# A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings

## ——MARL论文分享

林越 2022.02.15

DeepMind

*October 2021*

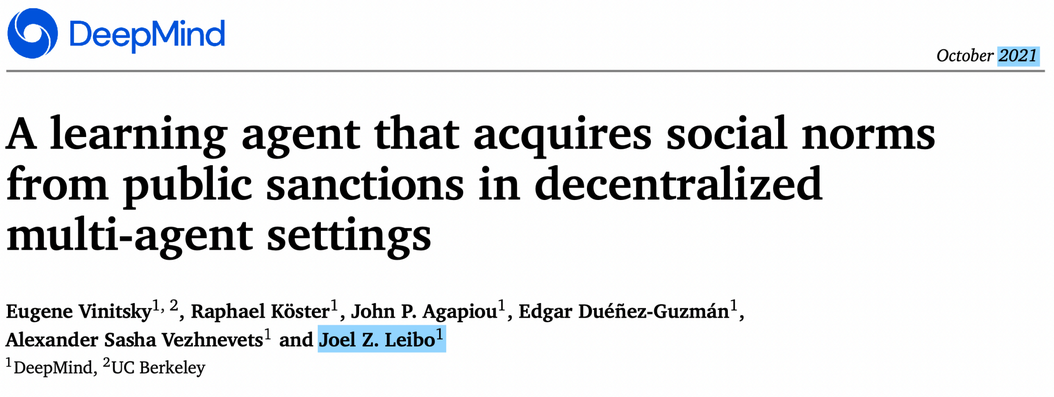
# A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings

**Eugene Vinitsky**[1, 2], **Raphael Köster**[1], **John P. Agapiou**[1], **Edgar Duéñez-Guzmán**[1], **Alexander Sasha Vezhnevets**[1] and **Joel Z. Leibo**[1]

[1]DeepMind, [2]UC Berkeley

# 快速介绍

## 面向老师和学长


DeepMind

October 2021

A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings

Eugene Vinitsky[1, 2], Raphael Köster[1], John P. Agapiou[1], Edgar Duéñez-Guzmán[1], Alexander Sasha Vezhnevets[1] and Joel Z. Leibo[1]
[1]DeepMind, [2]UC Berkeley

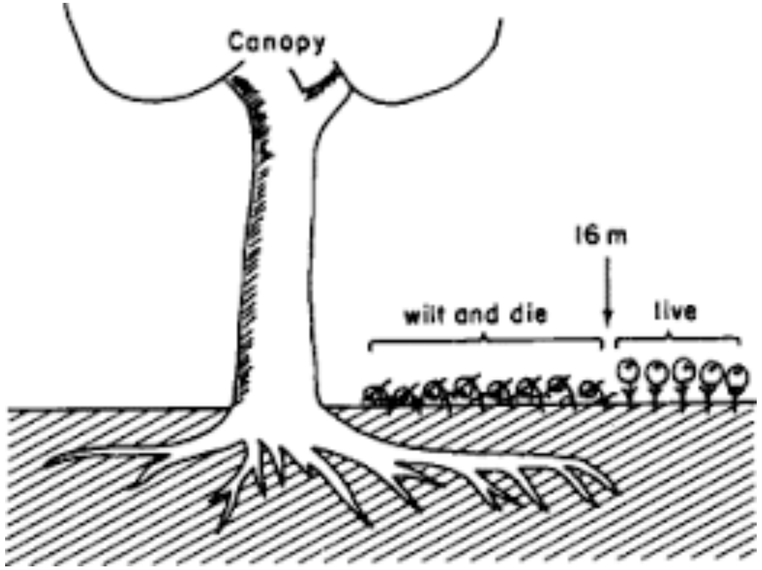
- Decentralized：Agent各自用不同的A3C（没有总体奖励或策略共享）
- Env：昨天乔丹学长提到的Allelopathic Harvest和Cleanup with Start-up Problem
- 惩罚机制Sanction：是一个动作，有机会制裁身边的人，制裁后冻结对方
- Sanction是唯一的public signal（没有群体奖励的设置，没有policy-sharing），每个人可以看到上一帧大家都是怎么选择制裁别人的（相当于补充了observation）
- Social Norm：二分类器classifier，何时应制裁别人，总结自Agents的制裁决定
- Social Norm影响Agent决定（classifier的输出结果传给observation）
- 边总结边影响，是一种演化的过程
- 演化最后出现了Social Norm，并且作者讨论了在一些情况下Social Norm是有益的

# Game 1

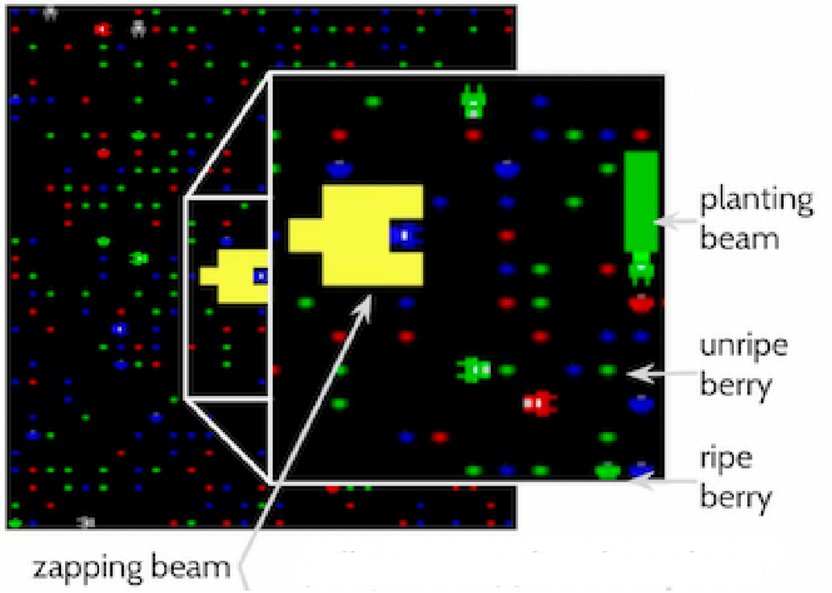
## Allelopathic Harvest

- Allelopathy is defined as the **effects (stimulatory and inhibitory) of a plant on the development of neighboring plants through the release of secondary compounds.**

- 植物通过释放secondary compounds来影响（刺激性和抑制性）邻近植物的发育

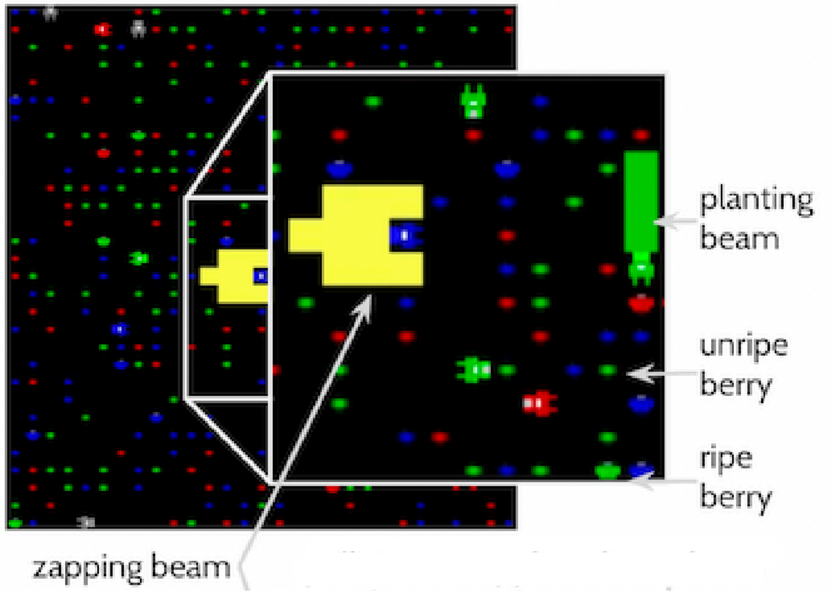


# Game 1
## Allelopathic Harvest




- Env
  - Grid World：(30, 29)个cell，每个cell (8, 8)像素
  - 超出屏幕边界，则到另一边，和一些版本的贪吃蛇一样，移动是不受限制
  - 16 Agents，各自用各自策略，一开始随机在出生点位置初始化（这些出生点在哪里则是固定的）
  - 870个格子中一共有384个固定的位置可以种浆果（每个位置都可种3种颜色）
  - 有3种颜色的浆果， 每个浆果品种的生长速度，线性地取决于该颜色占总数的分数
    - 每回合有(0.0000025*该颜色浆果当前数量/384)的概率生成浆果（Allelopathic的含义）
    - 被吃掉后，要过10回合才能开始计算这个概率
    - 如果有Agent站在上面，则不计算这个概率

# Game 1
## Allelopathic Harvest

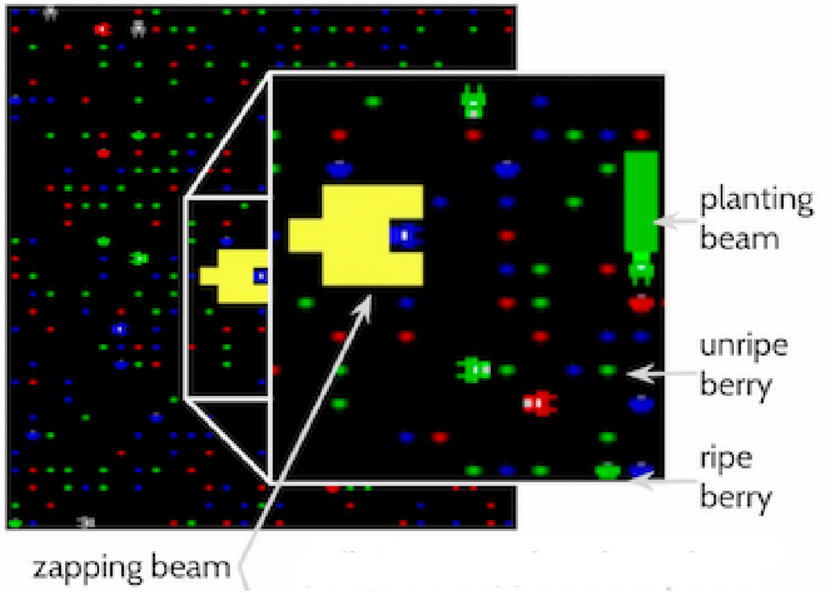



- Agent：
  - 移动：4种朝向，移动是ULRD，不同朝向的ULRD的结果不同
  - 动作：转向、移动（经过浆果就会吃掉）、制裁、种植
  - 观测：以自己为中心的(88, 88, 3) RGB图像（前9后1左右5）、Classifier的预测值（if used）
  - 可在自己面前选择种植哪种颜色的浆果：发射激光（2回合冷却），前方3格，会被阻挡，将其变为想要的颜色
  - 有一半Agent（8个）吃红色浆果的奖励（reward=2）是他们吃别的颜色浆果奖励（reward=1）的2倍；另一半也一样，但他们更喜欢吃绿色浆果
  - 做过什么，自己就会变色，并且别人看得到（相当于是一种简单的reputation标志？）
    - 一开始默认是灰色；如果吃了浆果，有(1-m)的概率变回灰色，如果产量多，变灰的概率小（Monoculture Fraction）
    - 种植了什么颜色就变成什么颜色

$$m = \max\{\frac{r}{r+g+b},, \frac{g}{r+g+b}, \frac{b}{r+g+b}\}$$

# Game 1

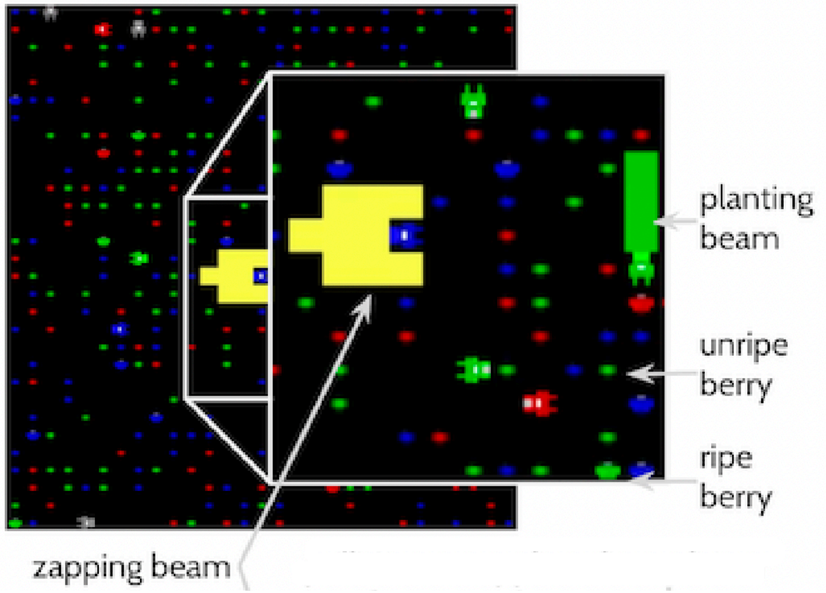
## Allelopathic Harvest




- 制裁的设置
  - 可以发射黄色激光制裁别人，激光长度是面前3格和身旁往前3格
  - 用制裁是一个动作，在一个回合内如果选择制裁了就不能移动
  - 如果被制裁了，就冻结25个steps不动，并且被打上标记
  - 如果被标记的Agent又被制裁了，那么其直接被removed并-10 reward
  - 如果被标记的Agent在其后50步内没被制裁，那么标记消失

