# Mega-Reward: Achieving Human-Level Play without Extrinsic Rewards

**Yuhang Song,[1] Jianyi Wang,[3] Thomas Lukasiewicz,[1] Zhenghua Xu,[1,2*] Shangtong Zhang,[1] Andrzej Wojcicki,[4] Mai Xu[3]**

[1]Department of Computer Science, University of Oxford, United Kingdom
[2]State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China
[3]School of Electronic and Information Engineering, Beihang University, China
[4]Lighthouse
{yuhang.song,thomas.lukasiewicz,shangtong.zhang}@cs.ox.ac.uk, {iceclearwjy,maixu}@buaa.edu.cn
zhenghua.xu@hebut.edu.cn, andrzej@wojcicki.xyz

## Abstract

Intrinsic rewards were introduced to simulate how human intelligence works; they are usually evaluated by intrinsically-motivated play, i.e., playing games without extrinsic rewards but evaluated with extrinsic rewards. However, none of the existing intrinsic reward approaches can achieve human-level performance under this very challenging setting of intrinsically-motivated play. In this work, we propose a novel megalomania-driven intrinsic reward (called *mega-reward*), which, to our knowledge, is the first approach that achieves human-level performance in intrinsically-motivated play. Intuitively, mega-reward comes from the observation that infants' intelligence develops when they try to gain more control on entities in an environment; therefore, mega-reward aims to maximize the control capabilities of agents on given entities in a given environment. To formalize mega-reward, a relational transition model is proposed to bridge the gaps between direct and latent control. Experimental studies show that mega-reward (i) can greatly outperform all state-of-the-art intrinsic reward approaches, (ii) generally achieves the same level of performance as Ex-PPO and professional human-level scores, and (iii) has also a superior performance when it is incorporated with extrinsic rewards.

## Introduction

Since humans can handle real-world problems without explicit extrinsic reward signals (Friston 2010), intrinsic rewards (Oudeyer and Kaplan 2009) are introduced to simulate how human intelligence works. Notable recent advances on intrinsic rewards include empowerment-driven (Klyubin, Polani, and Nehaniv 2005; 2008; Mohamed and Rezende 2015; Montúfar, Ghazi-Zahedi, and Ay 2016), count-based novelty-driven (Bellemare et al. 2016; Martin et al. 2017; Ostrovski et al. 2017; Tang et al. 2017), prediction-error-based novelty-driven (Achiam and Sastry 2017; Pathak et al. 2017; Burda et al. 2018; 2019), stochasticity-driven (Florensa, Duan, and Abbeel 2017), and diversity-driven (Song et al. 2019a) approaches. Intrinsic reward approaches are usually evaluated by *intrinsically-motivated play*, where proposed approaches are used to play games without extrinsic rewards but evaluated with extrinsic rewards. However, though proved to be able to learn some useful knowledge (Florensa, Duan, and Abbeel 2017; Song et al. 2019a) or to conduct a better exploration (Burda et al. 2018; 2019), none of the state-of-the-art intrinsic reward approaches achieves a performance that is comparable to human professional players under this very challenging setting of intrinsically-motivated play.

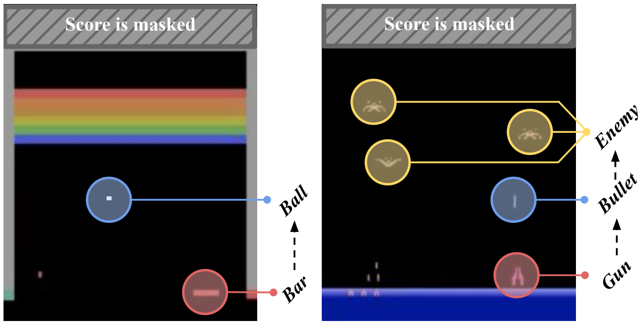

Figure 1: Latent control in *Breakout* (left) and *DemonAttack* (right).

In this work, we propose a novel megalomania-driven intrinsic reward (called *mega-reward*), which, to our knowledge, is the first approach that achieves human-level performance in intrinsically-motivated play. The idea of mega-reward originates from early psychology studies on *contingency awareness* (Watson 1966; Baeyens, Eelen, and van den Bergh 1990; Bellemare, Veness, and Bowling 2012), where infants are found to have awareness of how entities in their observation are potentially under their control. We notice that the way in which contingency awareness helps infants to develop their intelligence is to motivate them to have more control over the entities in the environment; therefore, we believe that having more control over the entities in the environment should be a very good intrinsic reward. Mega-reward follows this intuition, aiming to maximize the control capabilities of agents on given entities in a given environment.

Specifically, taking the game *Breakout* (shown in Fig. 1

[*] Corresponding author: Zhenghua Xu.

(left)) as an example, if an infant is learning to play this game, contingency awareness may first motivate the infant to realize that he/she can control the movement of an entity, *bar*; then, with the help of contingency awareness, he/she may continue to realize that blocking another entity, *ball*, with the bar can result in the ball also being under his/her control. Thus, the infant's skills on playing this game is gradually developed by having more control on entities in this game.

Furthermore, we also note that entities can be controlled by two different modes: *direct control* and *latent control*. Direct control means that an entity can be controlled directly (e.g., *bar* in *Breakout*), while latent control means that an entity can only be controlled indirectly by controlling another entity (e.g., *ball* is controlled indirectly by controlling *bar*). In addition, latent control usually forms a hierarchy in most of the games; the game *DemonAttack* as shown in Fig. 1 (right) is an example: there is a *gun*, which can be fired (direct control); then firing the gun controls *bullets* (1st-level latent control); finally, the *bullets* control *enemies* if they eliminate enemies (2nd-level latent control).

