MODELING BOUNDED RATIONALITY IN MULTI-AGENT SIMULATIONS USING RATIONALLY INATTENTIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent reinforcement learning (MARL) is a powerful framework for studying emergent behavior in complex agent-based simulations. However, RL agents are often assumed to be rational and behave optimally, which does not fully reflect human behavior. Here, we study more human-like RL agents which incorporate an established model of human-irrationality, the Rational Inattention (RI) model. RI models the cost of cognitive information processing using mutual information. Our RIRL framework generalizes and is more flexible than prior work by allowing for multi-timestep dynamics and information channels with heterogeneous processing costs. We evaluate RIRL in Principal-Agent (specifically manager-employee relations) problem settings of varying complexity where RI models information asymmetry (e.g. it may be costly for the manager to observe certain information about the employees). We show that using RIRL yields a rich spectrum of new equilibrium behaviors that differ from those found under rational assumptions. For instance, some forms of a Principal's inattention can increase Agent welfare due to increased compensation, while other forms of inattention can decrease Agent welfare by encouraging extra work effort. Additionally, new strategies emerge compared to those under rationality assumptions, e.g., Agents are incentivized to misrepresent their ability. These results suggest RIRL is a powerful tool towards building AI agents that can mimic real human behavior.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has shown great utility in complex agent-based model (ABM) simulations, e.g., in economics, games, and other fields (Zheng et al., 2020; Baker et al., 2020; Leibo et al., 2017; Yang et al., 2018). ABM is a field of research which has been found important and useful for quantifying the potential impacts of various policies in diverse real world applications (eg. market (Bozanta & Nasır, 2014) or environmental impact (Raihanian Mashhadi & Behdad, 2018) simulations) by examining real-world systems through simulation. In many of such simulations, the behavioral rules of agents may be too difficult for designers to specify. With MARL, designers instead specify the agents' objective functions and the reinforcement learning (RL) agents autonomously learn behaviors to optimize their objectives. However, this approach may be problematic when simulating systems of *human* agents, because in contrast to established models of human decision-making, it assumes that agents behave rationally and execute the objective-maximizing behavior. Prior literature demonstrates that using established models of human-irrationality during decision making yields results and implications that are significantly different from those obtained using rationality assumptions (Sims, 2003; Maćkowiak & Wiederholt, 2011; Jiang et al., 2019). Therefore, it is important to account for this irrationality when simulating systems involving human(-like) agents.

To this end, we introduce Rational Inattention Reinforcement Learning (RIRL), a MARL framework with agents that can be rationally inattentive. Rational Inattention (RI) is a model of bounded rationality (Sims, 2003) that is well established in behavioral economics (Mackowiak et al., 2021). It attributes human irrationality to the *costliness of mental effort* (e.g. attention) required to identify the optimal action. Mathematically, the RI framework measures these costs as the *mutual information* (MI) between the variables considered by the decision process and the decisions ultimately made.

This captures the intuition that a policy has a higher cognitive cost if its execution requires more information about the state of the world and thus more attention. MI-based rewards have been used in RL, (e.g. Leibfried & Braun (2016); Leibfried et al. (2018); Grau-Moya et al. (2018); Leibfried & Grau-Moya (2020) use RI-style MI costs to regularize learning) but not to model sub-optimal behavior. When used to model sub-optimal behavior, RIRL can rationalize seemingly sub-optimal behavior by including this cognitive cost in the reward function, i.e., by adding the MI cost(s). That is, the "rational" behaviors of an RIRL-actor can mimic human-like bounded rationality.

Contributions. The main contributions of this work are: (1) we propose RIRL, a framework for modeling bounded-rationality which can analyze complex scenarios not previously possible; and (2) we validate RIRL's modeling capabilities in classical economic settings intractable using the analytical methods of prior work. *To our knowledge, we are the first to study the effects of Rational Inattention in MARL-based simulations.*

RIRL extends the single-timestep framework proposed by Peng et al. (2017), which decomposes decision-making into stochastic perception followed by stochastic action, each subject to its own MI cost. RIRL provides a novel boundedly-rational policy class that generalizes this to multiple information channels with heterogeneous costs, hidden-state policy models, and sequential environments. To achieve this increased modeling power, our architecture uses methods significantly different from prior work, such as using deep mutual information estimation modules to allow for multiple channels and using an LSTM to account for multiple timesteps. This allows RIRL to analyze settings with rich cognitive cost structures, e.g., when information about state variables have different observation difficulty. For example, in hiring, a candidate's past job performance may be more relevant but harder to evaluate than her employment history.

We evaluate RIRL in two *Principal-Agent* (PA) problems, specifically manager-employee relations, where the Principal is boundedly rational. We use RIRL to model asymmetric information (i.e. hard to obtain information) between the Principal and the Agent they are incentivizing. Real-world PA experiments have shown that bounded rationality is key to explaining marked deviations between equilibria reached by human participants and theoretical predictions (Erlei & Schenk-Mathes, 2017). Prior work considered bounded rationality assumptions but found it to be analytically difficult (Mirrlees, 1976) and further research in this direction is sparse

We show that RIRL allows us to analyze generalized PA problems that are analytically intractable, such as a sequential PA problem with multiple Agents and heterogeneous information channels. Across all settings, we observe that equilibrium implications depends strongly on the cost(s) of attention and differs from that under rational assumptions sometimes in ways that are hard to intuitively predict. We observe that, depending on the channel, increasing Principal inattention can either increase Agent welfare due to increased compensation or decrease Agent welfare due to encouraging additional work. Compared to under rational assumptions, we additionally observe the emergence of different strategies, such as Agents choosing to misrepresent their ability, classically referred to as *signaling* (Spence, 1973). Our results show that RIRL can be a powerful tool to model boundedly rational behavior and analyze its emergent consequences in multi-agent systems.

