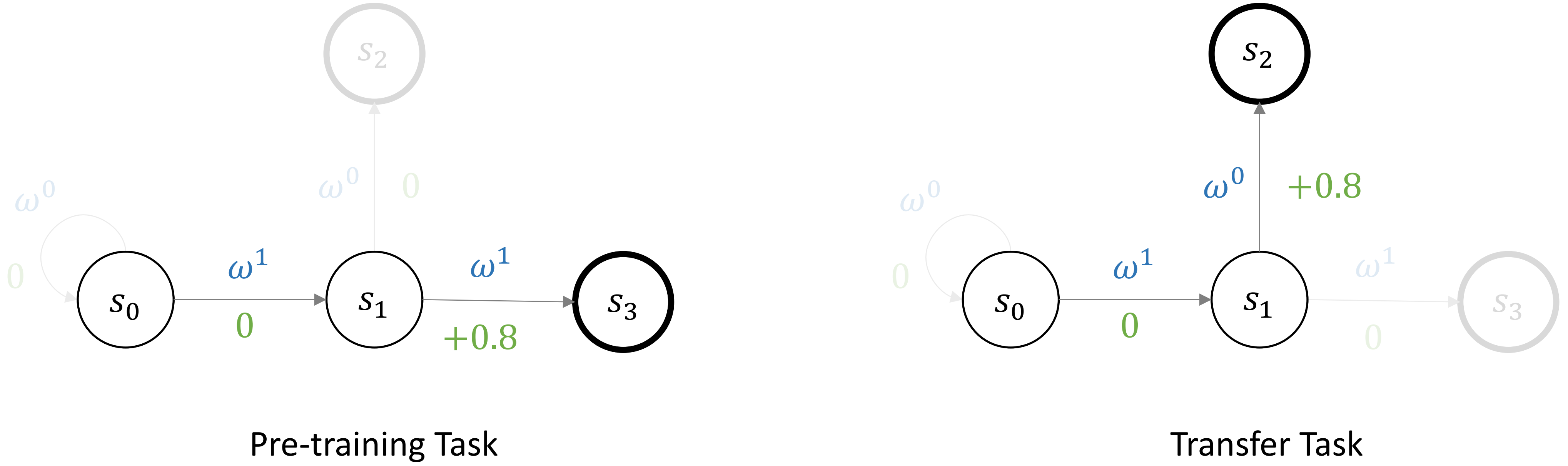
Pre-training Task



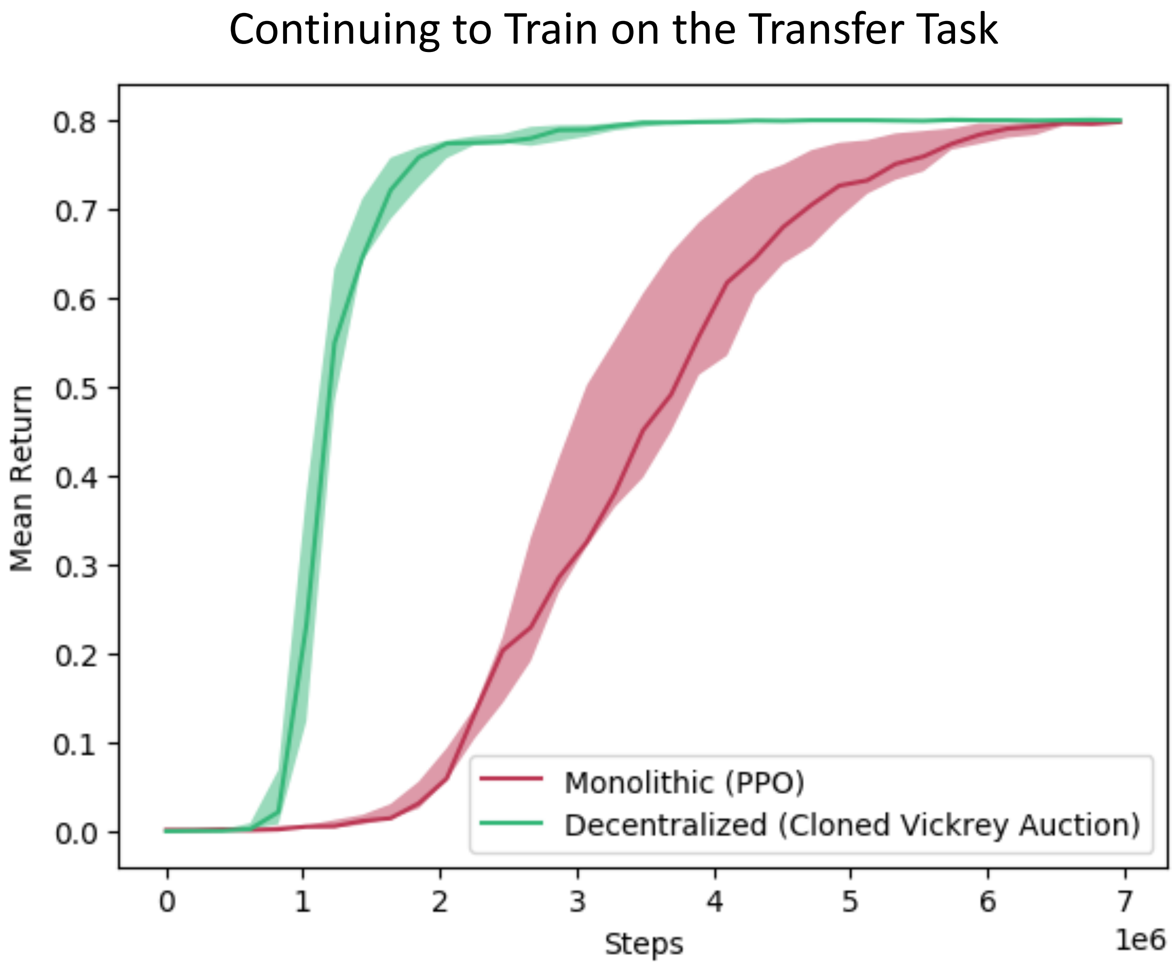
Transfer Task

Transfer

Optimal Policy for the Society



Transfer


Continuing to Train on the Transfer Task
Mean Return
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
4
5
6
7
Steps
1e6
Monolithic (PPO)
Decentralized (Cloned Vickrey Auction)

Contributions

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q^*(s_t, \omega^i)}$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	**Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

https://sites.google.com/view/clonedvickreysociety

Contributions

Cloned Vickrey Society

A society of agents that implements global decision making via local economic transactions.

Question	Key Idea
What should the optimal bids be for the solution of the Global MDP to emerge?	Define the optimal bid as the **optimal Q value** $\boldsymbol{Q^*(s_t, \omega^i)}$ for activating agent ω^i at state s_t.
For what auction mechanism would these optimal bids be an equilibrium strategy?	By defining the agents' valuations $v^i(s)$ as $Q^*(s, \omega^i)$, under the Vickrey auction it is a **dominant strategy** to truthfully bid $Q^*(s, \omega^i)$.
How can we adapt this auction mechanism for discrete-action MDPs?	**Temporally couple the agents in a market**: An agent's valuation of s_t is defined by how much it can sell the product s_{t+1} of executing its transformation on s_t.
How can we avoid suboptimal equilibria?	**Redundancy enforces credit conservation** that helps avoid suboptimal equilibria.
How can we translate the auction mechanism into a decentralized reinforcement learning algorithm?	Define the **auction utility** as the agents' reinforcement learning objective, yielding a **decentralized reinforcement learning algorithm** for the Global MDP.

https://sites.google.com/view/clonedvickreysociety