Obviously, gradually discovering and utilizing the hierarchy of latent control helps infants to develop their skills on such games. Consequently, mega-reward should be formalized by maximizing not only direct control, but also latent control on entities. This thus requests the formalization of both direct and latent control. However, although we can model direct control with an attentive dynamic model (Choi et al. 2019), there is no existing solution that can be used to formalize latent control. Therefore, we further propose a *relational transition model* (RTM) to bridge the gap between direct and latent control by learning how the transition of each entity is related to itself and other entities. For example, the agent's direct control on entity $A$ can be passed to entity $B$ as latent control if $A$ implies the transition of $B$. With the help of RTM, we are able to formalize mega-reward, which is computationally tractable.

Extensive experimental studies have been conducted on 18 Atari games and the "*noisy TV*" domain (Burda et al. 2018); the experimental results show that (i) mega-reward significantly outperforms all six state-of-the-art intrinsic reward approaches, (ii) even under the very challenging setting of intrinsically-motivated play, mega-reward (without extrinsic rewards) still achieves generally the same level of performance as two benchmarks (with extrinsic rewards), Ex-PPO and professional human-level scores, and (iii) the performance of mega-reward is also superior when it is incorporated with extrinsic rewards, outperforming state-of-the-art approaches in two different settings.

This paper's contributions are briefly as follows: (1) We propose a novel intrinsic reward, called mega-reward, which aims to maximize the control capabilities of agents on given entities in a given environment. (2) To realize mega-reward, we further propose a relational transition model (RTM) to bridge the gap between direct and latent control. (3) Experiments on 18 Atari games and the "*noisy TV*" domain show that mega-reward (i) greatly outperforms all state-of-the-art intrinsic reward approaches, (ii) generally achieves the same level of performance as two benchmarks, Ex-PPO and professional human-level scores, and (iii) also has a superior performance when it is incorporated with extrinsic rewards. Easy-to-run code is released in https://github.com/YuhangSong/Mega-Reward.

## Direct Control

We start with the notion of *direct control*. Generally, we consider the effect of the action $a_{t-1} \in \mathcal{A}$ on the state $s_t \in \mathcal{S}$ as direct control. In practice, we are more interested in how different parts of a visual state are being directly controlled by $a_{t-1}$. Thus, prevailing frameworks (Jaderberg et al. 2017; Choi et al. 2019) mesh $s_t$ into subimages, as shown in Fig. 2, where we denote a subimage of $s_t$ at the coordinates $(h, w)$ as $s_t^{h,w} \in \mathcal{S}^{H,W}$. The number of possible coordinates $(h, w)$ and space of the subimage $\mathcal{S}^{H,W}$ is determined by the granularity of the meshing $H, W$. Then, we can define the quantification of how likely each $s_t^{h,w}$ is being directly controlled by $a_{t-1}$ as $\alpha(s_t^{h,w}, a_{t-1}) \in \mathbb{R}$.

The state-of-the-art method (Choi et al. 2019) models $\alpha(s_t^{h,w}, a_{t-1})$ with an *attentive dynamic model (ADM)*, which predicts $a_{t-1}$ from two consecutive states $s_{t-1}$ and $s_t$. The key intuition is that ADM should attend to the most relevant part of the states $s_{t-1}$ and $s_t$, which is controllable by $a_{t-1}$, to be able to classify $a_{t-1}$. Thus, a *spatial attention mechanism* (Bahdanau, Cho, and Bengio 2015; Xu et al. 2015) can be applied to ADM to model $\alpha(s_t^{h,w}, a_{t-1})$:

$$e_t^{h,w} = \Theta\left(\left[s_t^{h,w} - s_{t-1}^{h,w}; s_t^{h,w}\right]\right) \in \mathbb{R}^{|\mathcal{A}|} \quad (1)$$

$$\bar{\alpha}(s_t^{h,w}, a_{t-1}) = \Lambda\left(s_t^{h,w}\right) \in \mathbb{R} \quad (2)$$

$$\alpha(s_t^{h,w}, a_{t-1}) = \text{sparsemax}\left(\bar{\alpha}(s_t^{h,w}, a_{t-1})\right) \in \mathbb{R} \quad (3)$$

$$p(\hat{a}_{t-1}|s_{t-1}, s_t) = \text{SoM}\left(\sum_{h' \in H, w' \in W} \alpha(s_t^{h,w}, a_{t-1}) \cdot e_t^{h,w}\right), \quad (4)$$

where $\Theta$ and $\Lambda$ are two parameterized models, $e_t^{h,w}$ is the logits of the probability of the predicted action $p(\hat{a}_{t-1}|s_{t-1}, s_t)$ before masking it by the spatial attention $\alpha(s_t^{h,w}, a_{t-1})$, SoM is the softmax operation, and $\bar{\alpha}(s_t^{h,w}, a_{t-1})$ is the spatial attention mask before converting it into a probability distribution $\alpha(s_t^{h,w}, a_{t-1})$ using the sparsemax operator (Martins and Astudillo 2016). The models can be optimized with the standard cross-entropy loss $\mathcal{L}_{\text{action}}(a_{t-1}, \hat{a}_{t-1})$ relative to the ground-truth action $a_{t-1}$ that the agent actually has taken. More details, including additional *attention entropy regularization losses*, can be found in (Choi et al. 2019).