2 Related Works

Multi-Agent Reinforcement Learning for Agent-Based Simulation. Agent-Based Models (ABM) are a popular tool use to study real-world systems (e.g., organizations or market systems) and discover emergent behavior through simulation with fixed or manually-specified agent behaviors (Bonabeau, 2002). Instead, Multi-Agent Reinforcement Learning (MARL)-based simulations use RL agents which autonomously learn utility-maximizing behavior, so designers do not need to specify behavioral rules. MARL has been used to study tax policy (Zheng et al., 2020; 2021; Trott et al., 2021), games (Baker et al., 2020), and social dilemmas (Leibo et al., 2017) among others (Yang et al., 2018). However, RL agents are mostly assumed rational which contradicts how humans make decisions. Some recent work considers bounded rationality in MARL by accounting for cognitive limitations when reasoning about the behaviors of other agents (Evans & Prokopenko, 2021; Wen et al., 2019; Łatek et al., 2009). We model a complementary source of bounded rationality: the cognitive costs of attending to all available information when forming a decision.

Models of Irrationality. Behavioral economics has shown extensively that human decision-making is not fully rational, but instead features many cognitive biases (Caverni et al., 1990). Bounded rationality (Simon, 1957) attributes these human irrationalities to resource limitations, e.g., bounded cognitive capabilities or costliness of using (more) time to make decisions. Rational Inattention (RI) is a well-established model of bounded rationality which models the cognitive cost of decisions as the mutual information between variables relevant to the decision and the decision itself (Sims, 2003; Mackowiak et al., 2021). RI has been tested in real world experiments (Dean & Neligh, 2017) and used to model human behavior in a wide variety of domains (Hoiles et al., 2020; Mackowiak et al., 2021). Closest to our work, Jiang et al. (2019) study a multi-agent system for traffic route choice under RI but where the agents do not react to each other's actions.

Another line of work uses human behavioral data with deep learning to model human cognition (Kubilius et al., 2019; Battleday et al., 2017; Ma & Peters, 2020) or predict human behavior (Bourgin et al., 2019; Kolumbus & Noti, 2019; Hartford, 2016). In contrast, our ABM approach examines the implications of RI in complex settings where experimental data are presently unavailable.

Mutual Information in Reinforcement Learning. MI has been extensively studied in the intrinsically motivated RL literature for curiosity-driven exploration and unsupervised skill and option discovery (Still & Precup, 2012; Campos et al., 2020; Mohamed & Rezende, 2015; Gregor et al., 2016; Eysenbach et al., 2019), often with differing techniques used to measure MI. Recently MI-based rewards have been used to regularize exploration (Grau-Moya et al., 2018; Leibfried & Grau-Moya, 2020). Comparatively, very little work has considered MI-based rewards for modeling boundedly rational behavior (Ortega et al., 2015; Peng et al., 2017), with no prior work, to our knowledge, considering the domain of multi-agent simulations.

Principal-Agent Problems. In addition to our guiding example, Principal-Agent problems have been used to describe various settings, including politics and insurance (Miller, 2005; Grossman & Hart, 1992). Prior economics literature mainly use analytical methods and narrow modeling assumptions, e.g., that all stochasticity follows Brownian motion (Sannikov, 2008) or certain separability conditions are satisfied (Grossman & Hart, 1992). Generally, Principal-Agent settings fall under the category of Stackelberg games where it has been shown computing optimal strategies can be NP-hard (Korzhyk et al., 2010). Instead, MARL can study complex setups that are analytically intractable. Shu & Tian (2019); Shi et al. (2019); Ahilan & Dayan (2019) studied coordinating cooperation in a Principal-Agent model, with an RL Principal learning to incentivize Agents to achieve an overall goal. However, they do not consider bounded rationality.

3 PRELIMINARIES

We consider multi-agent simulations in the form of partially-observable Markov Games (MGs), formally defined by $(S, A, r, \mathcal{T}, \gamma, O, \mathcal{I})$ (Sutton & Barto, 2018). Here S is the state space of the game, A is the combined action spaces of the actors, and $\mathcal{I}$ are actor indices. We use $o_i = O(s, i)$ to denote the portion of the game state s that actor i can observe. In addition, o_i may include a (possibly learnable) encoding of the observation history. Each game episode has a horizon of $T \geq 1$ timestep(s). Each timestep t, actor i selects action $a_{i,t}$, sampling from its policy $\pi_i\left(a_{i,t}|o_{i,t}\right)$. Given the sampled actions, the transition function $\mathcal{T}$ determines how the state evolves. Each actor's objective is encoded in its reward function $r_i(s, \boldsymbol{a})$, (boldface denotes concatenation across actors). When modeling economic behavior, the reward is the (marginal) utility $U_i(s_t, \boldsymbol{a}_t)$, an abstract measure of happiness. Each actor optimizes its policy to maximize its γ-discounted sum of future rewards.

While RL can be used to discover (approximately) utility-maximizing policies, such rational behavior fails to account for characteristic *irrational* human behavior. Behavioral economic models often model irrationality as consequences of *inattention* (Gabaix, 2017). Rational Inattention formalizes this intuition using a modified objective that includes a cost to the mutual information $I(a_i; o_i)$ between the (observable) state of the world o_i and the actions $a_i \sim \pi_i(\cdot|o_i)$. This definition captures the intuition that if the agent puts in more effort to pay attention to o_i, its action a_i likely becomes more correlated with the observation o_i, and thus high MI.

$$\pi_i^{\dagger} = \arg\max_{\pi_i} \left(\mathbb{E}_{\boldsymbol{\pi}} \left[U_i(s, \boldsymbol{a}) \right] - \lambda I(a_i; o_i) \right). \tag{1}$$

Note that this is equivalent to learning the optimal policy for an adjusted reward function: $r_i^\dagger(s_t, \boldsymbol{a}_t) = U_i(s_t, \boldsymbol{a}_t) - \lambda \tilde{I}(a_{i,t}; o_{i,t})$, where $\tilde{I}(a_{i,t}; o_{i,t}) = \log p(a_{i,t}, o_{i,t}) - \log p(a_{i,t})p(o_{i,t})$ is a Monte Carlo estimate of $I(a_i; o_i)$, and λ is the utility cost per bit of information. Here $p(a_i, o_i)$, $p(a_i)$, and $p(o_i)$ are the joint and marginal distributions over a_i and o_i induced by the environment and policies $\boldsymbol{\pi}$. Below, we describe our method for estimating these quantities during training.