# Game 1

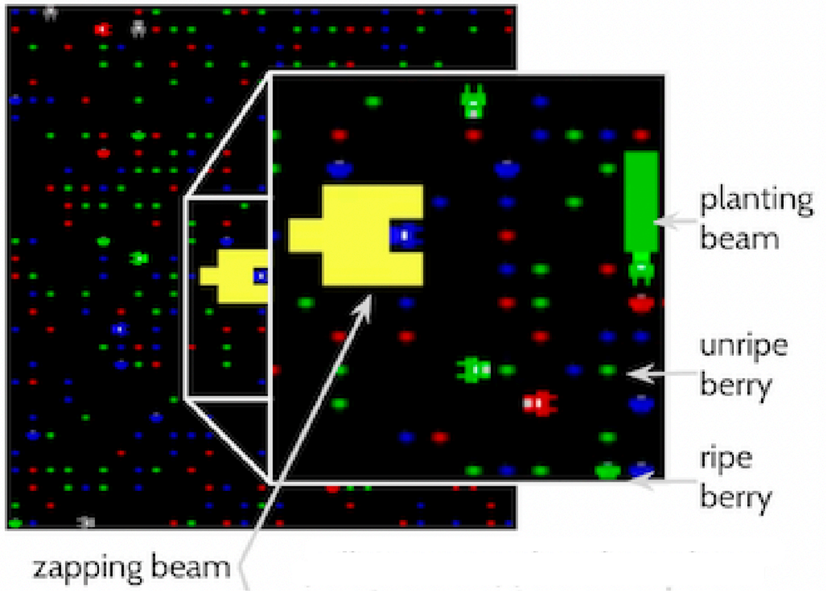
## Allelopathic Harvest




- 问题1：start-up problem
  - Before much learning has occurred, very few individuals work consistently toward any goal so defection is motivated by fear that too few others will contribute to successfully establish any norm.
  - Agent能通过只选择种植一种颜色的浆果，从而拿到更高的奖励
  - 但是他要选择哪一种颜色是很难协调的
  - 我的理解：每个人都怕建立不起来Norm，都等着别人先种浆果，自己先defect先只吃

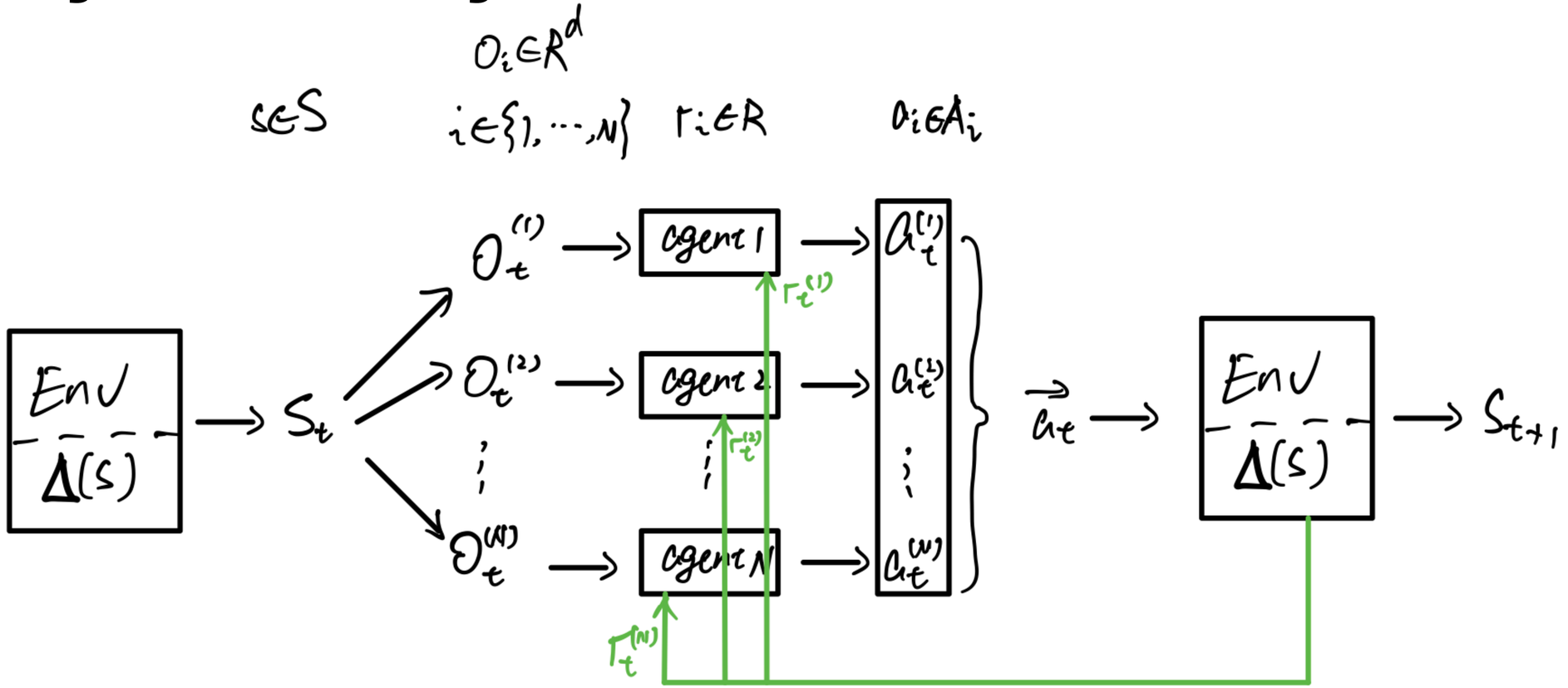
# Game 1

## Allelopathic Harvest




- 问题2：free-rider problem
  - In the later phase of learning, when most individuals are engaged, then the motivation to defect is greed since one can free-ride on the efforts of others.
  - 偷懒，自己从来不种，只吃
  - They also always prefer to eat berries over spending time planting.

# N-Player Partially Observed Decentralized Settings




transition $T$ : $S \times A_1 \times A_2 \times \cdots \times A_N \to \Delta(S)$

reward function $R$ : $S \times A_1 \times A_2 \times \cdots \times A_N \times S \times I \to R$

# Game Theory

- 每个人如果只注重优化自己能拿到的利益，能达到社会利益更大化吗?
  - 不一定，Social Dilemma就是表现了这种冲突的模型
  - 比如囚徒困境
    - Defect是占优策略
    - 但是一起Cooperate能带来更多总体效益
- 分析这个问题的数学工具是Game Theory

| P1 \ P2 | C | D |
|---|---|---|
| C | 3, 3 | 0, 4 |
| D | 4, 0 | 1, 1 |

# Social Norm

- Social norms: collective patterns of sanctioning that can prevent miscoordination and free-riding.
- 我的理解：
  - Social norm是一种标准，根据一个人以往的行为来评价这个人
    - reputation, indirect reciprocity, one-shot
  - 在这篇文章里它规定了做出了什么样行为的人应该被制裁
  - miscoordination: 多个equilibria的情况
  - free-riding

# Sanction

## Introduction

- 惩罚能加强合作
- 直观的理解：如果选择Defect，其获得的收益减去被罚去的收益（或者限制其行动以减小其收益），这样之后的收益期望要小于选择Cooperate的收益期望，那还不如选择Cooperate
- If the costs of being punished are large enough, moralistic strategies which cooperate, punish noncooperators, and punish those who do not punish noncooperators can be evolutionarily stable.
- Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol. Sociobiol. 13, 171–195. (doi:10.1016/0162-3095(92)90032-Y)

# Sanction

## Sanction Opportunity

- 制裁是一个个体对另一个个体的，但是每个个体是如何制裁别人的，会在下一状态被所有人知道（个体进行决定，其决定被广播，强调是decentralized）
- Agent_i对Agent_j有没有制裁的机会，取决于Agent_j在不在Agent_i身边的一定范围内
- 表示有制裁机会的集合是关于状态s的函数 $\mathcal{J}(s) \subseteq I^2$
- 这个集合是一个二元组的集合：{(1, 2), (4, 3), …, (i, j), …}
- 如果(i, j)在这个集合里，说明agent_i可以选择制裁或者不制裁agent_j

# Sanction

## the Context of Sanctioning Opportunity

- 语境表示的是，一个Agent选择制裁或者不制裁别人的依据（他看到了什么，做过什么）
- $C(s_t, i, j) = (o_{0:t}^{(i)}, a_{0:t-1}^{(i)})$
- $C(s_t, i, j) = o_t^{(i)}$
- 论文里的例子：一个小孩偷吃了巧克力，如果你遇到他的时候看到他脸上都是巧克力，我就可以惩罚他了，不需要看到他偷吃的行为（这里好像不对）
- 在实验里相应的操作是，Agent会根据自己最近的行为改变自己的颜色
- Agent_i能看到Agent_j的颜色

# Sanction

## Social Norms

- 一个Agent有制裁别人的机会后，其可以选择制裁或者不制裁
- Social Norm这里指的是，选择制裁或者不制裁的标准
  - 有制裁的机会，且看到别人是红色，就选择制裁别人
  - 有制裁的机会，但看到别人是绿色，就选择不制裁别人
- 制裁或不制裁的表示：$\mathcal{Z}(s, \vec{a}, i, j)$
  - 如果其为0，则表示在s状态下群体作出联合动作a，Agent_i不制裁Agent_j
  - 如果其为1，则表示在s状态下群体作出联合动作a，Agent_i制裁Agent_j

# Sanction

## Sanction-Observation Function

- A sanctioning observation $g \in \mathcal{G}$
- $\mathcal{B} : \mathcal{S} \rightarrow \mathcal{G}$
- $\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) | (i, j) \in \mathcal{J}(s_{t-1}) \wedge c$
  $= C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$

可选择是否制裁j的(i,j)组合

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) | (i, j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j)\}$$

基于什么观测(语境)

而作的是否制裁决定

在 $s_{t-1}$ 和 $\vec{a}_{t-1}$ 状态下

最终是否制裁j了?

# Sanction

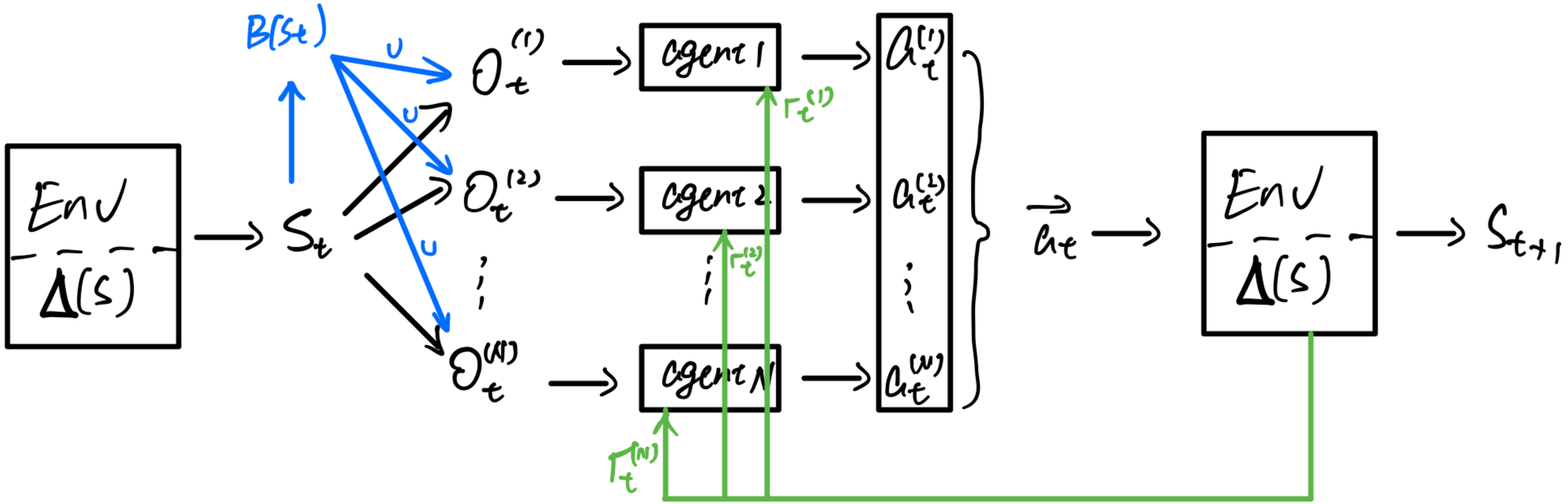
## Sanction-Observation Function

- A sanctioning observation $g \in \mathcal{G}$
- $\mathcal{B} : \mathcal{S} \rightarrow \mathcal{G}$
- $$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) | (i, j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$$
- 通过增加state以包含先前的observations，可以将其化简成$\mathcal{B}(s_t)$（原文的话）
- 我的理解：t-1时候做的动作会改变Agent的颜色，在t时刻外加观察一项颜色即可

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) | (i, j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$$

# Sanction

## Broadcasting Global Information on Sanctioning




- 每个人都知道：所有人中有权利制裁别人的那些人，在这个状态的语境下，最终有没有选择制裁别人

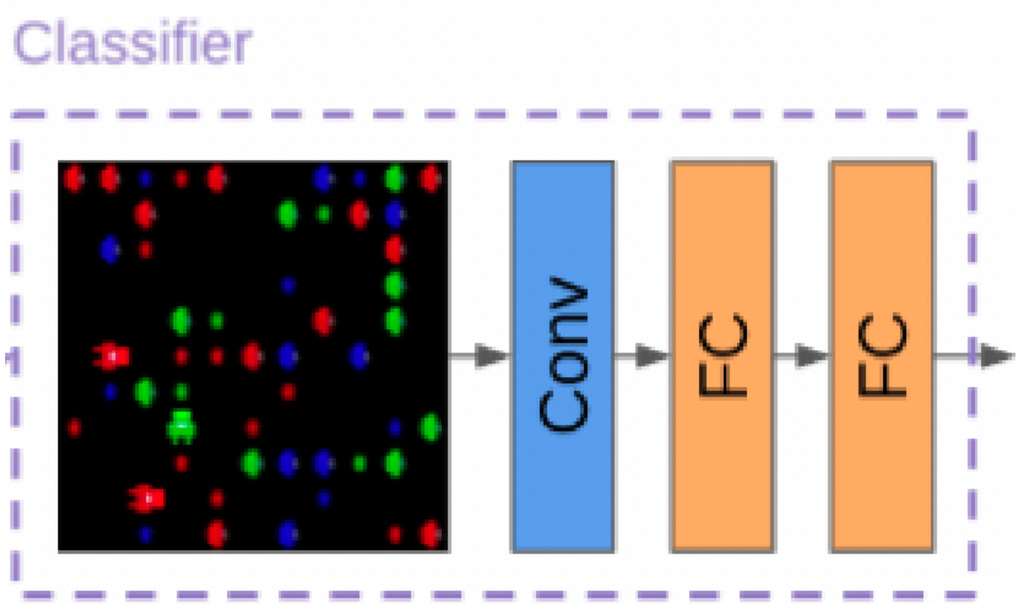
# Classifier $\Psi_\phi$

- We learn the **social norm** by training a classifier on the public sanctioning observation provided by $\mathcal{B}(s_{0:T})$
- Assuming the size of $\mathcal{T}(s)$ is M
- 输入语境，输出惩罚的概率
- Loss: binary cross-entropy loss $$\mathcal{L}_\phi = \frac{1}{M} \sum_{c,z \in \mathcal{B}} -z \log\left(\Psi_\phi(c)\right) - (1-z) \log\left(1 - \Psi_\phi(c)\right)$$
- Minimize this loss using Stochastic Gradient Descent
- 使classifier产生的分布接近z的分布，学习z的分布
- 大多数人做的这个映射关系就成了Social Norm

# Classifier $\Psi_\phi$

## Architecture

- The classifier consists of a convolutional backbone attached to a multi-layer perceptron (MLP)



# Classifier $\Psi_\phi$

## Potential Challenges

- A key issue
  - if a particular agent behavior is effectively suppressed,
    - the classifier will no longer receive samples of approval / disapproval of that behavior,
    - and unlearn its prior pattern of disapproval.
  - we **stop classifier learning** by setting the learning rate of the classifier to zero after some fixed number of time-steps.
    - This freezes the group norm as observed by the classifier at that point in time but does not prevent subsequent drift in agent behavior.