## From Direct Control to Latent Control

Built on the modelled direct control $\alpha(s_t^{h,w}, a_{t-1})$, we propose to study the notion of *latent control*, which means the effect of $a_{t-n}$ on the state $s_t$, where $n > 1$. Like in direct control, we are interested in how different parts of a visual state $s_t^{h,w}$ are being latently controlled, thus, we define the quantification of how likely $s_t^{h,w}$ is being latently controlled

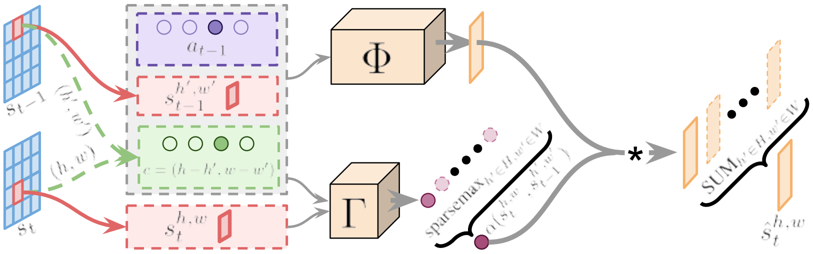

Figure 2: Relational transition model.

by $a_{t-n}$ as $\alpha(s_t^{h,w}, a_{t-n})$. As an example for latent control quantification $\alpha(s_t^{h,w}, a_{t-n})$, consider the game *DemonAttack* in Fig. 1 (right). In this case, consider $a_{t-n}$ to be the action at step $t - n$ that shoots a bullet from the *gun*, $s_t^{h,w}$ to be the subimage containing one of the *enemy* at step $t$. Clearly, $s_t^{h,w}$ is influenced by $a_{t-n}$. But this influence needs $n$ steps to take effect, where $n$ is unknown; also, it involves an intermediate entity *bullet*, i.e., the *gun* "controls" the *bullet*, the *bullet* "controls" the *enemy*. Due to the nature of delayed influence and involvement of intermediate entities, we call it latent control, in contrast to direct control.

To compute $\alpha(s_t^{h,w}, a_{t-n})$, one possible way is to extend ADM to longer interval, i.e., a variant of ADM takes in $s_{t-n}$, $s_t$ and makes predictions of a sequence of actions $a_{t-n}, a_{t-n+1}, ..., a_{t-1}$. However, $n$ is not known in advance, and we may need to enumerate over $n$. We propose an alternative solution, based on the observation that latent control always involves some intermediate entities. For example, the latent control from the *gun* to the *enemy* in *DemonAttack* in Fig. 1 (right) involves an intermediate entity *bullet*. Thus, how likely the *enemy* is latently controlled can be quantified if we can model that (1) an action directly controls the *gun*, (2) the *gun* directly controls the *bullet*, and (3) the *bullet* directly controls the *enemy*. In this solution, latent control is broken down into several direct controls, which avoids dealing with the unknown $n$. As can be seen, this solution requires modelling not only the direct control of an action on a state, but also the direct control of a state on the following state, which is a new problem. Formally speaking, we first quantify how likely $s_t^{h,w}$ is controlled by $s_{t-1}^{h',w'}$ with $\alpha(s_t^{h,w}, s_{t-1}^{h',w'}) \in \mathbb{R}$. Then, we can formally express the above idea by

$$\alpha(s_t^{h,w}, a_{t-n}) = \sum_{h' \in H, w' \in W} \alpha(s_t^{h,w}, s_{t-1}^{h',w'}) \alpha(s_{t-1}^{h',w'}, a_{t-n}). \quad (5)$$

Thus, Eq. (5) derives $\alpha(s_t^{h,w}, a_{t-n})$ from $\alpha(s_{t-1}^{h',w'}, a_{t-n})$. If we keep applying Eq. (5) on $\alpha(s_{t-1}^{h',w'}, a_{t-n})$, we can eventually derive it from $\alpha(s_{t-n+1}^{h',w'}, a_{t-n})$, which is the qualification of direct control defined in the last section and can be computed via Eqs. (1) to (4). Thus, $\alpha(s_t^{h,w}, a_{t-n})$ can be computed as long as we know $\alpha(s_{t'}^{h,w}, s_{t'-1}^{h',w'})$ for all $t' \in [t - n + 2, t]$. That is, $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$ bridges the gap between direct and latent control. Furthermore, since $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$ models how a part of the previous state $s_{t-1}^{h',w'}$ implies a part of the current state $s_t^{h,w}$, it reveals the need of a new form of transition model, which contains information about the relationships between different parts of the state underlying the transition of full states. Thus, we call it a *relational transition model (RTM)*. In the next section, we introduce our method to learn RTM efficiently.

## Relational Transition Model

To produce an approximation of $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$, we propose relational transition models (RTMs), the general idea behind which is introducing a spatial attention mechanism to the *transition model*. Specifically, Fig. 2 shows the structure of an RTM, which consists of two parameterized models, namely, $\Phi$ for relational transition modeling and $\Gamma$ for attention mask estimation. We first define the forward function of $\Phi$; it makes a prediction of the transition from $s_{t-1}^{h',w'}$ to $s_t^{h,w}$:

$$\hat{s}_t^{h,w} = \sum_{h' \in H, w' \in W} \alpha(s_t^{h,w}, s_{t-1}^{h',w'}) \Phi\left(\left[s_{t-1}^{h',w'}, a_{t-1}, c\right]\right). \quad (6)$$

Here, $\hat{s}_t^{h,w}$ represents the prediction of $s_t^{h,w}$. Also, note that apart from taking in $s_{t-1}^{h',w'}$, $\Phi$ also takes in the relative coordinates $c = (h - h', w - w')$ and $a_{t-1}$, both as one-hot vectors, so that the model $\Phi$ knows the relative position of the part to predict and the action taken. Furthermore, $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$ is the estimated attention mask of predicting $s_t^{h,w}$ from $s_{t-1}^{h',w'}$, which models how informative each $s_{t-1}^{h',w'}$ of different $h', w'$ is for the prediction of $s_t^{h,w}$, i.e., how likely $s_{t-1}^{h',w'}$ controls $s_t^{h,w}$. $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$ is estimated by the model $\Gamma$. Specifically, $\Gamma$ first estimates $\bar{\alpha}(s_t^{h,w}, s_{t-1}^{h',w'})$ via

$$\bar{\alpha}(s_t^{h,w}, s_{t-1}^{h',w'}) = \Gamma\left(\left[s_t^{h,w}, s_{t-1}^{h',w'}, a_{t-1}, c\right]\right), \quad (7)$$

which is later sparsemaxed over $h' \in H, w' \in W$ to compute

$$\alpha(s_t^{h,w}, s_{t-1}^{h',w'}) = \text{sparsemax}\left(\bar{\alpha}(s_t^{h,w}, s_{t-1}^{h',w'})\right). \quad (8)$$

We train RTM end-to-end with $L_{\text{transition}} = \text{MSE}(\hat{s}_t^{h,w}, s_t^{h,w})$. As an intuitive explaination of RTM, taking the game *Breakout* (shown in Fig. 1 (left)) as an example, $\Phi$ makes three predictions of the current ball based on the previous ball, bar, and brick. Since the final prediction of the current ball is the weighted combination of these three predictions, $\Gamma$ is further used to estimate the weights of this combination, measuring different control effects that the previous ball, bar, and brick have on the current ball. We thus propose $\Phi$ and $\Gamma$ as relational transition models.