4 Modeling and Training Boundedly Rational Actors with RIRL

MI Estimation. RIRL uses a general-purpose module for estimating $\tilde{I}_{\pi_i}(a_i; o_i)$, i.e., the single-sample MC estimate of the mutual information between o_i and $a_i \sim \pi_i(a_i|o_i)$. Given the pair (a_i, o_i), we estimate $\tilde{I}_{\pi_i}(a_i; o_i)$ from the ratio between $\log p(a_i, o_i)$ (the log-odds under the joint distribution) and $\log p(a_i)p(o_i)$ (the log-odds under the factorized distributions). This ratio can be estimated using a discriminator $d_{\pi_i}(a_i, o_i)$ that learns to classify which of the two distributions the sample (a_i, o_i) came from. Samples from $p(a_i, o_i)$ are generated during on-policy rollouts, while samples from $p(a_i)p(o_i)$ can be generated by shuffling a batch of samples from the joint distribution. As such, on-policy rollout data can be used both to optimize π and to train d_{π_i}, and compute $r_{i,t}^\dagger = r_{i,t} - \lambda \tilde{I}_{\pi_i}(a_{i,t}; o_{i,t})$.

While we favor the above approach for its simplicity, other techniques for estimating MI have also been proposed, e.g. Belghazi et al. (2018). Our framework is not restricted to the choice of MI estimator used here, but a comparison of these techniques is outside the scope of the current work.

Action-Perception Decoupling with Multiple Information Channels. Penalizing $I(a_i; o_i)$ models the intuition that it is costly to obtain or use information about the world, e.g., to reduce uncertainty about $o_{i,t}$ or to identify the utility-maximizing action given $o_{i,t}$. To support richer modeling, we also extend our framework with *multiple channels of information with heterogeneous cognitive costs*. For example, when buying a used car it is easier to see car prices than their actual condition.

To that end, we extend the "action-perception decoupling" strategy of Peng et al. (2017), which models $\pi(a|s)$ as a stochastic perception module $q(y|s)$ followed by an action module $p(a|y)$, jointly trained to optimize an RI-style reward $r(s, a) - \lambda_q I_q(y; s) - \lambda_p I_p(a; y)$. We extend this using a policy class that can flexibly model M information channels (i.e., partitions of o) with different processing costs λ^m. We also use recurrent policies which may allocate processing costs strategically over time.

More formally, we assume that an observation o_t can be decomposed as a set of $M \geq 1$ observations $o_t = \{o_t^1, \ldots, o_t^M\}$, with o_t^m being the observation from information channel m. In addition, we assume that each channel has an associated information cost: $\lambda = \{\lambda^1, \ldots, \lambda^M\}$. For each channel, we learn an encoder $f^m(y_t^m | o_t^m, \psi_t)$, which takes o_t^m and recurrent state ψ_t (see below) as inputs and outputs the parameters (means and standard deviations) of a stochastic encoding y_t^m. In practice, we implement each f^m as a residual-style encoder, with samples given by:

$$\mu_t^m, \sigma_t^m = f^m(o_t^m, \psi_t), \quad y_t^m = o_t^m + \mu_t^m + \sigma_t^m \cdot \epsilon_t^m, \tag{2}$$

where ϵ_t^m is a random sample from a spherical Gaussian with dimensionality equal to that of y^m and o^m. For each f^m, we also train a discriminator $d_{f^m}(y_t^m, [o_t^m, \psi_t])$ used to estimate $\tilde{I}_{f^m}(y_t^m; [o_t^m, \psi_t])$. The full encoding $y_t = \left[y_t^1, \ldots, y_t^M\right]$ of o_t concatenates all M encoder samples. Before sampling an action, we use an LSTM (Hochreiter & Schmidhuber, 1997) to maintain a history of (encoded) information: $\psi_{t+1} = \texttt{LSTM}(y_t, \psi_t)$. The encoding y_t and updated LSTM state ψ_{t+1} are used as inputs to the stochastic action module $\omega(a_t | y_t, \psi_{t+1})$ which outputs a distribution over actions.

Training. We train this architecture with policy gradients (Williams, 1992):

$$\Delta\pi \propto \mathbb{E}\left(\nabla \log \pi(y_t^1, \ldots, y_t^M, a_t | s_t, \psi_t, \psi_{t+1}) r_t^\dagger\right), \tag{3}$$

$$\log \pi(y_t^1, \ldots, y_t^M, a_t | s_t, \psi_t, \psi_{t+1}) = \log \omega(a_t | y_t, \psi_{t+1}) + \sum_{m=1}^{M} \log f^m(y_t^m | o_t^m, \psi_t), \tag{4}$$

$$r_t^\dagger = U(s_t, \boldsymbol{a}_t) - \lambda^\omega \tilde{I}_\omega(a_t; [y_t, \psi_{t+1}]) - \sum_{m=1}^{M} \lambda^m \tilde{I}_{f^m}(y_t^m; [o_t^m, \psi_t]). \tag{5}$$

5 VALIDATING RIRL IN PRINCIPAL-AGENT PROBLEMS

We demonstrate RIRL's modeling capabilities in two Principal-Agent (PA) problems of varying complexity. In both settings, a boundedly rational manager (the Principal) decides how to compensate its workers (the Agents). In general, labor is costly to the Agents but beneficial to the Principal, while the Agents benefit from income that is costly for the Principal to provide. Therefore, the Principal aims to find a wage schedule $\mathcal{W}$ that maximizes its profits. PA problems often consider how *information asymmetry* between the Principal and Agent(s) influences equilibrium schedules. RIRL can analyze a meaningful extension in which we model how the *cost of information* influences equilibrium schedules, i.e., where the Principal can spend effort to reduce information asymmetry.