# Learning Polices

## An Intrinsic Motivation

- The core idea of the **CNM (Classifier Norm Model)** agent is that an individual embedded in a wider group **is motivated to sanction in accord with the group's joint pattern** of approval and disapproval.
- a_t是disapproval的意思是，Agent有机会惩罚别人且已经选择惩罚别人了
- 如果我惩罚了别人，并且这是符合社会规范的，那我加分
- 如果我惩罚了别人，但社会规范要求在这种情况下不该惩罚别人，那么我扣分

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } \alpha, \beta \in \mathbb{R}_0^+$$

# Learning Polices

## An Intrinsic Motivation

- a_t是disapproval的意思是，Agent有机会惩罚别人且已经选择惩罚别人了

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } \alpha, \beta \in \mathbb{R}_0^+$$

$$V_{\theta,\phi}^{\vec{\pi}}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, \vec{a}_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)}) \middle| \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right]$$

# Learning Polices

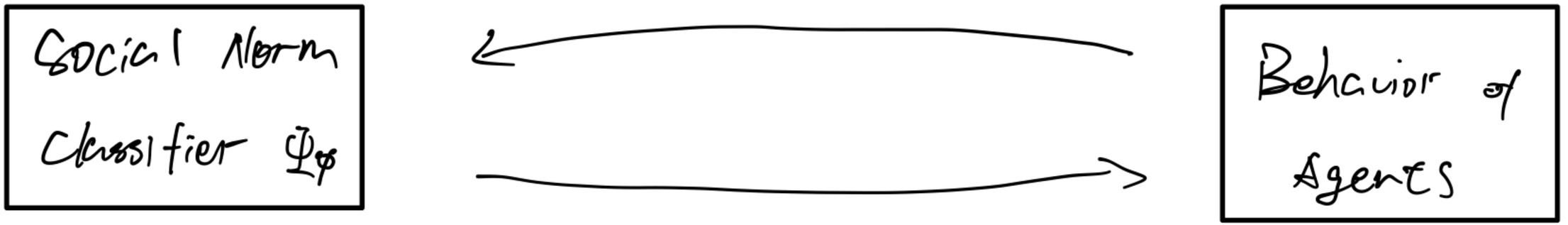
## Decentralized Learning

- Each agent $i$ learns a parameterized behavior policy that is conditioned solely only the history of its own individual observations and actions and its estimate of the social norm
- $\pi_\theta(a_t^{(i)} | o_{0:t}^{(i)}, a_{0:t-1}^{(i)}, p_t)$
  - $p_t = \text{stop}\left(\mathbb{1}[\Psi_\phi(o_t) \geq 0.5]\right)$
  - $stop(\cdot)$ is the stop gradient operator，优化π时不计算这梯度
- The RL algo used for each agent is A3C with a V-Trace loss for computing the advantage
  - For more details, refer to the Appendix.

# 两种学习

学习总结社会规则

$\Psi_\phi(c) \to \geq$



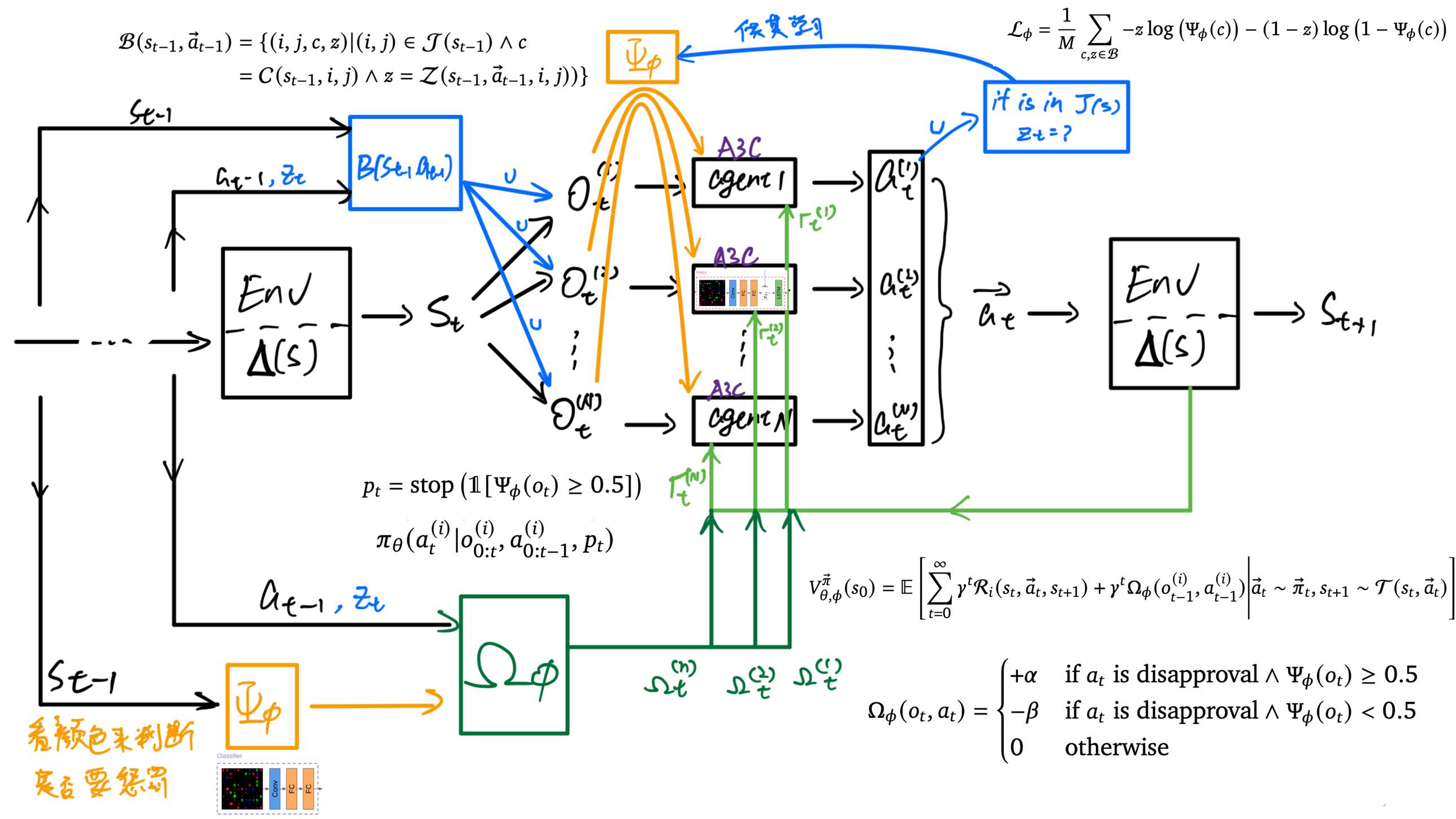

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) < 0.5 \\ 0, & \text{otherwise} \end{cases}$$

$$V(s_0) = E\left[\sum_{t:0}^{\infty} \gamma^t R_i(s_t, a_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)})\right]$$

学习如何惩罚

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i,j,c,z) | (i,j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$$

$$\mathcal{L}_\phi = \frac{1}{M} \sum_{c,z \in \mathcal{B}} -z \log\left(\Psi_\phi(c)\right) - (1-z) \log\left(1 - \Psi_\phi(c)\right)$$

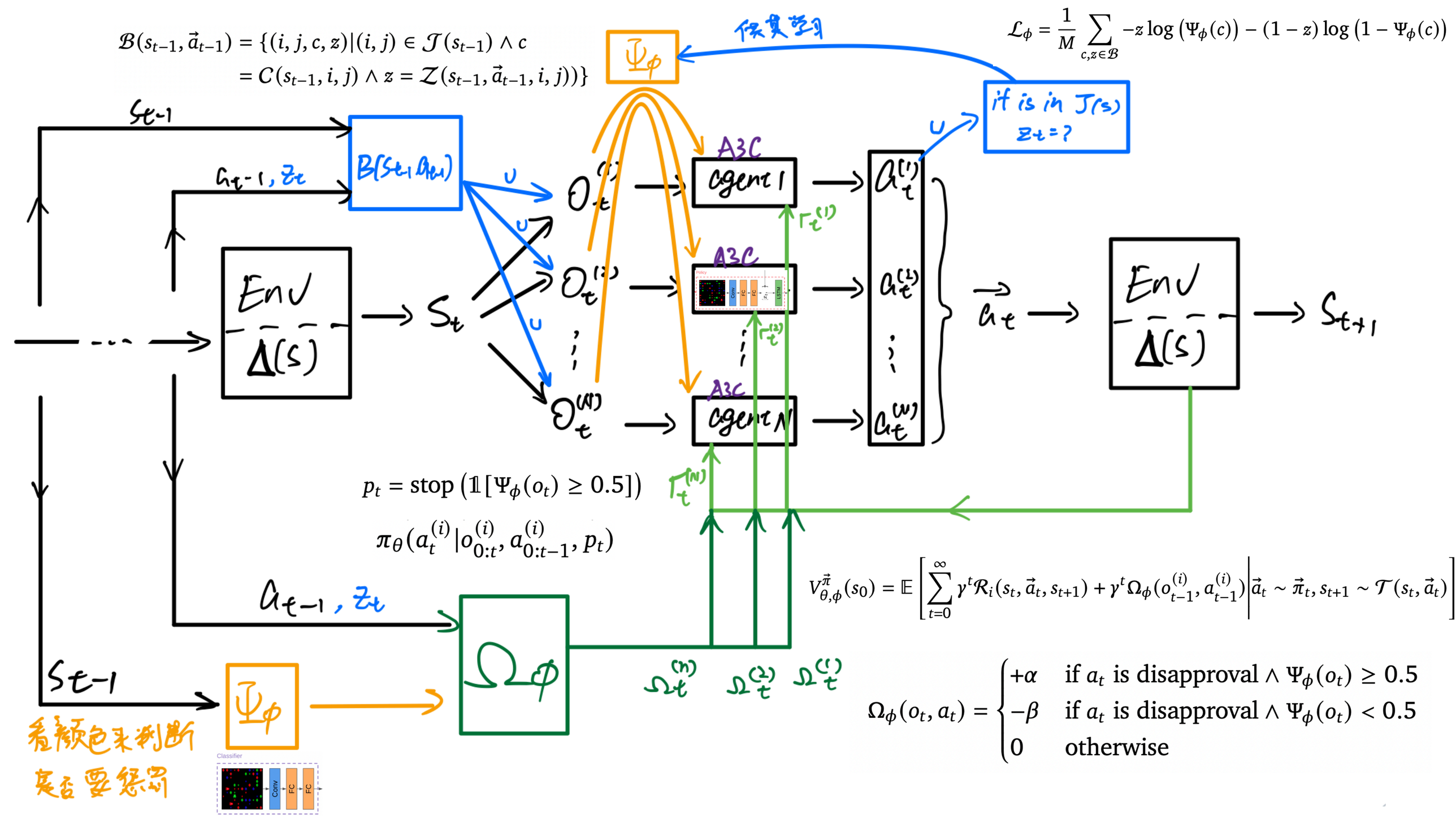



$$V_{\theta,\phi}^{\vec{\pi}}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, \vec{a}_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)}) \middle| \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right]$$

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i,j,c,z) | (i,j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$$

$$\mathcal{L}_\phi = \frac{1}{M} \sum_{c,z \in \mathcal{B}} -z \log\left(\Psi_\phi(c)\right) - (1-z) \log\left(1 - \Psi_\phi(c)\right)$$