RTM has introduced separated forwards over every $h' \in H$, $w' \in W$, $h \in H$, and $w \in W$; however, by putting the separated forwards into the batch axis, the computing is well parallelized. We reported the running times and included code in the extended paper (Song et al. 2019b).

## Formalizing Intrinsic Rewards

Summarizing the previous sections, ADM and RTM model $\alpha(s_{t-n+1}^{h',w'}, a_{t-n})$ and $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$, respectively. Based on this, $\alpha(s_t^{h,w}, a_{t-n})$ can be modelled via Eq. (5). In this section, we formalize the intrinsic reward from $\alpha(s_t^{h,w}, a_{t-n})$.

First, $\{\alpha(s_t^{h,w}, a_{t-n})\}^{n \in [1,t]}$ contains all the information about what is being controlled by the agent in the current state, considering all the historical actions with both direct and latent control. Clearly, computing all components in the above set is intractable as $t$ increases. Thus, we define a quantification of accumulated latent control $g_t^{h,w} \in \mathbb{R}$, which is a discounted sum of $\alpha(s_t^{h,w}, a_{t-n})$ over $n$:

$$g_t^{h,w} = \sum_{n \in [1,t]} \rho^{n-1} \alpha(s_t^{h,w}, a_{t-n}), \tag{9}$$

where $\rho$ is a discount factor, making $\alpha(s_t^{h,w}, a_{t-n})$ with $n >> 1$ have a lower contribution to the estimation of $g_t^{h,w}$. Then, we show that $g_t^{h,w}$ can be computed from $g_{t-1}^{h,w}$ and $\alpha(s_t^{h,w}, a_{t-1})$ without enumerating over $n$ (see proof of Lemma 1 in the extended paper (Song et al. 2019b)):

$$g_t^{h,w} = \rho \sum_{h' \in H, w' \in W} \alpha(s_t^{h,w}, s_{t-1}^{h',w'}) g_{t-1}^{h',w'} + \alpha(s_t^{h,w}, a_{t-1}), \tag{10}$$

which reveals that we can maintain an $H \times W$ memory for $g^{h,w}$, and then update $g_{t-1}^{h,w}$ to $g_t^{h,w}$ at each step with $\alpha(s_t^{h,w}, s_{t-1}^{h',w'})$ and $\alpha(s_t^{h,w}, a_{t-1})$ according to (10). The intuitive integration of $g_t^{h,w}$ is an overall estimation of what is being controlled currently, both directly and latently, considering the effect of all historical actions. This also coincides with the intuition that humans do not explicitly know what they latently control for each historical action. Instead, we maintain an overall estimation of what is under the historical actions' control, both directly and latently. At last, to maximize $\sum_{h \in H, w \in W} g_{t=T}^{h,w}$, where $T$ is the terminal step, the intrinsic reward (our mega-reward) at each step $t$ should be:

$$r_t^{\text{meg}} = \textstyle\sum_{h \in H, w \in W} \left( g_t^{h,w} - g_{t-1}^{h,w} \right) . \tag{11}$$

## Experiments

In extensive experiments, we evaluated the performance of mega-reward. We first report on the evaluation on 18 Atari games under the very challenging settings of intrinsically-motivated play, where a case study is used to visualize how each part of mega-reward works, and mega-reward is compared with six state-of-the-art intrinsic rewards, the benchmark of a PPO agent with access to extrinsic rewards (*Ex-PPO*), and the benchmark of professional human-level scores, to show its superior performance. Then, we further investigate two possible ways to integrate mega-reward with extrinsic rewards. Finally, a few failure cases of mega-reward are studied, showing possible topics for future research.

Mega-reward is implemented on PPO in (Schulman et al. 2017) with the same set of hyper-parameters, along with $H \times W = 4 \times 4$ and $\rho = 0.99$. $H \times W = 4 \times 4$ is a trade-off between efficiency and accuracy. An ablation study on value settings of $H \times W$ over the game *Breakout* is available in the extended paper (Song et al. 2019b), showing that $4 \times 4$ is sufficient to achieve a reasonable accuracy, while having the best efficiency. The network structures of $\Phi$ and $\Gamma$ are provided in the extended paper (Song et al. 2019b). The hyper-parameters of the other baseline methods are set as in the corresponding original papers. The environment is wrapped as in (Burda et al. 2018; Mnih et al. 2015).

Due to the page limit, running times, additional ablation studies (e.g., of components in mega-reward), and additional comparisons under other settings (e.g., the setting when agents have access to both intrinsic and extrinsic rewards) are provided in the extended paper (Song et al. 2019b).

### Intrinsically-Motivated Play of Mega-Reward

Intrinsically-motivated play is an evaluation setting where the agents are trained by intrinsic rewards only, and the performance is evaluated using extrinsic rewards. To make sure that the agent cannot gain extra information about extrinsic rewards, the displayed score in each game is masked out. To ensure a fair comparison, all baselines are also provided with a feature map $g_t^{h,w}$ as an additional channel. Here, all agents are run for 80M steps, with the last 50 episodes averaged as the final scores and reported in Table 1. The evaluation is conducted over 18 Atari games.