We first study a bandit-style PA setting with a Principal that optimizes a wage schedule for each level of Agent output, subject to a cognitive cost for observing output. Because some aspects of the solution are tractable with prior analytical methods, this simple setting provides a useful validation of RIRL in a multi-Agent game. However note, even in this simple setting, aspects of the solution (such as the constants) are not tractable analytically. Secondly, we apply RIRL to analyze a complex sequential PA problem with multiple Agents and heterogeneous information channels that is outside the scope of analytical methods of prior work.

Our experimental results were able to uncover several non-trivial insights when bounded rationality is assumed that are hard to intuitively predict, demonstrating the importance of modeling bounded rationality in such simulations. Insights include (1) Agent utility can increase with principal inattention and a more rational Agent benefits more from Principal inattention due to the Principal setting higher pay schedules. To our knowledge, no prior work has identified the consequences of bounded rationality in both the Principal and Agent. (2) In the case of where there are multiple Agents, a boundedly rational Principal can induce Prisoner's Dilemma like incentives for the Agents, reducing the welfare of all agents. We discuss these insights along with others in detail below.

5.1 PRINCIPAL-AGENT WITH A SINGLE AGENT, SINGLE TIMESTEP (BANDIT SETTING)

We first explore RIRL in a PA setting with a single Agent, following classic work (Mirrlees, 1976; Holmstrom & Milgrom, 1987; Spremann, 1987; Haubrich, 1994). The Agent's labor output $z \sim h(e)$ is a stochastic function of its effort action e. The Agent chooses how much to work based on its payment schedule $\mathcal{W}$, which is controlled by the Principal and yields payment w as a stochastic function of output z, i.e., $w \sim \mathcal{W}(z) = \mathcal{N}(\mu_z, \sigma_z)$. Here, the stochasticity in the pay schedule aims to model the Principal's uncertainty over z, e.g., due to limited attention. By assumption, the Agent accurately anticipates the shape of the payment schedule $\mathcal{W}$ when making its decisions, e.g., based on prior experience or reputation. As such, the RIRL-Principal sets the payment schedule:

$$\mathcal{W}^* = \arg\max_{\mathcal{W}} \left[\mathbb{E}\left[U_p(w,z)\right] - \lambda I_{\mathcal{W}}(w;z)\right]_{w\sim\mathcal{W}(z), z\sim h(e), e\sim\pi_a^\beta(e|\mathcal{W})}. \tag{6}$$

Here, $\lambda I_{\mathcal{W}}(w;z)$ is the attention cost for $\mathcal{W}$. For the Agent, we use a soft-Q policy $\pi_a^\beta(e|\mathcal{W}) = e^{\beta U_a(e|\mathcal{W})}/Z$ with normalization Z, where $U_a(e|\mathcal{W})$ denotes the *expected* utility of effort e given schedule $\mathcal{W}$. Note that π_a^β can be calculated directly. Importantly, *the soft-Q policy models a form of bounded rationality*, i.e., the log-odds of effort e is proportional to its expected utility, with higher β making "better" actions e more likely. Hence, low β corresponds to less-than-rational behavior.

The utility of the Principal (U_p) and Agent (U_a) follow standard economic utility functions:

$$U_a(w,e) = \underbrace{\texttt{CRRA}(w;\rho)}_{\text{Income Utility}} - \underbrace{e}_{\text{Work Disutility}} \quad \text{and} \quad \underbrace{U_p(w,z)}_{\text{Profit}} = \underbrace{z}_{\text{Revenue}} - \underbrace{w}_{\text{Amount Paid}} \tag{7}$$

Here CRRA is the concave Constant Relative Risk Adverse function (Pratt, 1978); the risk aversion parameter ρ sets its concavity (we use $\rho = 2$). More environment details are in the Appendix. We use N-sample policy gradients on estimates of Equation 6 to learn the parameters μ_z and σ_z of $\mathcal{W}^*$

$$r^\dagger_{\mathcal{W}} = \frac{1}{N}\sum_{i=1}^{N}\left[U_p(w_i,z_i) - \lambda\tilde{I}_{\mathcal{W}}(w_i;z_i)\right], \quad w_i \sim \mathcal{W}(z_i), z_i \sim h(e_i), e_i \sim \pi_a^\beta(e|\mathcal{W}_i). \tag{8}$$

Results. Figure 1 shows the effects of the Principal's inattention on simulated outcomes. We also compare RIRL outcomes against those found including entropy-based rewards. MI-based and

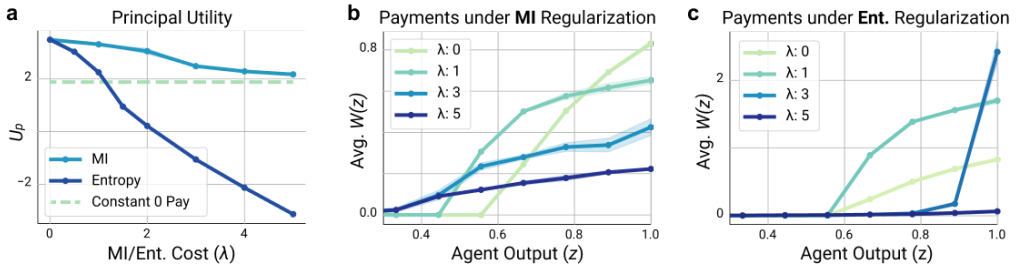

Figure 1: Bandit Experiment Results. All results are averaged over 5 random seeds and plotted with 95% confidence intervals. **(a)** Comparing MI and Entropy regularization on Principal utility. The constant 0-pay schedule provides a meaningful lower bound. **(b, c)** Pay schedule means under MI (b) and Entropy (c) regularization (pay schedule standard deviations are plotted in the Appendix).