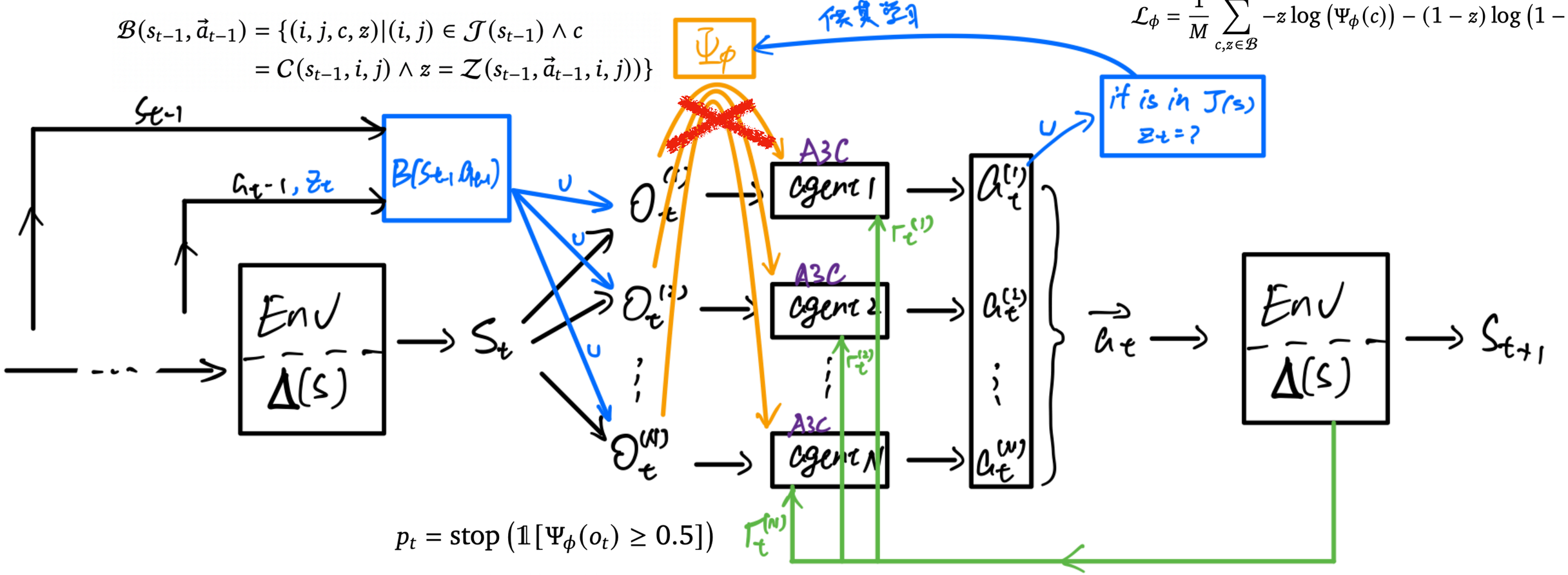



$$V_{\theta,\phi}^{\vec{\pi}}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, \vec{a}_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)}) \middle| \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right]$$

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) | (i, j) \in \mathcal{J}(s_{t-1}) \wedge c = C(s_{t-1}, i, j) \wedge z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j))\}$$

$$\mathcal{L}_\phi = \frac{1}{M} \sum_{c,z \in \mathcal{B}} -z \log\left(\Psi_\phi(c)\right) - (1 - z) \log\left(1 - \Psi_\phi(c)\right)$$




$$p_t = \text{stop}\left(\mathbb{1}[\Psi_\phi(o_t) \geq 0.5]\right)$$

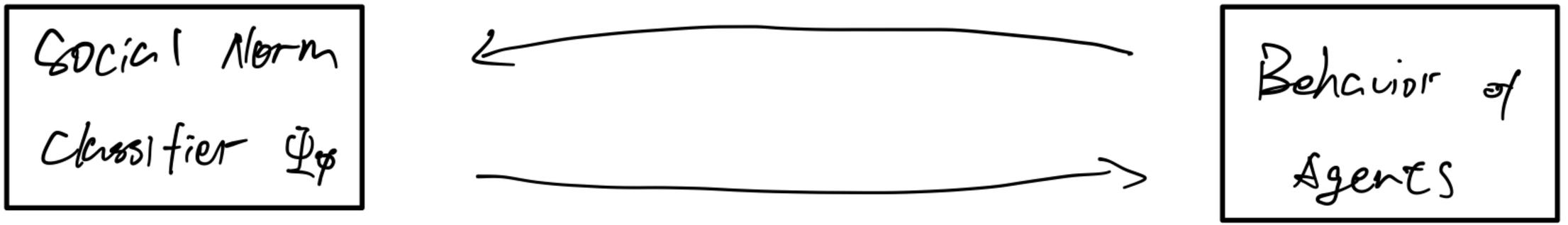
$$\pi_\theta(a_t^{(i)} | o_{0:t}^{(i)}, a_{0:t-1}^{(i)}, p_{\cdot})$$

$$V_{\theta,\phi}^{\vec{\pi}}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, \vec{a}_t, s_{t+1}) + \gamma^t \Omega(a_{t-1}^{(i)}, a_{t-1}^{(i)}) \middle| \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right]$$

# 两种学习

学习总结社会规则

$\Psi_\phi(c) \to \geq$




$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) < 0.5 \\ 0, & \text{otherwise} \end{cases}$$

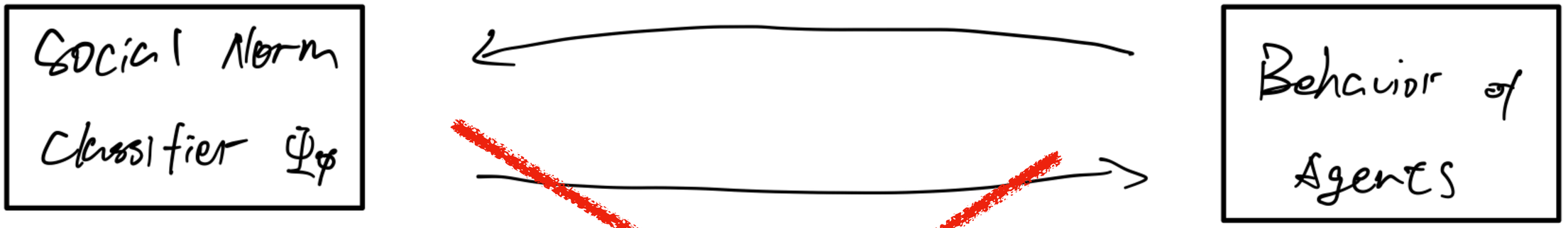
$$V(s_0) = E\left[\sum_{t:0}^{\infty} \gamma^t R_i(s_t, a_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)})\right]$$

学习如何惩罚

# 两种学习

学习总结社会规则

$\Psi_\phi(c) \rightarrow \geq$

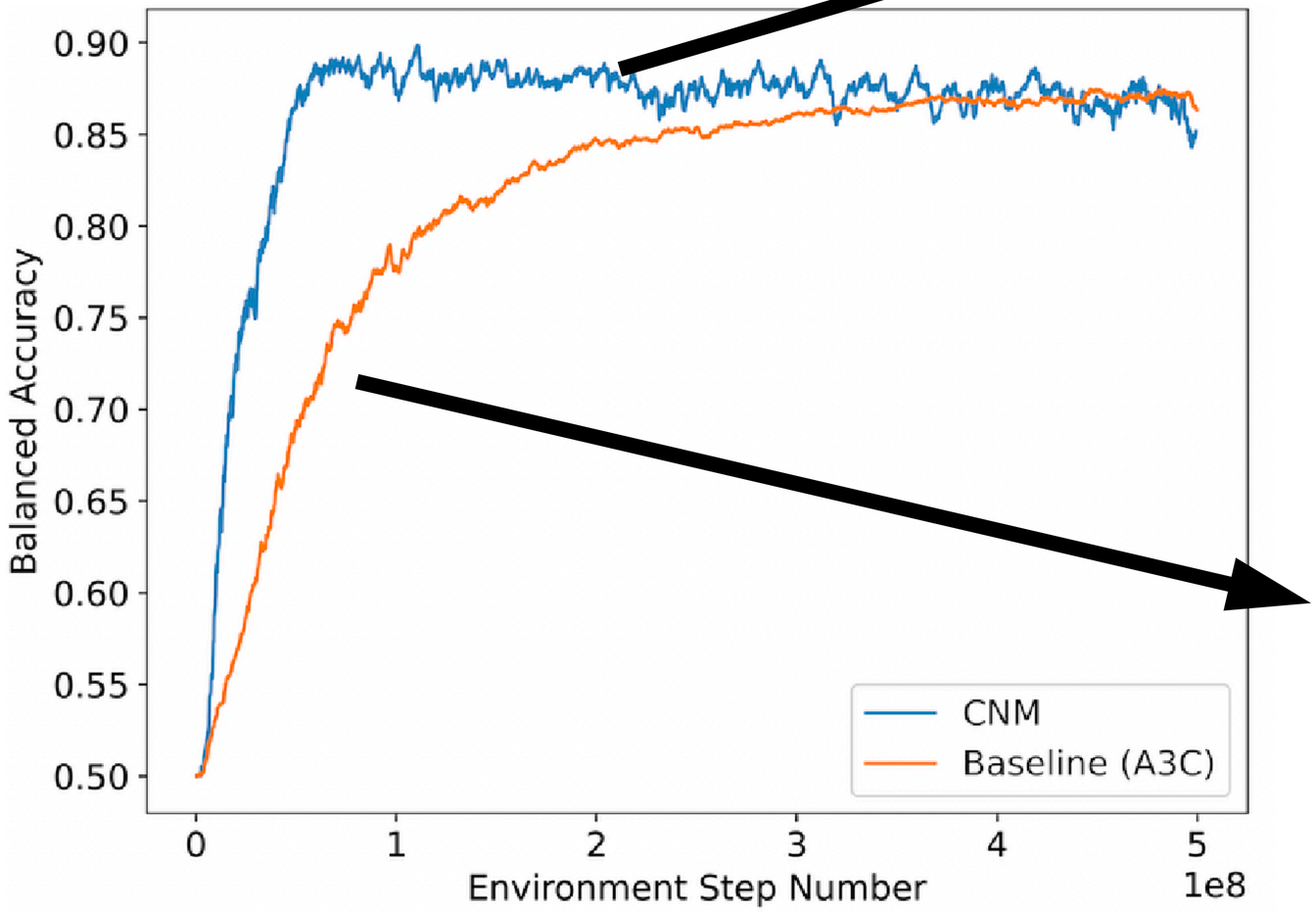


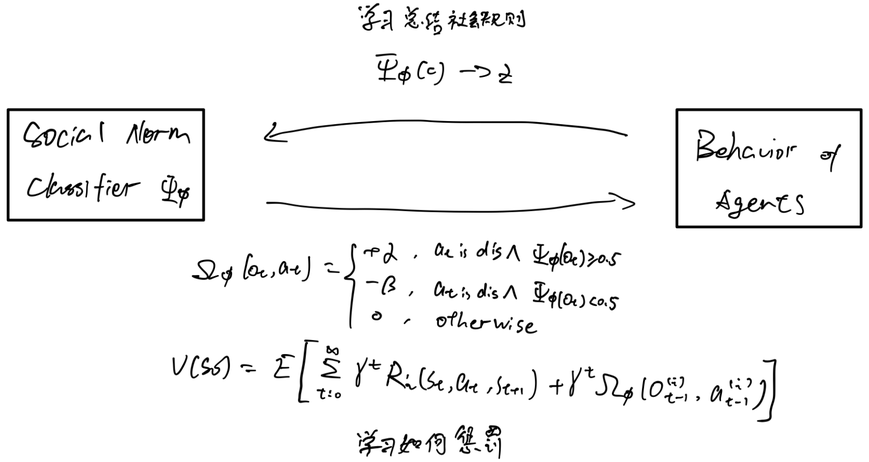

$$\Omega_\phi(o_t, a_t) = \begin{cases} r_0, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta, & a_t \text{ is dis} \wedge \Psi_\phi(o_t) < 0.5 \\ 0, & \text{otherwise} \end{cases}$$

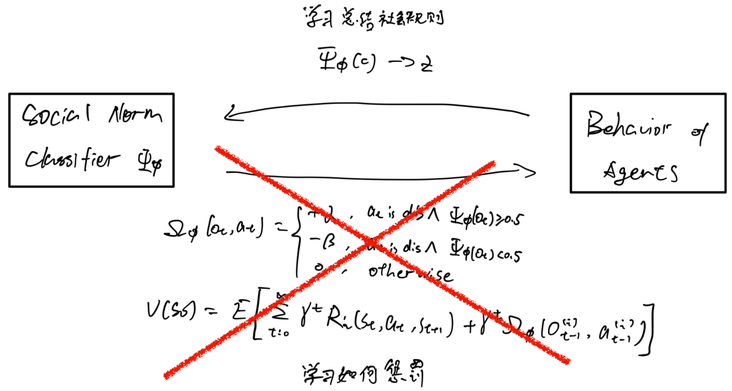
$$V(s_0) = \mathbb{E}\left[\sum_{t:0}^{\infty} \gamma^t R_i(s_t, a_t, s_{t+1}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)})\right]$$