**Case Study.** Fig. 3 visualizes how each component in our method works as expected. The 1st row is a frame sequence. The 2nd row is the corresponding direct control map $\alpha(s_t^{h,w}, a_{t-1})$, indicating how likely each grid being directly controlled by $a_{t-1}$. As expected, the learned map shows the grid containing the bar being directly controlled. The 3rd row is the accumulated latent control map $g_t^{h,w}$, indicating how likely each grid being controlled (both directly and latently) by historical actions. As expected, the learned map shows: (1) only the bar is under control before the bar hits the ball (frames 1–5); (2) both the bar and the ball are under control after the bar has hit the ball (frames 6–10); and (3) the bar, ball, and displayed score are all under control if the opponent missed the ball (frame 11). The 4th row is mega-reward $r_t^{\text{meg}}$, obtained by Eq. (11) from the map in the 3rd row. As expected, it is high when the agent controls a new

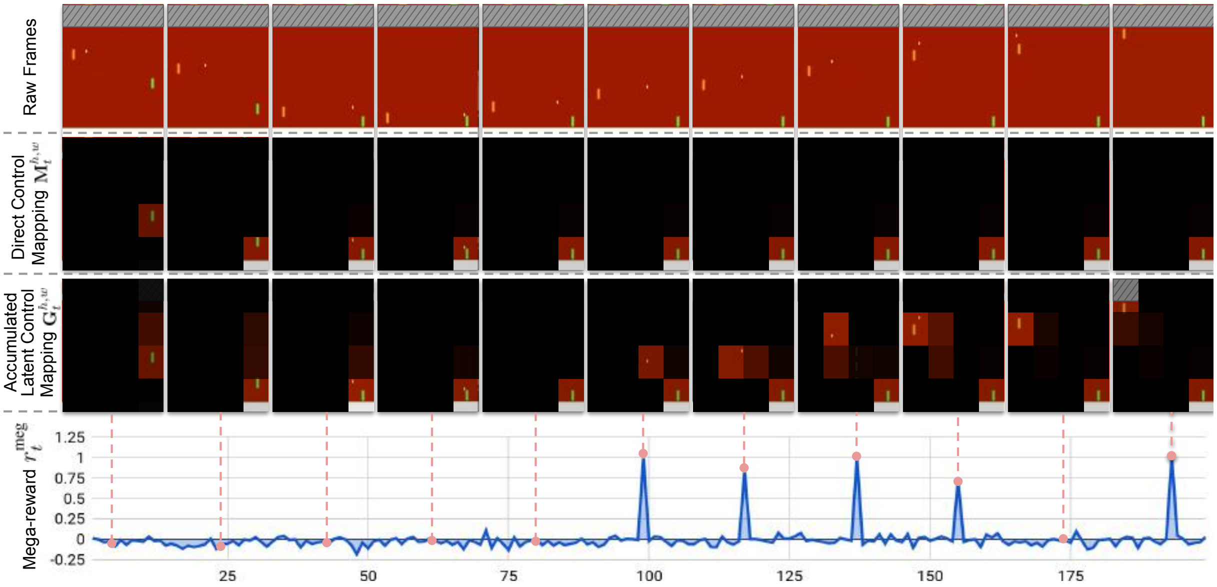

Figure 3: Case study: the example of *Pong*.

Table 1: Comparison of mega-reward against six baselines.

| Game | *Emp* | *Cur* | *RND* | *Sto* | *Div* | *Dir* | *Meg* |
|---|---|---|---|---|---|---|---|
| *Seaquest* | 612.2 | 422.2 | 324.2 | 103.5 | 129.2 | 323.1 | **645.2** |
| *Bowling* | 103.4 | **156.2** | 77.23 | 86.23 | 79.21 | 113.3 | 82.72 |
| *Venture* | 62.34 | 0.0 | 83.12 | 61.32 | 95.67 | 86.21 | **116.6** |
| *WizardOfWor* | 526.2 | 562.3 | 702.5 | 227.1 | 263.1 | 723.7 | **1030** |
| *Asterix* | 1536 | 1003 | 462.3 | 304.2 | 345.6 | 1823 | **2520** |
| *Robotank* | **5.369** | 3.518 | 3.619 | 4.164 | 2.639 | 1.422 | 2.310 |
| *BeamRider* | 944.1 | 864.2 | 516.3 | 352.1 | 381.2 | 1273 | **1363** |
| *BattleZone* | 3637 | 4625 | **8313** | 0.0 | 0.0 | 2262 | 3514 |
| *KungFuMaster* | 424.9 | **3042** | 652.1 | 245.1 | 523.9 | 423.7 | 352.4 |
| *Centipede* | 1572 | 3262 | **4275** | 1832 | 1357 | 2034 | 2001 |
| *Pong* | -7.234 | -8.234 | -17.42 | -16.52 | -14.53 | -17.62 | **-3.290** |
| *AirRaid* | 1484 | 1252 | 942.3 | 723.4 | 1426 | 1583 | **2112** |
| *DoubleDunk* | -18.26 | -20.42 | -17.34 | -19.34 | -18.35 | -17.72 | **-13.58** |
| *DemonAttack* | 9259 | 69.14 | 412.4 | 57.14 | 90.23 | 7838 | **10294** |
| *Berzerk* | 735.7 | 363.1 | 462.4 | 157.2 | 185.2 | 413.3 | **764.6** |
| *Breakout* | 201.4 | 145.3 | 125.5 | 113.5 | 1.352 | 125.2 | **225.3** |
| *Jamesbond* | 523.2 | 603.0 | 201.2 | 0.0 | 0.0 | 1383 | **3223** |
| *UpNDown* | 8358 | 8002 | 2352 | 331.3 | 463.3 | 60528 | **124423** |

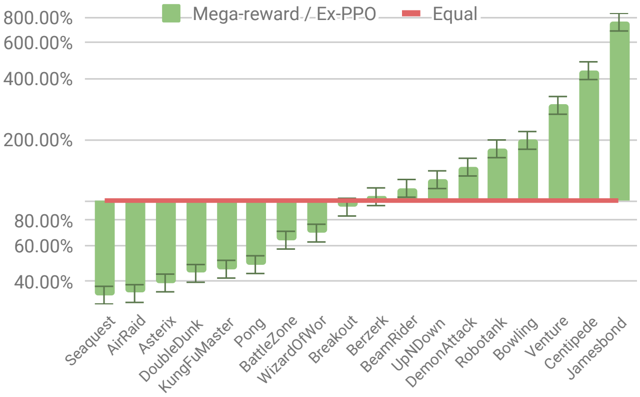

Figure 4: Mega-reward against the benchmark of Ex-PPO.

grid in the 3rd row (achieving more control over the grids in the state).