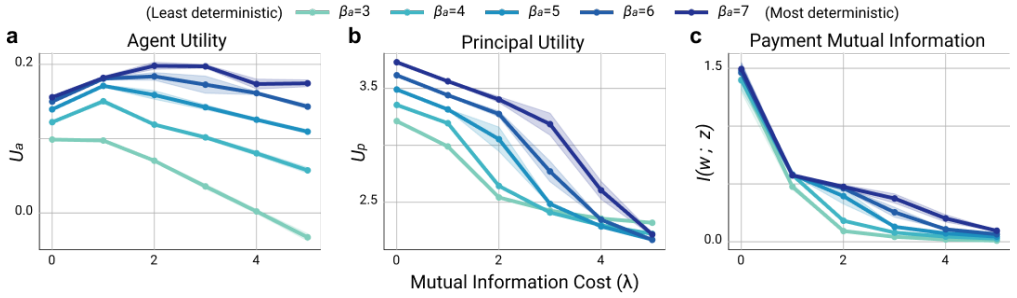

Figure 2: Additional Bandit Experiment Results. All results are averaged over 5 random seeds and plotted with 95% confidence intervals. Results are shown for each level of MI cost (λ, x-axis) across multiple levels of Agent policy temperatures (β, color). **(a, b)** Agent and principal utility. **(c)** Mutual Information between output z and payment w.

entropy-based regularization are closely related (Grau-Moya et al., 2018; Leibfried & Grau-Moya, 2020), with subtle but important differences[1]. Nevertheless, they yield markedly different outcomes.

To highlight an important difference, consider the constant 0-pay schedule. This provides no incentive for the Agent to work and costs the Principal nothing, and it therefore provides a reasonable lower bound on U_p. This lower bound should hold under bounded rationality, since no attention is required when $\mathcal{W}$ treats all outputs identically. Indeed, under RIRL, U_p approaches this lower bound as increasing λ leads the Principal to trade more profitable $\mathcal{W}$'s for ones with smaller demands on attention (Figs. 1a, b, 2c). In contrast, under entropy regularization, increasing λ quickly yields $\mathcal{W}$'s that violate the U_p lower bound (Fig. 1a, c).

Figure 2 shows how Agent and Principal utility and $I(w; z)$ change as a function of λ, across a range of Agent β's. Note that higher β increases the odds that the Agent selects the optimal action. This reveals an interesting multi-Agent interaction. As before, we observe that the Principal's utility U_p (without the attention cost) decreases with increasing λ. Furthermore, U_p is generally lower when β is lower (Fig. 2b), as the Principal must pay more to influence π_a^β towards different actions.

Interestingly, we observe that *the Agent's utility U_a can actually increase when the Principal is inattentive*, and that higher β tends to increase the beneficial range of inattention (Fig. 2a). Lastly, β determines how the Principal responds to increasing attention costs. Agents that behave optimally more often (higher β) incentivize the Principal more to pay attention (Fig. 2c).

Comparison With Prior Literature. For this setting, Mirrlees (1976) found the theoretically optimal pay schedule should be of the form $\mathcal{W}(z) = \max(Az + B, C)^{\frac{1}{\rho}}$. While they assume a slightly

[1]Under some conditions, they are identical. However, the key difference is that entropy regularization penalizes the policy for deviating from a *fixed, uniform* prior, whereas RI-style MI regularization penalizes deviations from an *optimal* prior.

Figure 3: A depiction of a single timestep in an episode of our Sequential Multi-Agent Setting. Green variables (such as cumulative output) are not costly while blue variables (such as individual output) are costly for the principal to observe. The red variable (Ability) cannot be observed.

different model of costly attention[2], the emergent pay schedules match closely with this work: when fitting A, B, and ρ to our data, we measure $r^2 = 0.99$ for all β. The best fit ρ was close to the theoretical value for some values of β (true ρ: 2, best fit ρ: 2.19 for $\beta = 5$). Interestingly, we do observe that the best fitting ρ tends to increase with β in our model. Notably, the influence of Agent sub-optimality was outside the scope of this prior work. We are able to observe this relationship between ρ and β through the modelling complexity enabled by our RIRL framework.

5.2 Principal-Agent in the Sequential Multi-Agent Setting

We show that RIRL enables modeling sequential Principal-Agent problems with multiple Agents with $T > 1$ timesteps. We consider horizons $[2, 10]$ and teams of $n_a = 4$ Agents. We assume there are $K = 5$ possible Agent abilities k. The Principal cannot see the Agents' abilities. Each Agent's ability is sampled randomly at the start of each episode. At each timestep, Agent i's output is:

$$z_{i,t} = h_{i,t} \cdot (\nu_i^k + e_{i,t}), \tag{9}$$

where it works $h_{i,t}$ hours and exerts effort $e_{i,t}$. The Principal moves first and sets a wage $w_{i,t}$. Each Agent moves second: it knows $w_{i,t}$ before choosing $h_{i,t}$ and $e_{i,t}$, and earning income $w_{i,t} \cdot h_{i,t}$. As before, Principal utility U_p measures profit. We define U_a following standard utility functions, where the optimal h increases with w. As a consequence of this configuration (demonstrated below), the profit-maximizing wage w_i for Agent i increases with its ability ν_i^k. The Agent utility is:

$$U_a(w, h, e) = \underbrace{\texttt{CRRA}(w \cdot h; \rho)}_{\text{Income Utility}} - \underbrace{c_l h^\alpha (1 + e)}_{\text{Work Disutility}}, \quad \underbrace{U_p(\boldsymbol{w}, \boldsymbol{h}, \boldsymbol{z})}_{\text{Profit}} = \underbrace{\sum_{i \in [n_a]} z_i}_{\text{Revenue}} - \underbrace{\sum_{i \in [n_a]} w_i h_i}_{\text{Wages Paid}}, \tag{10}$$

where ρ, c_l and α are constants governing the shape of U_a.