学习如何惩罚

# Experiments



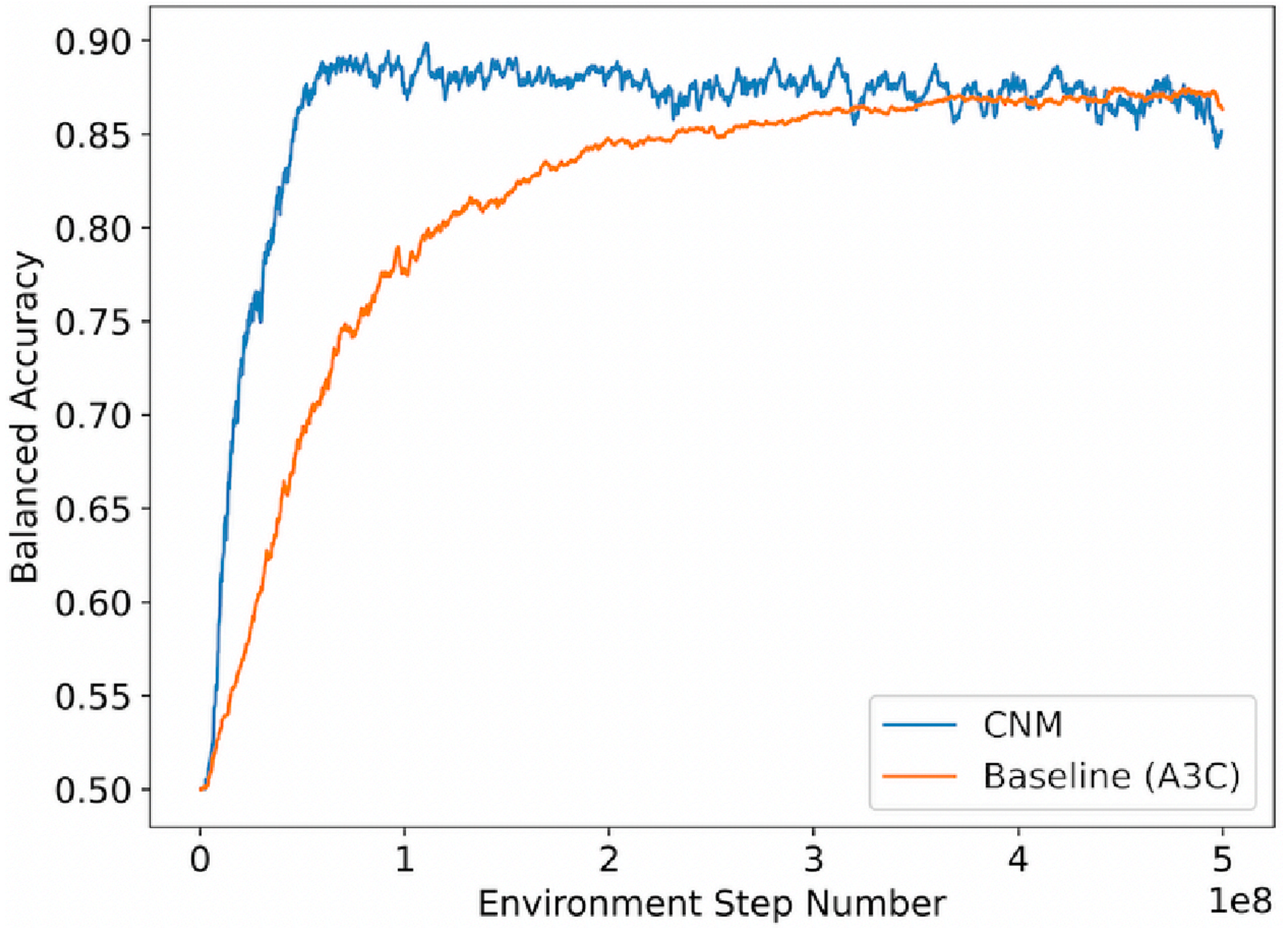




- 相当于只是总结学习SN，不影响行为

# Experiments

## Existence of the Emergent Social Norm

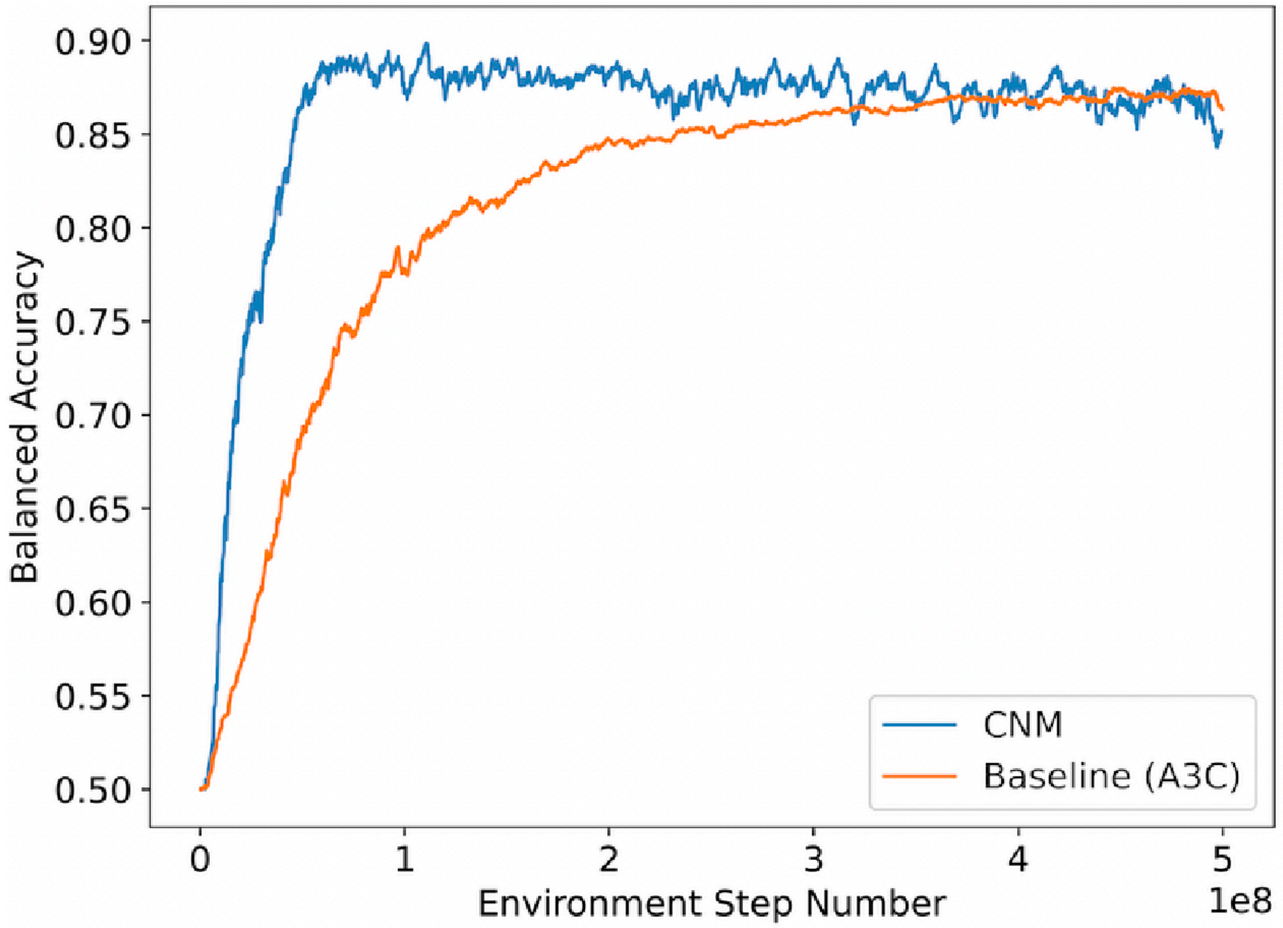



## 结论1

- 这样的设置下，classifier可以学到很高的balanced accuracy
- 我对balanced accuracy的理解：两个互相学习，互相影响，classifier总结惩罚的规律，agent又根据这个来做出动作，这样演化后会达到一种平衡关系
- 就算只有一帧作为语境也有很好效果，说明最开始的规范的行为可能很简单，比如：“制裁那些可能和我竞争的人”或“制裁某种特定颜色的人”

# Experiments

## Existence of the Emergent Social Norm

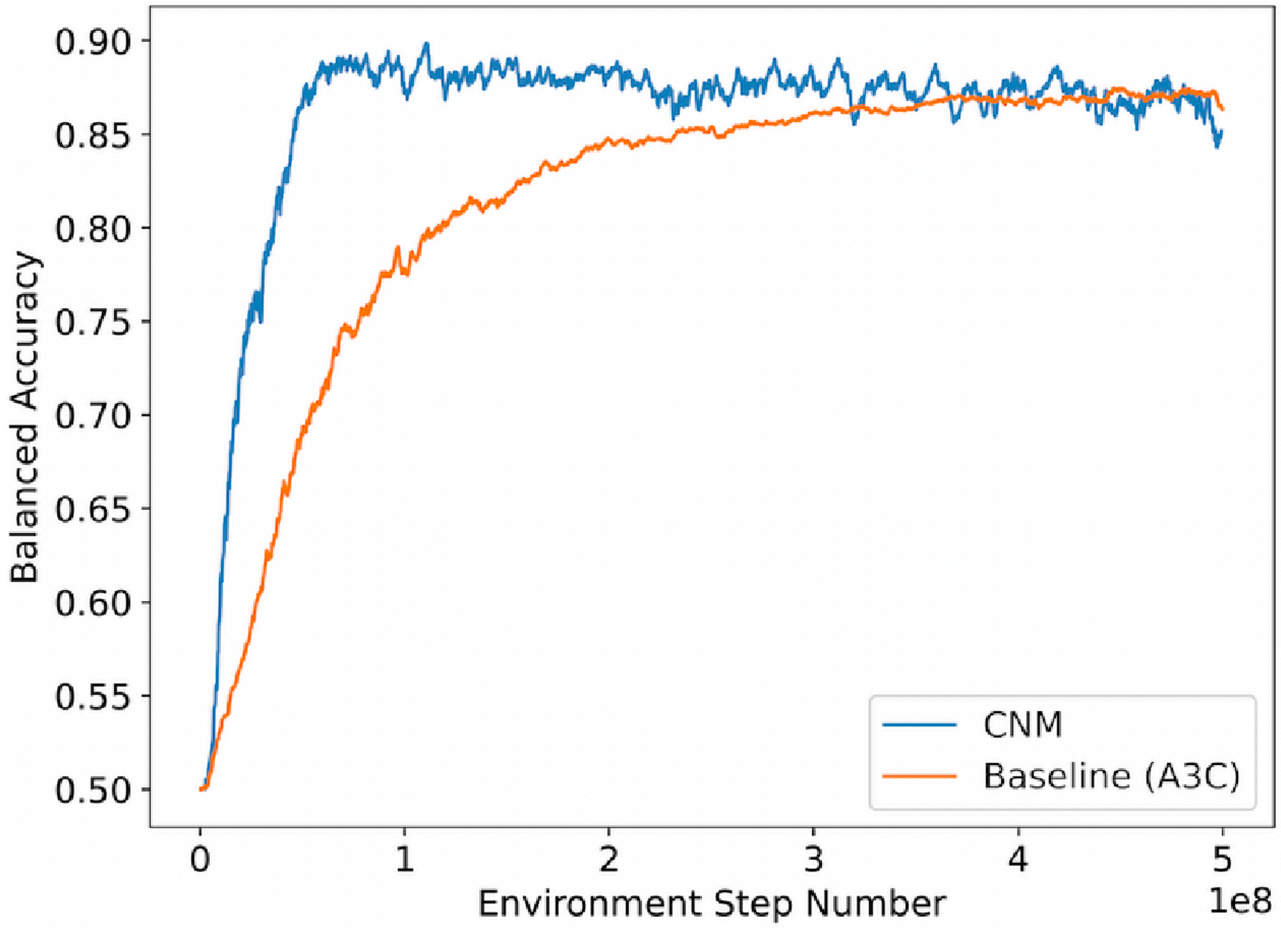



## 结论2

- The intrinsic motivation to punish in accord with the classifier's predictions causes its accuracy to rapidly converge.
- 如果我惩罚了别人，并且这是符合社会规范的，那我加分
- Agent会趋于迎合Social Norm
- 我的理解是，要达到平衡状态，那么像设置那样双向奔赴是更好的更快能收敛的

# Experiments

## Existence of the Emergent Social Norm




## 结论3

- 在5e7步后冻结classifier，不再更新里面的参数
- 尽管如此，训练期间的balanced accuracy还是很高
- 固定下来Social Norm后，Agent的行为对于已有的Social Norm不会发生太大偏移

# Experiments

## Beneficial Effects of the Emergent Social Norms

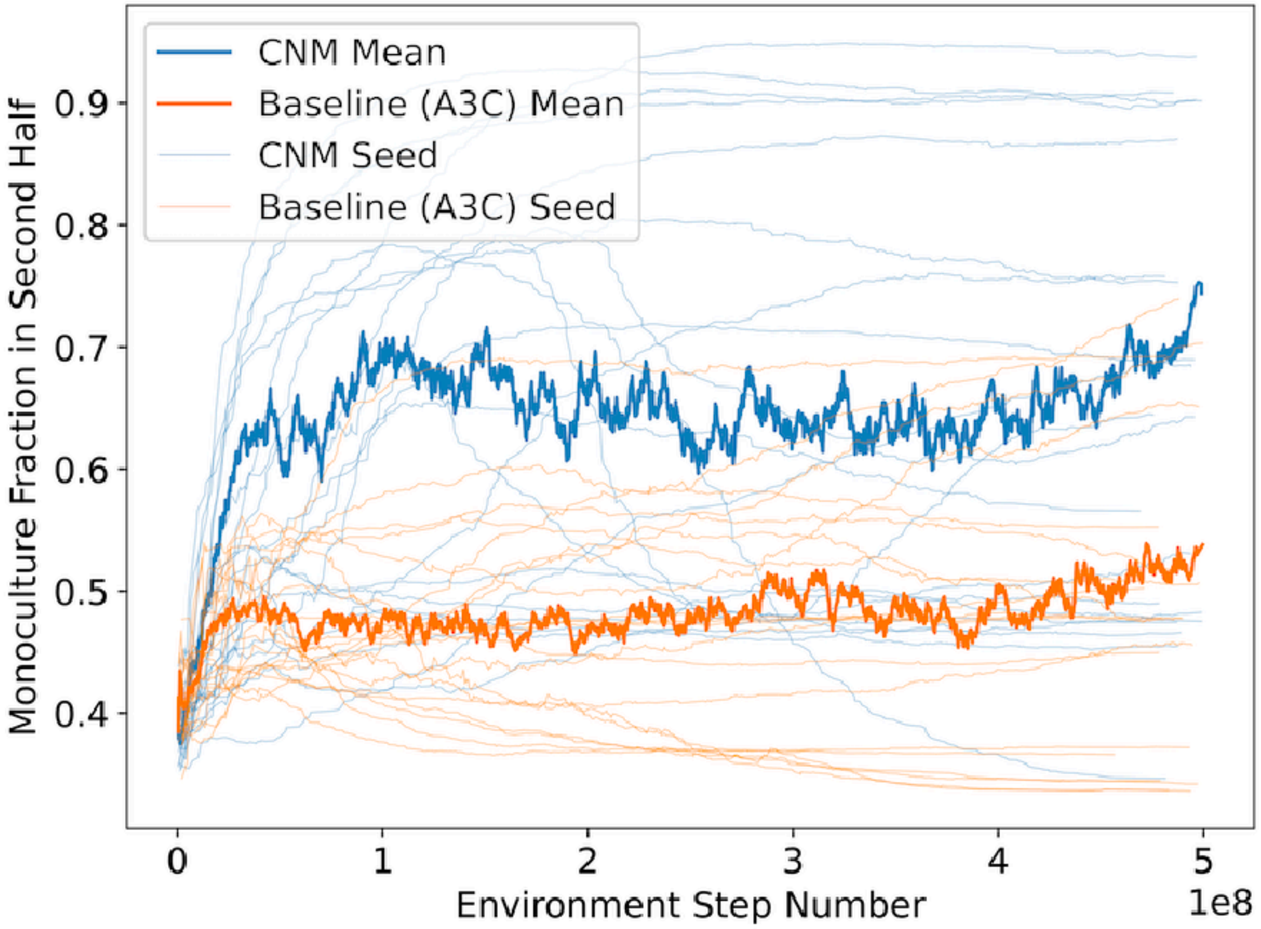
- Whether the use of **CNM (Classifier Norm Model)** leads to better outcomes
- The effect of norms on avoiding startup problems and overcoming free-rider problems.
- Run 20 seeds

$$m = \max\{\frac{r}{r+g+b},, \frac{g}{r+g+b}, \frac{b}{r+g+b}\}$$

- The measure of success is the **monoculture fraction**, the percentage of the color that corresponds to the largest number of berry spawning sites.
- Agent能通过只选择种植一种颜色的浆果，从而拿到更高的奖励