**Against Other Intrinsic Rewards.** To show the superior performance of mega-reward (denoted *Meg*), we first compare its performance with those of six state-of-the-art intrinsic rewards, i.e., *empowerment-driven* (denoted *Emp*) (Mohamed and Rezende 2015), *curiosity-driven* (denoted *Cur*) (Burda et al. 2018), *RND* (Burda et al. 2019), *stochasticity-driven* (denoted *Sto*) (Florensa, Duan, and Abbeel 2017), *diversity-driven* (denoted *Div*) (Song et al. 2019a), and a mega-reward variant with only direct control (denoted *Dir*). Results for more baselines can be found in the extended paper (Song et al. 2019b). By the experimental results in Table 1, mega-reward outperforms all six baselines substantially. In addition, we also have the following findings: (i) *Sto* and *Div* are designed for games with explicit hierarchical structures, so applying them on Atari games with no obvious temporal hierarchical structure will result in the worst performance among all baselines. (ii) *Dir* is also much worse than the other baselines, proving the necessity of latent control in the formalization of mega-reward. (iii) The failure of the empowerment-driven approach states that applying information theory objectives to complex video games like Atari ones is an open problem. A detailed discussion of the benefits of mega-reward over other intrinsically motivated approaches can be found in the extended paper (Song et al. 2019b). Videos demonstrating the benefits of mega-reward on all 57 Atari games can be found in the released code (Song 2019).

**Against Two Benchmarks.** In general, the purpose of evaluating intrinsic rewards in intrinsically-motivated play is to investigate if the proposed intrinsic reward approaches can achieve the same level of performance as two bench-

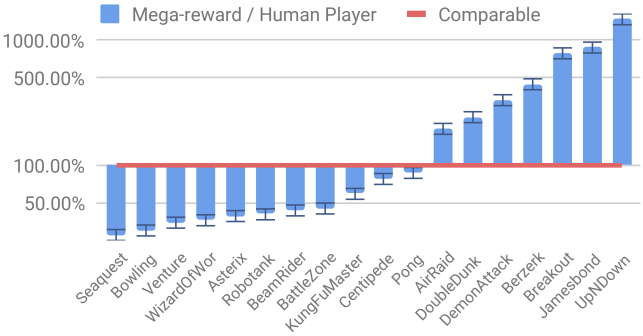

Figure 5: Mega-reward against the benchmark of human player.

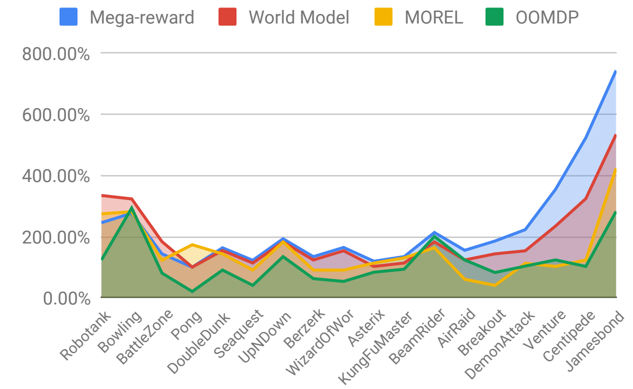

Figure 6: Relative improvements of the score when pretrained with mega-reward and world model (*Delta* $=$ *Mega-reward-World Model*).

marks: PPO agents with access to extrinsic rewards (denoted *Ex-PPO*) and professional human players. Therefore, we evaluate mega-reward using a relative score against two such benchmarks, which can be formally defined as

$$S_{\text{Relative}} = \frac{S_{\text{Mega-reward}} - S_{\text{Random}}}{S_{\text{Benchmark}} - S_{\text{Random}}} \times 100\%, \qquad (12)$$

where $S_{\text{Relative}} > 100\%$ means that mega-reward achieves a better performance than the corresponding benchmark, $S_{\text{Relative}} < 100\%$ that it achieves a worse performance, and $S_{\text{Relative}} = 0\%$ is random play.

Fig. 4 shows the comparative performance of mega-reward against Ex-PPO on 18 Atari games, where mega-reward greatly outperforms the Ex-PPO benchmark in 8 games, and is close to the benchmark in 2 games. These results show that mega-reward generally achieves the same level of or a comparable performance as Ex-PPO (though strong on some games and weak on others); thus, the proposed mega-reward is as informative as the human-engineered extrinsic rewards.

Similarly, Fig. 5 shows the comparative performance of mega-reward against professional human players. As the performance of professional human players (i.e., professional human-player scores) on 16 out of 18 Atari games have already been measured by (Mnih et al. 2015), we measure the professional human-player scores on *AirRaid* and *Berzerk* using the same protocol. Generally, in Fig. 5, mega-reward greatly outperforms the professional human-player benchmark in 7 games, and is close to the benchmark in 2 games. As the professional players are equipped with strong prior knowledge about the game and the scores displayed in the state, they show a relatively high-level of human skills on the corresponding games. Thus, the results sufficiently prove that mega-reward has generally reached the same level of (or a comparable) performance as a human player.