Attention Costs. Because a strategic Principal must *infer* private Agent features, e.g., ability, its equilibrium behavior depends on any inference costs it experiences, e.g., attention costs. Note that, unlike in the above bandit setting, here we train *both* the Principal and the Agent policies, each of which is modeled using the RIRL-actor architecture (Section 4). However to isolate and explore the effects of distinct Principal attention costs, we do not impose any attention costs on the Agents. Therefore, the Agents' reward is only their utility $r_{i,t} = U_a(w_{i,t}, h_{i,t}, e_{i,t})$.

For the Principal, we model $M = 3$ information channels: one is "easy" and low-cost to observe and two are "hard" and high-cost. The low-cost channel o_p^f includes information that we regard as freely available ($\lambda^f = 0$), e.g., the time t, the hours worked $\boldsymbol{h}$ (workers often fill out timesheets which makes h_i easy to see), and the *total* output, $Z = \sum_{i \in [n_a]} z_i$ (managers can see the final result). However, it is high-cost to see *individual* contributions: we use a high-cost channel o_p^e for efforts $\boldsymbol{e}$, and a high-cost channel o_p^z for outputs $\boldsymbol{z}$. This models a Principal who can spend time and attention to observe individual Agents to reduce uncertainty about their true ability, e.g., their working styles and productivity (Figure 3). Modeling the attention cost of output and effort separately lets us study the effects and interactions of unequal observation costs. The Principal's reward is

$$r_{p,t}^\dagger = U_p(\boldsymbol{w}_t, \boldsymbol{h}_t, \boldsymbol{z}_t) - \underbrace{\lambda^z \tilde{I}(y_t^z; \boldsymbol{z}_t)}_{\text{Individual Output and}} - \underbrace{\lambda^e \tilde{I}(y_t^e; \boldsymbol{e}_t)}_{\text{Effort Perception Cost}} \tag{11}$$

[2]Their Principal pays a cost to reduce the *noise* added to its observed output, as opposed to our MI cost; see the Appendix for full details.

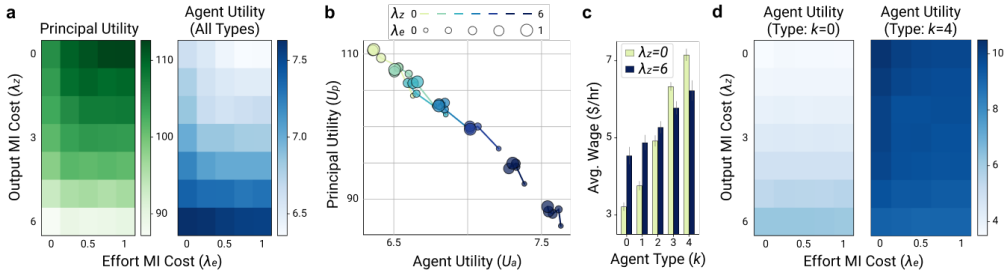

Figure 4: Sequential, multi-agent experiment results. All results are averaged over 20 runs and generated with $T = 5$. **(a, b)** Principal and Agent utility heatmaps (a) and scatter plot (b), for each λ^z and λ^e. **(c)** Average wage for each Agent type, under a rational ($\lambda^z = 0$, yellow) and a boundedly rational ($\lambda^z = 6$, blue) Principal. Error bars denote to 95% confidence intervals. **(d)** Utility for the lowest ($k = 0$) and highest ($k = 4$) ability Agent types, for each λ^z and λ^e.

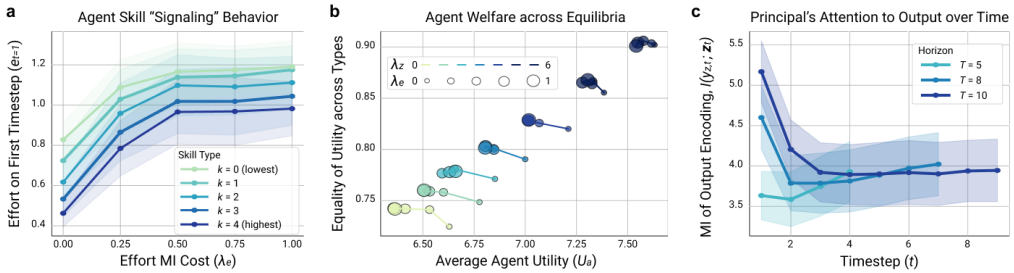

Figure 5: Additional sequential, multi-agent experiment results. All results are averaged over 20 runs. Shaded regions denote to 95% confidence. All results except (c) were generated with $T = 5$. **(a)** Average effort at $t = 1$ for each Agent type. **(b)** Equality across Agent types vs. Agent utility, for each λ^z and λ^e. **(c)** Output information $\tilde{I}_{f^z}(y_t^z; \boldsymbol{z}_t)$ encoded over time (when $\lambda^z = 3$).

Thus, the Principal's bounded rationality is modeled through the cost to get information about effort and individual outputs. The Principal's RIRL-actor architecture would also let us use attention costs $\tilde{I}(y_t^f; o_t^f)$ and $\tilde{I}(\boldsymbol{w}_t; y_t)$, but we omit those here[3]. Additional training details are in the Appendix.

Results. We highlight several noteworthy observations revealed by our RIRL framework: (1) Principal inattention to individual outputs $\boldsymbol{z}$ and inattention to efforts $\boldsymbol{e}$ have distinct, nearly opposite, consequences on actor utilities; (2) Agents respond to the incentive structures created by Principal inattention to effort, with parallels to a prisoner's dilemma; (3) The temporal patterns of Principal attention reflect the value of information over time.

Welfare Implications of Modeling Bounded Rationality. We analyze welfare related to actors' utilities. At equilibrium, the Principal and Agent utilities are negatively correlated (Figure 4a,b), comparing across varying levels of λ^z and λ^e. Note that the "rational" model has $\lambda^z = \lambda^e = 0$. This indicates that the Principal's bounded rationality has opposing implications for the Principal and Agents. From Figure 4a, we see that Agents' average utility increases and the Principal's utility decreases when the Principal's attention cost for individual outputs λ^z increases. Conversely, Agent utility decreases with increasing Principal attention cost on effort λ^e.