# Experiments

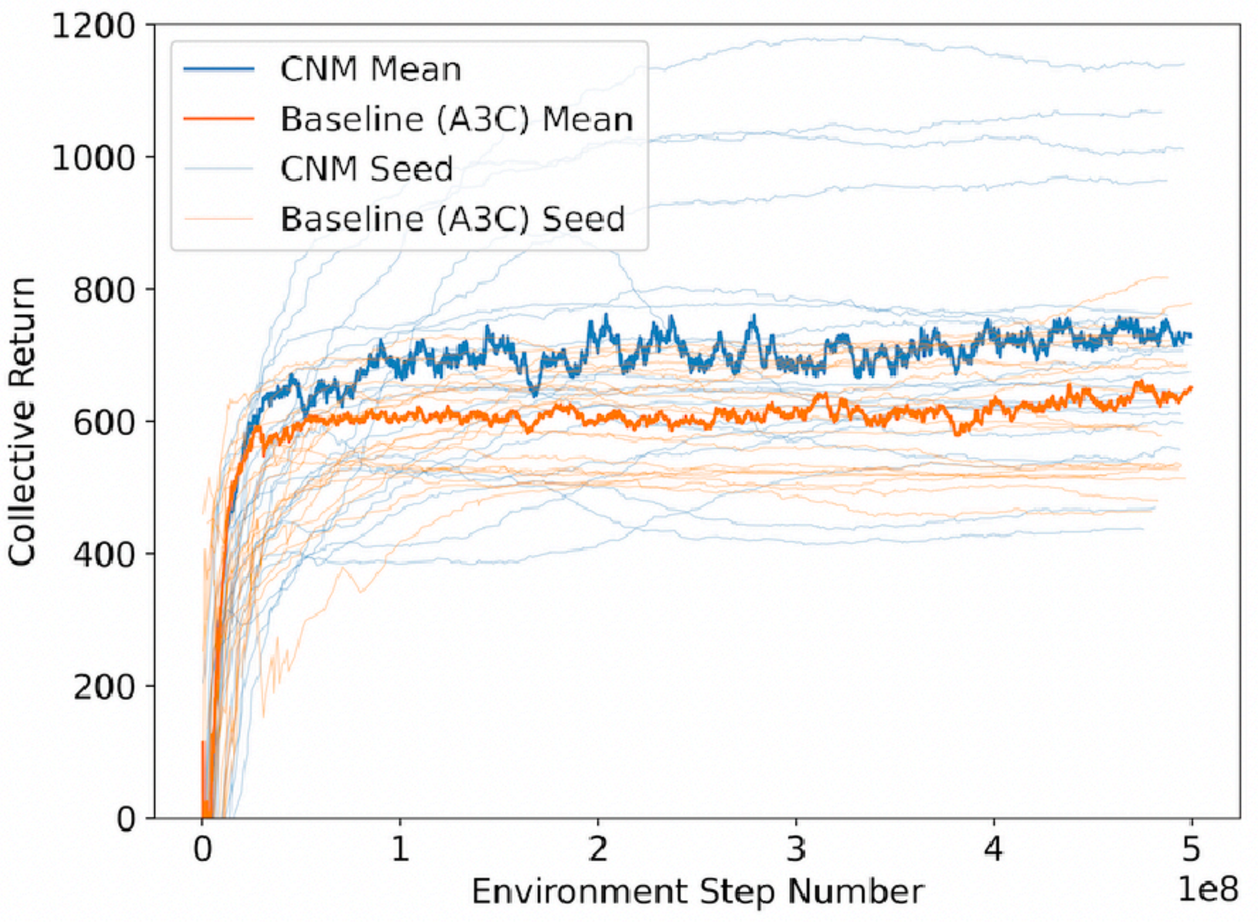
## Beneficial Effects of the Emergent Social Norms




- CNM increases the monoculture fraction above 50%
- indicating that agents on average are converging to a single preferred color

# Experiments

## Beneficial Effects of the Emergent Social Norms

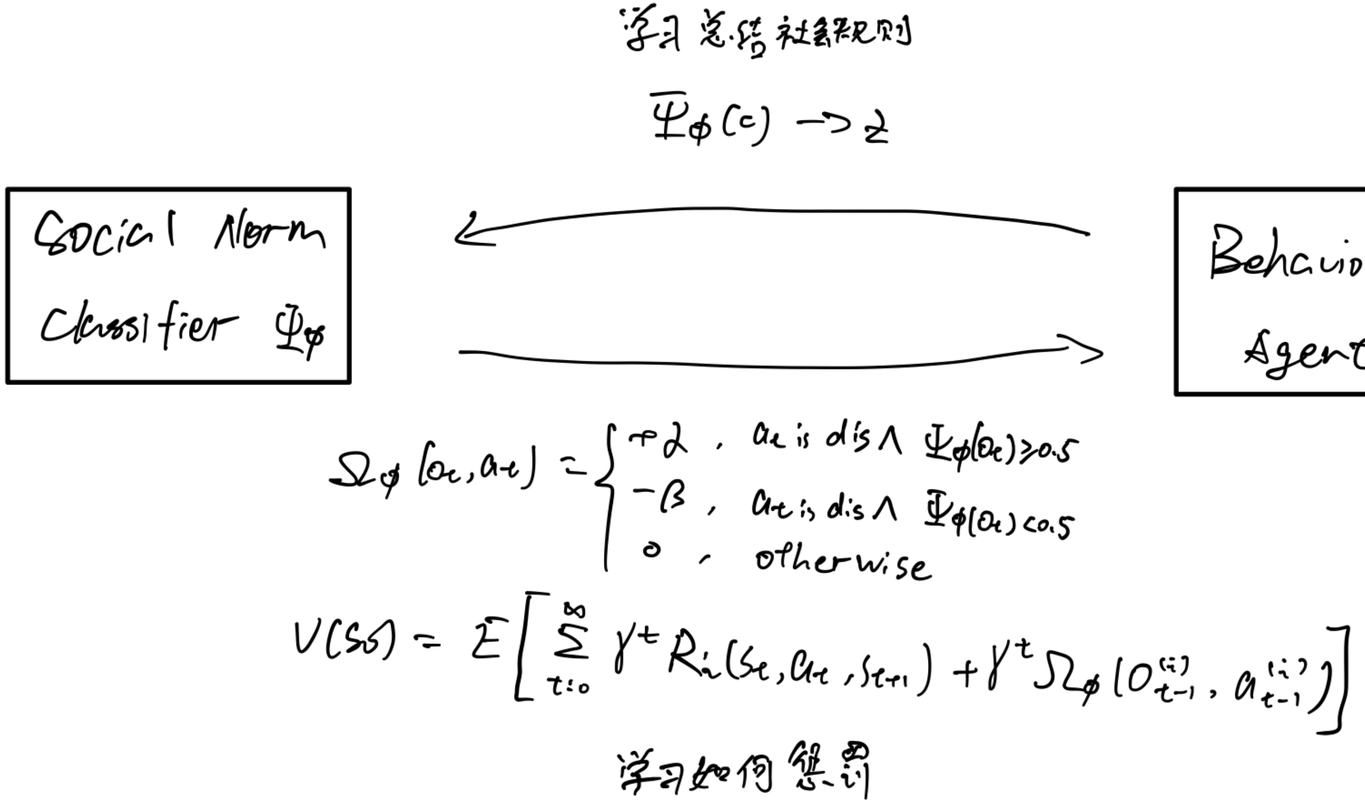



- and also increases the net agent return
- indicating that the costs of norm enforcement (punishing violators) are overcome by increased berry consumption
- 制裁会带来减产，但是这被制裁带来的收益克服了
- 奖励不包括intrinsic motivation

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } \alpha, \beta \in \mathbb{R}_0^+$$

# Experiments

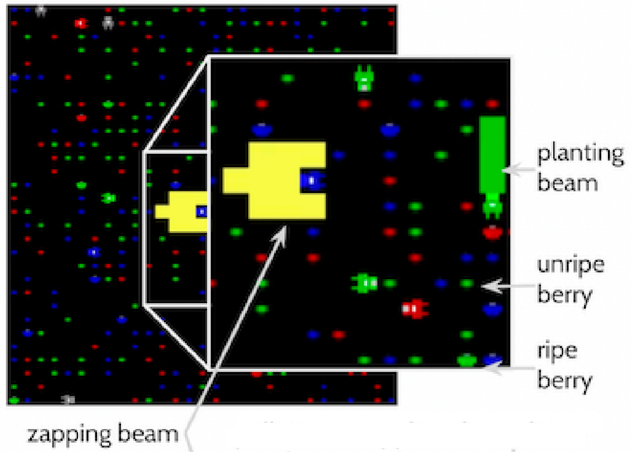
## Beneficial Effects of the Emergent Social Norms