## Pretraining with Mega-Reward

In many real-world cases, the agent may have access to the dynamics of the environment before the extrinsic rewards are available (Ha and Schmidhuber 2018). This means that an agent can only play with the dynamics of the environment to pretrain itself before being assigned with a specific task (i.e., having access to extrinsic rewards). Therefore, we further investigate the first way to integrate mega-reward with extrinsic rewards (i.e., using mega-reward to pretrain the agent) and compare the pretrained agent with that in the state-of-the-art world model (Ha and Schmidhuber 2018), as well as two state-of-the-art methods of unsupervised representation learning for RL: MOREL (Goel, Weng, and Poupart 2018) and OOMDP (Diuk, Cohen, and Littman 2008).

The evaluation is based on a relative improvement of the score, which is formally defined as

$$S_{\text{Improve}} = \frac{S_{\text{Pretrain}} - S_{\text{Random}}}{S_{\text{Scratch}} - S_{\text{Random}}} \times 100\%\,, \qquad (13)$$

where $S_{\text{Pretrain}}$ is the score after 20M steps with the first 10M steps pretrained without access to extrinsic rewards, and $S_{\text{Scratch}}$ is the score after 10M steps of training from scratch. In 14, 15, and 17 out of 18 games (see Fig. 6), pretraining using mega-reward achieves more relative improvements than pretraining using the world model, MOREL and OOMDP, respectively. This shows that mega-reward is also very helpful for agents to achieve a superior performance when used in a domain with extrinsic rewards.

## Attention with Mega-Reward

Furthermore, "*noisy TV*" is a long-standing open problem in novelty-driven approaches (Burda et al. 2018; 2019); it means that if there is a TV in the state that displays randomly generated noise at every step, the novelty-driven agent will find that watching at the noisy TV produces great interest. A possible way to solve this problem is to have an attention mask to remove the state changes that are irrelevant to the agent, and we believe the accumulated latent control map $g_t^{h,w}$ can be used as such an attention mask. Specifically, we estimate a running mean for each grid in $g_t^{h,w}$, which is then used to binarize $g_t^{h,w}$. The binarized $g_t^{h,w}$ is used to mask the state used in the state-of-the-art novelty-driven work, *RND* (Burda et al. 2019), making RND generate novelty scores only related to the agent's control (both direct

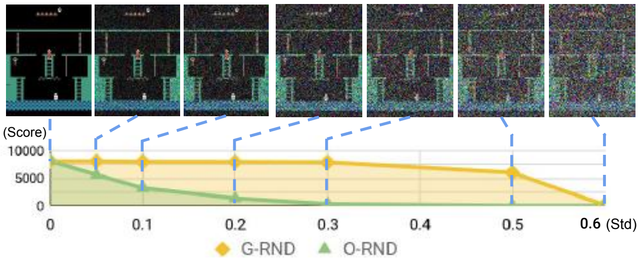

Figure 7: Comparing G-RND with O-RND over *MontezumaRevenge* of different noises.

Table 2: Comparing G-RND with O-RND and other baselines over 6 hard exploration Atari games with STD of $0.3$.

| Game | G-RND | A-RND | M-RND | O-RND |
|---|---|---|---|---|
| *MontezumaRevenge* | **7934** | 7138 | 2385 | 313 |
| *Gravitar* | **3552** | 3485 | 1634 | 1323 |
| *Pitfall* | -15.48 | -14.43 | **-13.64** | -14.23 |
| *PrivateEye* | 7273 | **7347** | 6128 | 2132 |
| *Solaris* | **3045** | 2857 | 2253 | 2232 |
| *Venture* | 1700 | **1701** | 1599 | 1572 |

or latent). There are two additional baselines, ADM (Choi et al. 2019) and MOREL (Goel, Weng, and Poupart 2018) that also generate segmentation masks, which can be used to mask the state in RND. Thus, we compare $g_t^{h,w}$ in our mega-reward with these baselines in terms of generating better masks to address the "*noisy TV*" problem.

Experiments are first conducted on *MontezumaRevenge*, following the same settings as in (Burda et al. 2019). Fig. 7 shows the performance of the original RND (O-RND) and $g_t^{h,w}$-masked RND (G-RND) with different degrees of noise (measured by the STD of the normal noise). The result shows that as the noise degree increases, the performance score of RND decreases catastrophically, while the performance drop of G-RND is marginal until the noise is so strong (STD $= 0.6$) that it ruins the state representation. Fig. 2 shows the performance of O-RND and RND masked with different baselines (G-RND for $g_t^{h,w}$-masked RND, A-RND for ADM-masked RND, and M-RND for MOREL-masked RND) over 6 hard exploration Atari games with STD of $0.3$. Results show that G-RND outperforms all other baselines, which means that $g_t^{h,w}$ generated in our mega-reward is the best mask to address the "*noisy TV*" problem. This further supports our conclusion that mega-reward can also achieve a superior performance when it is used together with extrinsic rewards.

## Failure Cases

Some failure cases of mega-reward are also noticed. We find that mega-reward works well on most games with a meshing size of $4 \times 4$; however, some of the games with extremely small or big entities may fail with this size. In addition, mega-reward also fails when the game terminates with a few seconds of flashing screen, because this will make the agent mistakenly believe that killing itself will flash the screen, which seems like having control on all entities for the agent. Another failure case is that when the camera can be moved by the agent, such as in the game *Pitfall* in Table 2. The experiment's first step, i.e., modeling direct control $\alpha(s_t^{h,w}, a_{t-1})$ via Eqs. (1) to (4) fails, as all grids are under direct control when the agent moves its camera. One possible solution for above failures is extracting the entities from the states using semantic segmentation (Goel, Weng, and Poupart 2018), then applying our method on the semantically segmented entities instead of each grid.

## Related Work

We now discuss related works on intrinsic rewards. Further related work on contingency awareness, empowerment, variational intrinsic control, and relation-based networks is presented in the extended paper (Song et al. 2019b).