The cause for this is as follows. The Principal has a different optimal wage for each ability ν^k as shown in Figure 4c. When the Principal is fully rational it can use output and effort to accurately infer ability. Similar to what was seen in the bandit setting (above), increasing attention costs (in this case, on output λ^z) lead the Principal to set wages in a manner that is less profitable but also less attentionally demanding. At the resulting equilibria, *the Principal has more uncertainty over each Agent's type and adopts a "better safe than sorry approach" and increases the average wage to ensure output.* In sum, the same force that creates a lower-utility equilibrium for the Principal has higher average utility for the Agents.

[3]We do add entropy regularization over $\omega(a|y)$ for both the Principal and Agent to encourage exploration.

However, while the *average* Agent utility increases with λ^z, it does not increase for all Agent types. Specifically, the utility of the (highest) lowest-ability Agent's (decreases) increases. Hence, the Principal's uncertainty over individual outputs decreases the wage (and utility) differences between Agents of different ability (Fig. 4c,d). This is particularly relevant when considering the *equality* of utility. A common inequality metric is the Gini coefficient (Gini, 1971), computed as a normalized sum of income differences; equality can be defined as $\texttt{eq} = 1 - \frac{N}{N-1}\texttt{gini}$. Figure 5b shows that Agents' average utility *and* equality across types both increase for higher output attention cost λ^z.

Bounded Rationality induces Social Dilemma Dynamics. This setup gives rise to interesting temporal strategic behaviors akin to *signalling* (Spence, 1973). Note that, while effort e increases an Agent's work disutility without increasing its income, Agents may still choose to exert non-zero effort $e > 0$. To provide an intuition for this, first, note that spending effort increases output rate (z per hour h), resembling higher ability ν^k. Second, the profit-maximizing wage w_i increases with Agent ability[4] ν_i^k. The Principal can't see each ν_i^k but can see the output z_i and hours h_i. Hence, output rate is a "signal" of ability and Agents can misrepresent their ability by spending effort.

When λ^e increases, it becomes costly for the Principal to distinguish between the Agent's ability and effort-action. Interestingly, this leads to *signaling equilibria* in which Agents choose to use more effort (Fig. 4a), resulting in lower average Agent utility and higher Principal utility (Fig. 5a). In effect, this Agent behavior has parallels to the "defect-defect" equilibrium in the prisoner's dilemma (Shoham & Leyton-Brown, 2008). When higher-ability Agents are not using effort, lower-ability Agents are incentivized to use effort to misrepresent themselves as having higher ability. This incentivizes the higher-ability Agents to also use effort to distinguish themselves from the lower-ability Agents. Figure 5a shows this effect: effort increases with λ^e at all ability levels. In effect, increasing λ^e creates equilibria where the Principal enjoys higher Agent effort essentially for free. Interestingly, the effects of effort being costly to observe are nullified if the individual outputs are also hard to observe (Fig. 5b), showing these two cost parameters interact.

RIRL-actors Learn the Time-Value of Information. We also show that RIRL can discover the time-value of information-acquisition and how it depends on the time horizon. Figure 5c shows output information that the Principal encodes $\tilde{I}_{f^z}(y_t^z; \boldsymbol{z}_t)$ over time. There is an initial spike in the amount of information at the beginning of the episode and this decreases over time. Intuitively, it is most efficient for the Principal to pay attention to outputs at the beginning of the episode, as this information can be used throughout the episode. Additionally, the initial spike is larger for longer horizons, as initial information has more value for longer episodes.

Limitations. While RI is a general model of boundedly rationality, it does not cover all models of human irrationality. For simulations that wish to use more specialized models of bounded rationality, RIRL may not be sufficient. Examples include hyperbolic discounting (Kirby & Herrnstein, 1995) to model time-inconsistent delay discounting (humans tend to prefer rewards in the near future much more than rewards farther in the future) or prospect theory (Tversky & Kahneman, 1992) to model loss-aversion. Future research may extend RIRL to include such notions of bounded rationality.

6 Conclusion

We propose a novel framework, RIRL, and associated policy class for modeling bounded-rationality in MARL simulations with complexity beyond the scope of prior techniques. Our method incorporates *Rational Inattention*, an established model of human bounded rationality which uses mutual information to model the cognitive cost of processing information. Mutual information has been used extensively in RL for exploration and skill discovery, but, to our knowledge, we are the first to incorporate it for modeling human-like behavior in multi-agent simulations. We evaluate our method in two Principal-Agent problem settings, including a complex multi-agent setting, with multiple information channels with heterogeneous costs. Incorporating bounded rationality leads to different actor strategies and welfare outcomes as compared to under rational assumptions. These results establish RIRL as a promising framework for using MARL to analyze systems of human agents.

[4]This is a simple consequence of how utility functions are defined. See Fig 4c for confirmation.

7 REPRODUCIBILITY AND ETHICS STATEMENTS

Ethics Statement. Our work proposes a framework to model bounded rationality in Multi-Agent Reinforcement Learning based simulations. Thus our framework may be used to draw implications in simulations modeled after real-world systems. While addressing the rationality gap between human actors and RL actors is an important step towards this goal, it is hardly sufficient for achieving the realism required for real-world decision making based on AI simulations. We anticipate a much longer path forward before this technology matures to such a degree and note that any premature usage of our framework towards real-world decision making would contradict its intended research purpose. Furthermore, our framework should not be used to explore methods to increase discrimination or unfairness in real-world systems, instead it should be use to investigate how to decrease such biases. We acknowledge that our PA settings identify a tension between employer (Principal) profit and employee (Agent) utility/equality. Our results are not meant to provide any actionable insight into how this tension could be manipulated for the benefit of one party. Any such application of our results or our proposed framework would constitute an ethical violation.

Reproducibility Statement. We discuss specific training details for all experiments such as hyperparameters, random seeds used, and our process for setting the random seed in the Appendix. Upon acceptance, all code for this project will be made open source and publicly available for reproducibility purposes and further research. We will share our code with reviewers and ACs for review upon the opening of the discussion forums.