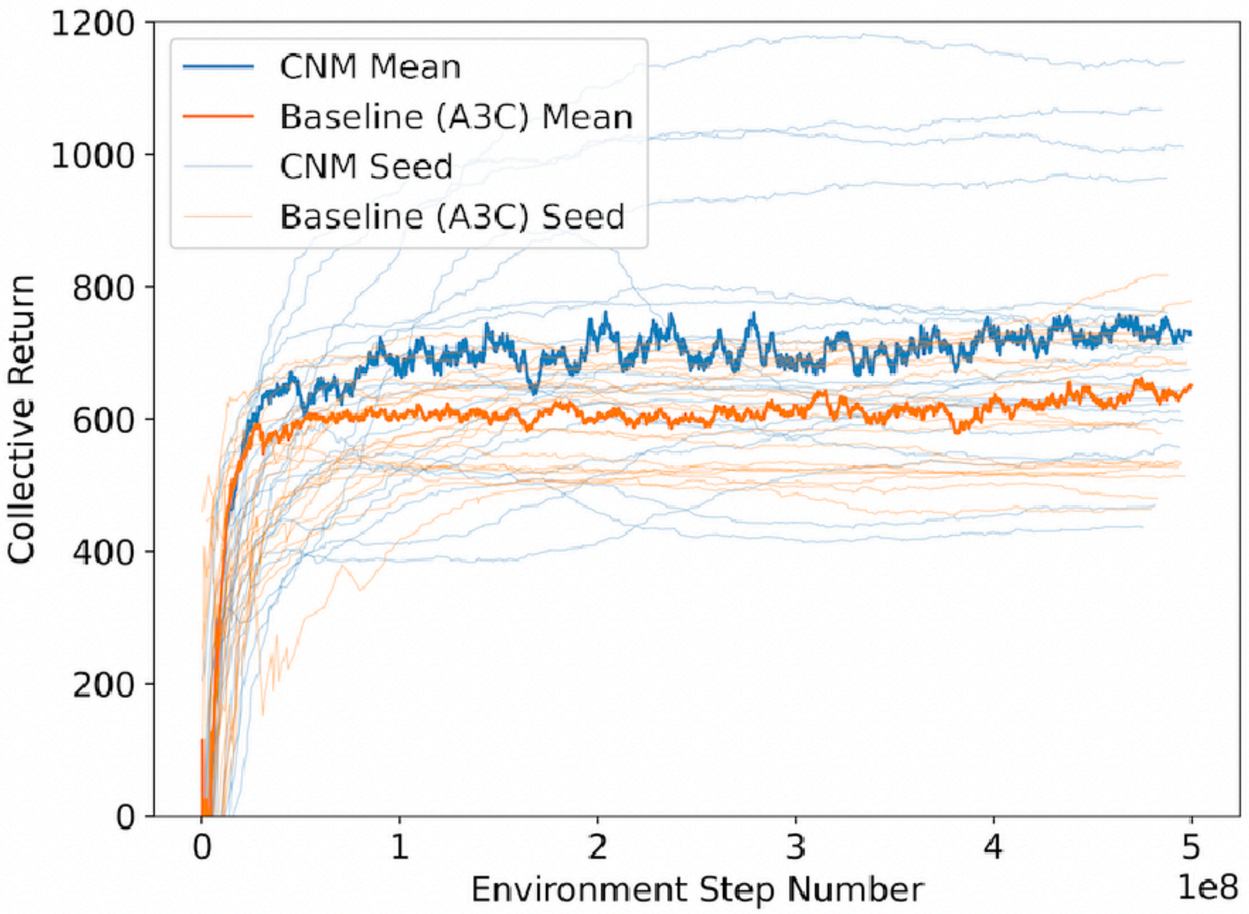



- Agents会随大流
- 会放大一开始很弱的制裁模式
- 制裁行为一开始都是随机的探索行为
- 不能保证他们会收敛到最好的那个equilibrium

# Experiments

## Beneficial Effects of the Emergent Social Norms

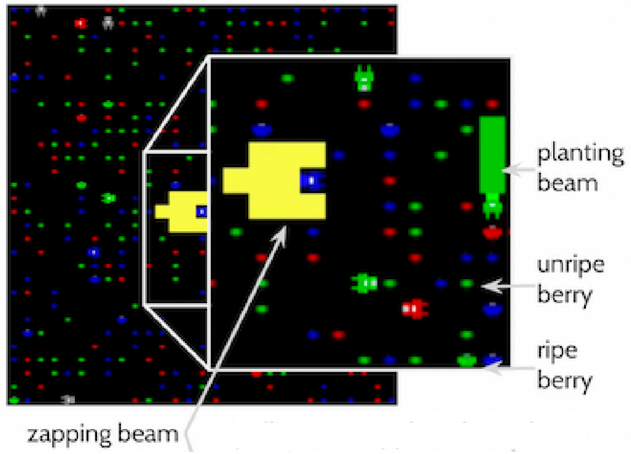


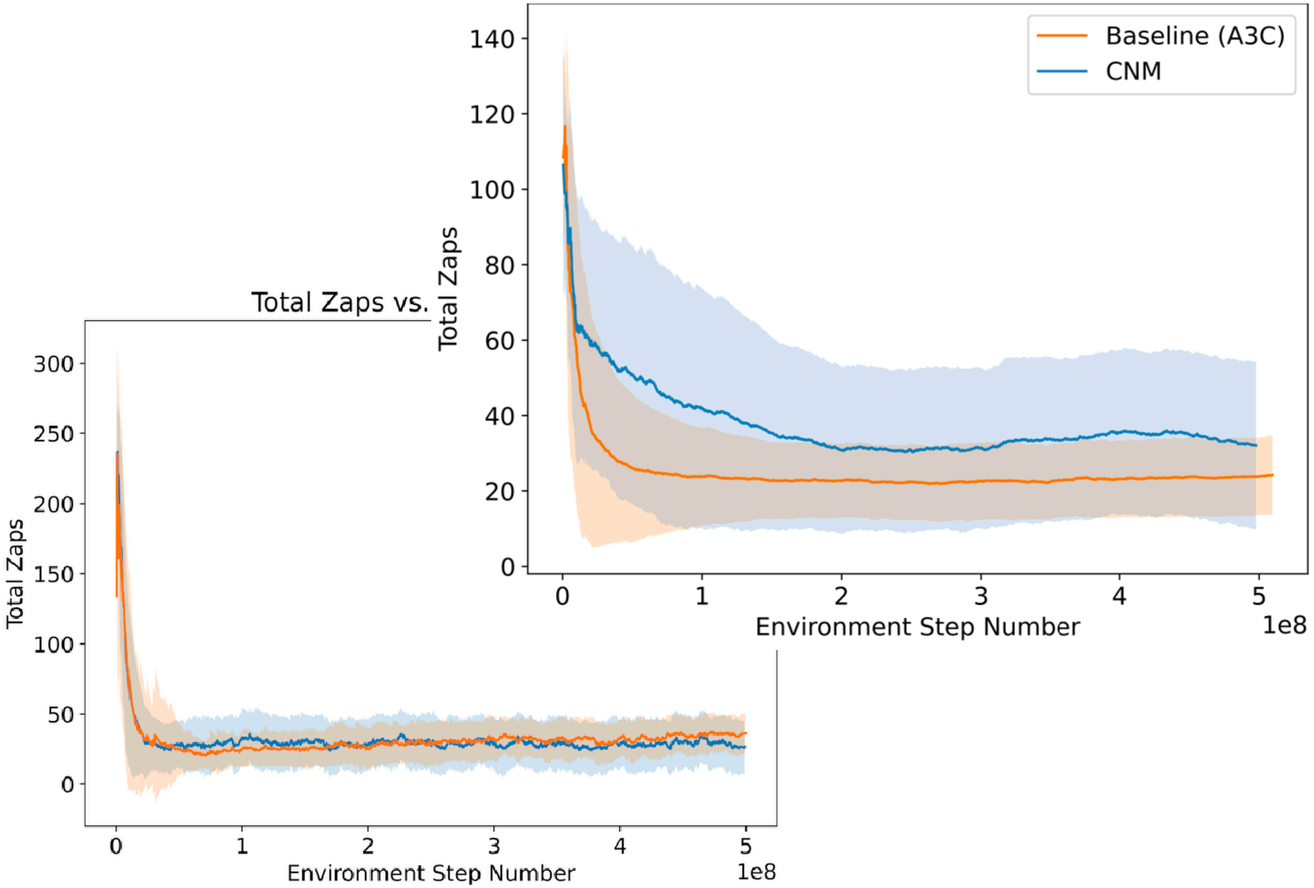




- 如果Agents一开始有趋势去种没人喜欢的蓝色浆果，并惩罚free-riders
- 那么classifier会学习这种行为，并认为这是合理的
- 那么就会收敛到blue equilibrium
- 所以运行了不同的seeds结果会有很大差别，variation很大

# Experiments

## Beneficial Effects of the Emergent Social Norms



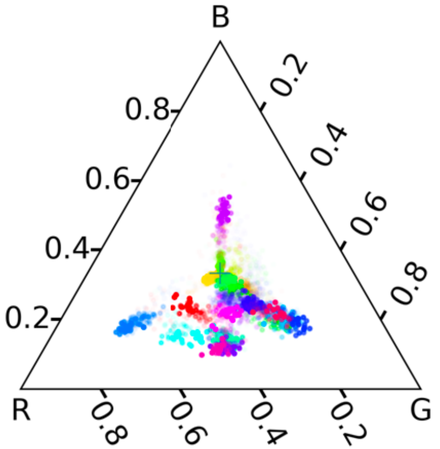




- 他们确认：奖励的改善并没有以某种方式发生，由于惩罚行为的压制，以及其造成后续奖励的减少（个体间的惩罚越少则对整体越好）（存疑）
- 用了CNM模型，惩罚的次数还变多了
- 之前提过，Agent可能会去制裁那些和自己有竞争的人来最大化自己利益
- 因此，集体回报的改善来自使用惩罚方式的变化（不能允许制裁那些竞争的人）

# Experiments

## How does CNM establish social norms?

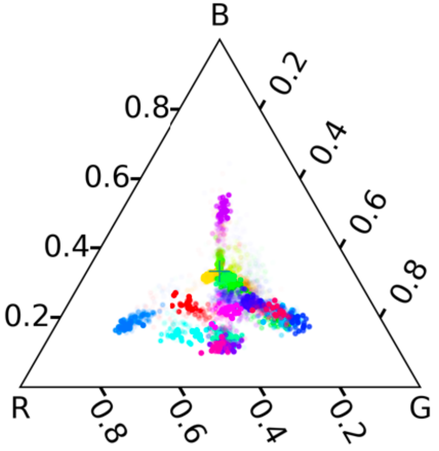
- 表面CNM刺激了Agent去遵守Social Norm
- 如果他们没有按现在的equilibrium去做，比如free-ride或者朝着别的equilibrium发展，他们就会被制裁
- equilibria → the corners of the **berry fraction simplex**（拓扑学，每个x大于等于0，相加等于1，三角形里面的所有点，三种浆果的数量比例）
- 如果是朝着角落去，说明收敛了
- monoculture fraction $m = \max\{\frac{r}{r+g+b},, \frac{g}{r+g+b}, \frac{b}{r+g+b}\}$
- Agent能通过只选择种植一种颜色的浆果，从而拿到更高的奖励



# Experiments

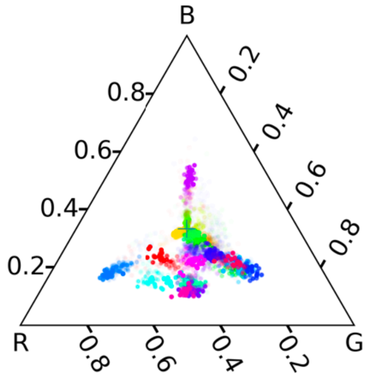
## How does CNM establish social norms?

- 中心点说明：全员偷懒，或全在撤销对方的操作（我发射激光把未成熟的浆果从红色变成蓝色，你又发射激光把未成熟的浆果从蓝色变成红色）
- 这个simplex说明了一个演化的过程，一个点是一个run的采样；同一种颜色是一个run（我的理解），从浅变到深，代表着这个run回合数越来越往后
- 可以看见不同的run向着不同的corner过去，说明收敛到不同的equilibrium

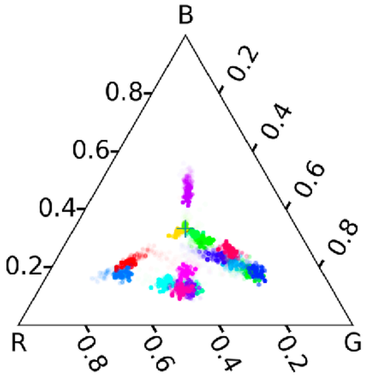


# Experiments

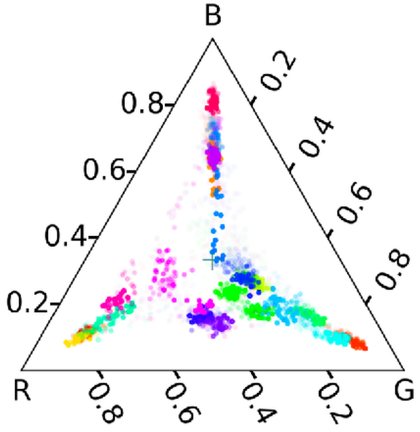
## How does CNM establish social norms?



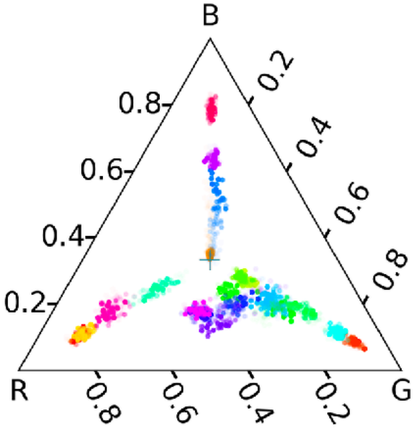

(a)




(b)




(c)




(d)

- (a) 无CNM，前2e8步
- (b) 无CNM，后2.5e8步
- (c) 有CNM，前2e8步
- (d) 有CNM，后2.5e8步
- 有CNM的明显变化会快，收敛更快
- Blue那个角的收敛情况相对少，可能是因为没人喜欢蓝色

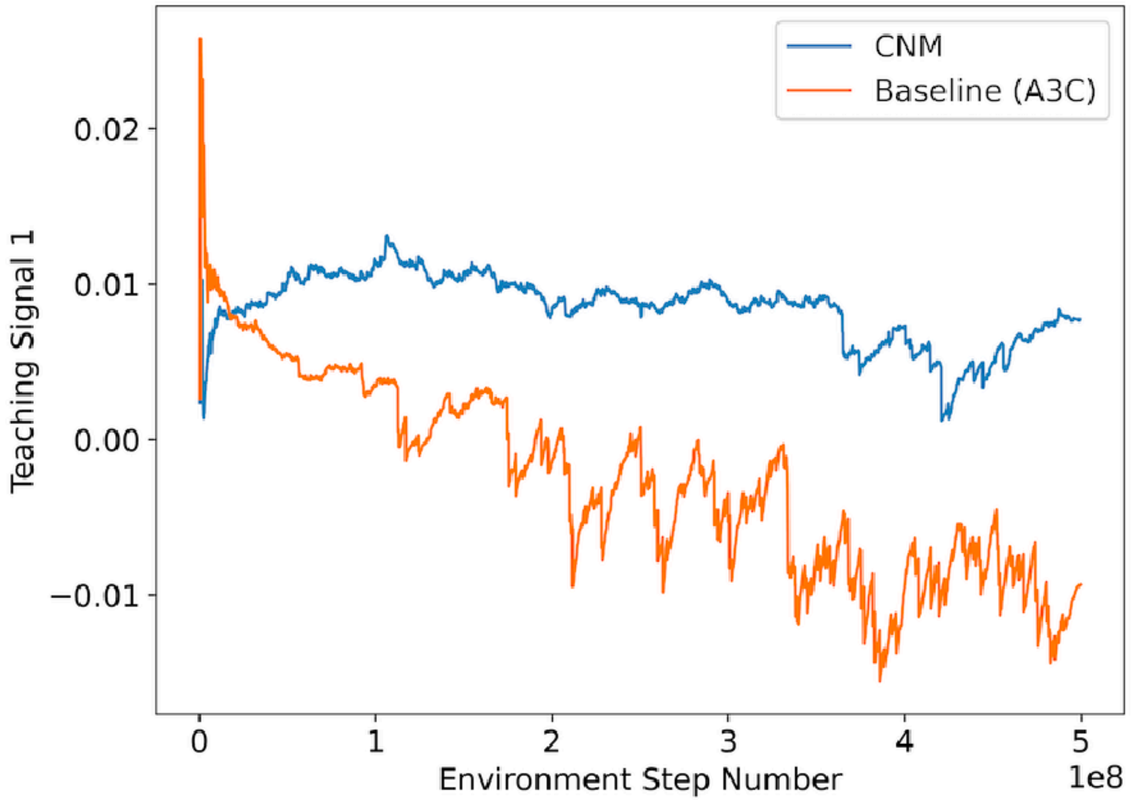
# Experiments

## How does CNM establish social norms?

- 偏离当前equilibrium的行为要被惩罚
- 用Bayes's rule计算每个颜色的p(zapped|color)，每种颜色被惩罚的可能性
- 以此来看看惩罚是怎么支持特定的equilibrium的
- 比较的是，比的是建立或保持equilibrium时，被惩罚的log likelihood的差异
- 如果一个颜色的likelihood的差异很大，那么学习算法会识别到，支持这个颜色的equilibrium的行为，很可能会导致别人的不赞成（被制裁）
- 这种差异就作为了一种teaching signal，使得Agent去种一种颜色而不是别的颜色

# Experiments

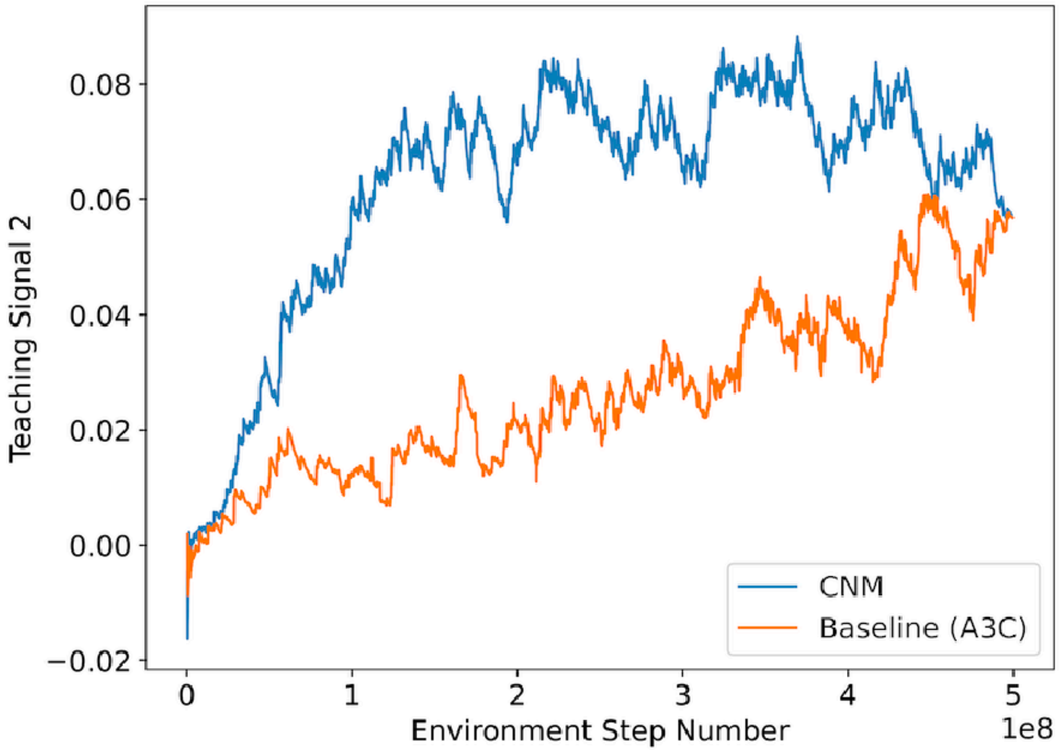
## How does CNM establish social norms?