Intrinsic rewards (Oudeyer and Kaplan 2009) are the rewards generated by the agent itself, in contrast to extrinsic rewards, which are provided by the environment. Most previous work on intrinsic rewards is based on the general idea of "novelty-drivenness", i.e., higher intrinsic rewards are given to states that occur relatively rarely in the history of an agent. The general idea is also called "surprise" or "curiosity". Based on how to measure the novelty of a state, there are two classes of methods: count-based methods (Bellemare et al. 2016; Martin et al. 2017; Ostrovski et al. 2017; Tang et al. 2017) and prediction-error-based methods (Achiam and Sastry 2017; Pathak et al. 2017; Burda et al. 2018; 2019). Another popular idea to generate intrinsic rewards is "difference-drivenness", meaning that higher intrinsic rewards are given to the states that are different from the resulting states of other subpolicies (Florensa, Duan, and Abbeel 2017; Song et al. 2019a). To evaluate intrinsic rewards, intrinsically-motivated play has been adopted in several state-of-the-art works. However, it may be an ill-defined problem, i.e., if we flip the extrinsic rewards, the agent only trained by the intrinsic rewards is likely to perform worse than a random agent in terms of the flipped extrinsic rewards. Discarding the possible bug in defining the problem, intrinsically-motivated play indeed helps in many scenarios, such as pretraining, improving exploration, as well as understanding human intelligence.

## Summary

In this work, we proposed a novel and powerful intrinsic reward, called mega-reward, to maximize the control over given entities in a given environment. To our knowledge, mega-reward is the first approach that achieves the same level of performance as professional human players in intrinsically-motivated play. To formalize mega-reward, we proposed a relational transition model to bridge the gap between direct and latent control. Extensive experimental studies are conducted to show the superior performance of mega-reward in both intrinsically-motivated play and real-world scenarios with extrinsic rewards.

**Acknowledgments.** This work was supported by the China Scholarship Council under the State Scholarship

Fund, by the Graduate Travel and Special Project Grants from the Somerville College of the University of Oxford, by the Alan Turing Institute under the UK EPSRC grant EP/N510129/1, by the AXA Reseach Fund, by the National Natural Science Foundation of China under the grants 61906063, 61876013, and 61922009, by the Natural Science Foundation of Tianjin under the grant 19JCQNJC00400, by the "100 Talents Plan" of Hebei Province, and by the Yuanguang Scholar Fund of Hebei University of Technology.

## References

Achiam, J., and Sastry, S. 2017. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

Baeyens, F.; Eelen, P.; and van den Bergh, O. 1990. Contingency awareness in evaluative conditioning: A case for unaware affective-evaluative learning. *Cognition And Emotion*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. In *NIPS*.

Bellemare, M. G.; Veness, J.; and Bowling, M. 2012. Investigating contingency awareness using Atari 2600 games. In *AAAI*.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2018. Large-scale study of curiosity-driven learning. In *NIPS*.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2019. Exploration by random network distillation. In *ICLR*.

Choi, J.; Guo, Y.; Moczulski, M.; Oh, J.; Wu, N.; Norouzi, M.; and Lee, H. 2019. Contingency-aware exploration in reinforcement learning. In *ICLR*.

Diuk, C.; Cohen, A.; and Littman, M. L. 2008. An object-oriented representation for efficient reinforcement learning. In *ICML*.

Florensa, C.; Duan, Y.; and Abbeel, P. 2017. Stochastic neural networks for hierarchical reinforcement learning. In *ICLR*.

Friston, K. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*.

Goel, V.; Weng, J.; and Poupart, P. 2018. Unsupervised video object segmentation for deep reinforcement learning. In *NIPS*.

Ha, D., and Schmidhuber, J. 2018. World models. In *NIPS*.

Jaderberg, M.; Mnih, V.; Czarnecki, W. M.; Schaul, T.; Leibo, J. Z.; Silver, D.; and Kavukcuoglu, K. 2017. Reinforcement learning with unsupervised auxiliary tasks. In *ICLR*.

Klyubin, A. S.; Polani, D.; and Nehaniv, C. L. 2005. All else being equal be empowered. In *ECAL*.

Klyubin, A. S.; Polani, D.; and Nehaniv, C. L. 2008. Keep your options open: An information-based driving principle for sensorimotor systems. *PloS one*.

Martin, J.; Sasikumar, S. N.; Everitt, T.; and Hutter, M. 2017. Count-based exploration in feature space for reinforcement learning. In *IJCAI*.

Martins, A., and Astudillo, R. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Mohamed, S., and Rezende, D. J. 2015. Variational information maximisation for intrinsically motivated reinforcement learning. In *NIPS*.

Montúfar, G.; Ghazi-Zahedi, K.; and Ay, N. 2016. Information theoretically aided reinforcement learning for embodied agents. *arXiv preprint arXiv:1605.09735*.

Ostrovski, G.; Bellemare, M. G.; van den Oord, A.; and Munos, R. 2017. Count-based exploration with neural density models. In *ICML*.

Oudeyer, P.-Y., and Kaplan, F. 2009. What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *ICML*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Song, Y.; Wang, J.; Lukasiewicz, T.; Xu, Z.; and Xu, M. 2019a. Diversity-driven extensible hierarchical reinforcement learning. In *AAAI*.

Song, Y.; Wang, J.; Lukasiewicz, T.; Xu, Z.; Zhang, S.; and Xu, M. 2019b. Mega-reward: Achieving human-level play without extrinsic rewards. *arXiv preprint arXiv:1905.04640*.

Song, Y. 2019. Released code. https://github.com/YuhangSong/Mega-Reward.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Chen, X.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. #Exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*.

Watson, J. S. 1966. The development and generalization of "contingency awareness" in early infancy: Some hypotheses. *Merrill-Palmer Quarterly of Behavior and Development*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.