- Teaching Signal 1：作为free-rider被惩罚likelihood - 支持当前dominant color而被惩罚的likelihood
- 如果这个信号是大的且是正的，对于学习算法就很容易识别出：如果支持当前的dominant color，受到惩罚的可能性会比当free-rider要好

# Experiments

## How does CNM establish social norms?




- Teaching Signal 2：the relative likelihood of getting punished，比的是，支持当前最多颜色的equilibrium，和切换到支持第二多颜色的equilibrium
- 如果这个信号是大的，对于学习算法就很容易识别出：如果支持当前的dominant color，受到惩罚的可能性会比切换去支持第二个多颜色要小

# Experiments

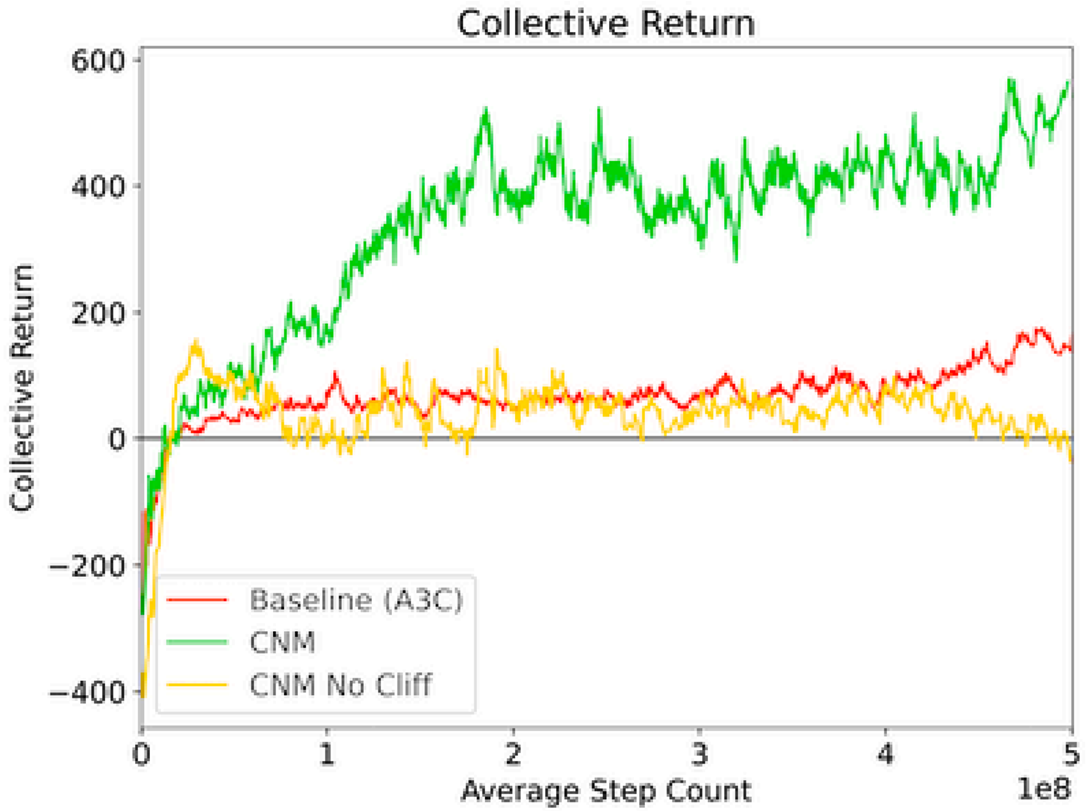
## Ablations on Architecture Components (CSP)

- (1) Is freezing the classifier necessary?
  - we allow the classifier to continue learning throughout training.
- (2) Is it essential to learn social norms from global sanctions or will local sanctions observed by each individual themselves suffice?
  - we train the classifier using only the sanctioning events directly observed by each agent.
- (3) Is our result sensitive to the relative scale between approval and disapproval pseudorewards?

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

# Experiments

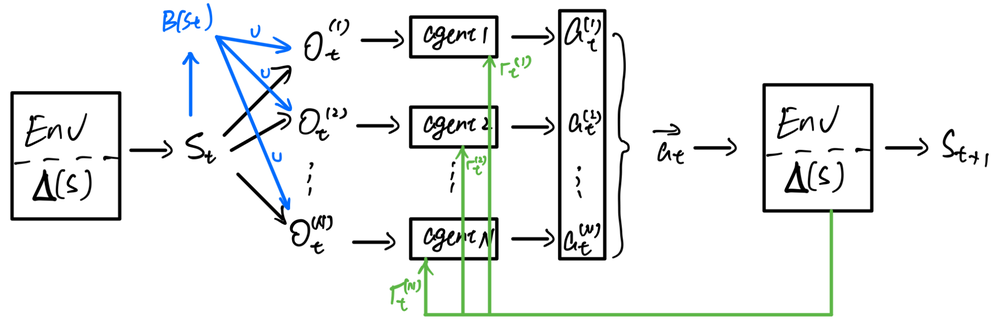
## Ablations on Architecture Components

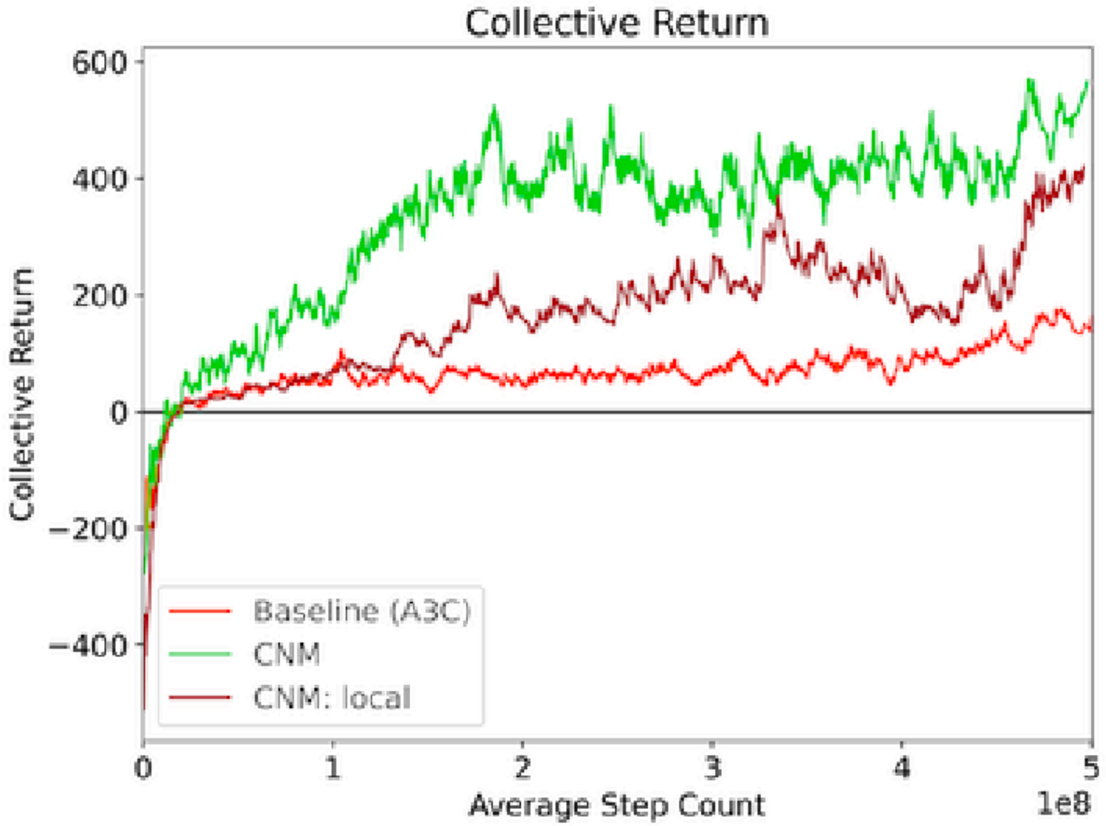



- 黄色是不冻结classifier的结果
- 看起来似乎合理的解释：
- 一开始的1e8步里搞定了free-riding
- 如果不被及时地限制classifier，后面合作的行为还是会被惩罚
- 因为有Agent有探索噪声，后面classifier学得可能不对，可能会去制裁那些目前策略已经是合作性质的Agent
- catastrophic forgetting：一开始一种颜色可能会被压制，很难去记住如何制裁一个很少出现的颜色。但是后面这些颜色可能又会出现（因为有噪声），这样一个被压制的策略可能又重新出现了

# Experiments

## Ablations on Architecture Components

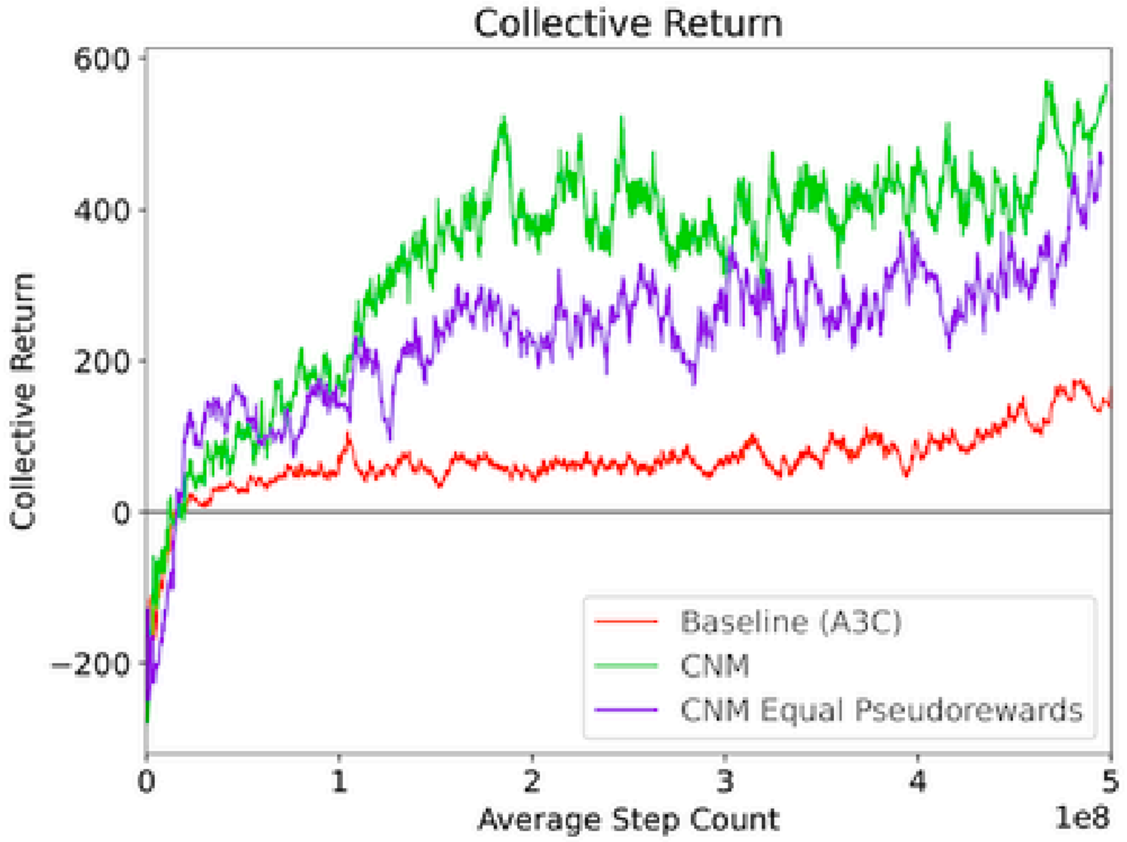



- 每个人只观察自己能观察到的sanction行为，不再是全局的
- 怎么界定每个人能观察到的sanction，好像没提
- 每个Agent现在要通过他们和别人的恰巧的互动，来自己推断Norm
- 这样classifier的样本数一下就少了很多，更难学了

# Experiments

## Ablations on Architecture Components




- 原来CNM是绿色的，beta是alpha的2倍
- 紫色的CNM，beta=alpha

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

 Thank You!