REFERENCES

Sanjeevan Ahilan and Peter Dayan. Feudal multi-agent hierarchies for cooperative reinforcement learning. *Workshop on "Structure & Priors in Reinforcement Learning" at ICLR*, 2019.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *ICLR*, 2020.

Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Modeling human categorization of natural images using deep feature representations. *arXiv preprint arXiv:1711.04855*, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *ICML*, 2018.

Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 2002.

David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pp. 5133–5141. PMLR, 2019.

Aysun Bozanta and Aslıhan Nasır. Usage of agent-based modeling and simulation in marketing. *Journal of Advanced Management Science Vol*, 2(3), 2014.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, pp. 1317–1327. PMLR, 2020.

J-P Caverni, J-M Fabre, and Michel Gonzalez. *Cognitive biases*. Elsevier, 1990.

Mark Dean and Nate Leigh Neligh. Experimental tests of rational inattention. 2017.

Mathias Erlei and Heike Schenk-Mathes. Bounded rationality in principal-agent relationships. *German Economic Review*, 18(4):411–443, 2017.

Benjamin Patrick Evans and Mikhail Prokopenko. Bounded rationality for relaxing best response and mutual consistency: The quantal hierarchy model of decision-making. *arXiv preprint arXiv:2106.15844*, 2021.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *ICLR*, 2019.

Xavier Gabaix. Behavioral Inattention. In *Handbook of Behavioral Economics*. 2017.

Corrado W Gini. Variability and mutability, contribution to the study of statistical distributions and relations. studi cconomico-giuridici della r. universita de cagliari (1912). reviewed in: Light, rj, margolin, bh: An analysis of variance for categorical data. *J. American Statistical Association*, 66:534–544, 1971.

Jordi Grau-Moya, Felix Leibfried, and Peter Vrancx. Soft q-learning with mutual-information regularization. In *ICML*, 2018.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics*, pp. 302–340. Springer, 1992.

Jason Siyanda Hartford. *Deep learning for predicting human strategic behavior*. PhD thesis, University of British Columbia, 2016.

Joseph G Haubrich. Risk aversion, performance pay, and the principal-agent problem. *Journal of Political Economy*, 102(2):258–276, 1994.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997.

William Hoiles, Vikram Krishnamurthy, and Kunal Pattanayak. Rationally inattentive inverse reinforcement learning explains youtube commenting behavior. *J. Mach. Learn. Res.*, 21(170):1–39, 2020.

Bengt Holmstrom and Paul Milgrom. Aggregation and Linearity in the Provision of Intertemporal Incentives. *Econometrica*, 1987.

Gege Jiang, Mogens Fosgerau, and Hong K Lo. Route choice, travel time variability, and rational inattention. *Transportation Research Procedia*, 38:482–502, 2019.

Kris N Kirby and Richard J Herrnstein. Preference reversals due to myopic discounting of delayed reward. *Psychological science*, 6(2):83–89, 1995.

Yoav Kolumbus and Gali Noti. Neural networks for predicting human interactions in repeated games. *arXiv preprint arXiv:1911.03233*, 2019.

Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Complexity of computing optimal stackelberg strategies in security resource allocation games. In *Twenty-fourth aaai conference on artificial intelligence*, 2010.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *arXiv preprint arXiv:1909.06161*, 2019.

Maciej Łatek, RL Axtell, and Bogumil Kaminski. Bounded rationality via recursion. In *Proceedings of Eighth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2009)*, pp. 457–464, 2009.

Felix Leibfried and Daniel A. Braun. Bounded rational decision-making in feedforward neural networks. In *UAI*, 2016.

Felix Leibfried and Jordi Grau-Moya. Mutual-information regularization in markov decision processes and actor-critic learning. In *CoRL*, pp. 360–373. PMLR, 2020.

Felix Leibfried, Jordi Grau-Moya, and Haitham Bou-Ammar. An Information-Theoretic Optimality Principle for Deep Reinforcement Learning. *NeurIPS Workshop on Deep Reinforcement Learning*, 2018.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*, 2017.

Wei Ji Ma and Benjamin Peters. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*, 2020.

Bartosz Maćkowiak and Mirko Wiederholt. Busines Cycle Dynamics Under Rational Inattention. *European Central Bank Working Paper Series*, 2011.

Bartosz Mackowiak, Filip Matejka, and Mirko Wiederholt. Rational inattention: A review. 2021.

Gary J Miller. The political evolution of principal-agent models. *Annu. Rev. Polit. Sci.*, 8:203–225, 2005.

James A Mirrlees. The optimal structure of incentives and authority within an organization. *The Bell Journal of Economics*, pp. 105–131, 1976.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *NeurIPS*, 2015.

Pedro A. Ortega, Daniel A. Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-Theoretic Bounded Rationality. *arXiv preprint arXiv:1512.06789*, 2015.

Zhen Peng, Tim Genewein, Felix Leibfried, and Daniel A Braun. An information-theoretic on-line update principle for perception-action coupling. In *2017 IEEE/RSJ International Conference on intelligent robots and systems (IROS)*, pp. 789–796. IEEE, 2017.

John W Pratt. Risk aversion in the small and in the large. In *Uncertainty in economics*, pp. 59–79. Elsevier, 1978.

Ardeshir Raihanian Mashhadi and Sara Behdad. Environmental impact assessment of the heterogeneity in consumers' usage behavior: An agent-based modeling approach. *Journal of Industrial Ecology*, 22(4):706–719, 2018.

Yuliy Sannikov. A continuous-time version of the principal-agent problem. *The Review of Economic Studies*, 75(3):957–984, 2008.

Zhenyu Shi, Runsheng Yu, Xinrun Wang, Rundong Wang, Youzhi Zhang, Hanjiang Lai, and Bo An. Learning expensive coordination: An event-based deep rl approach. In *ICLR*, 2019.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Tianmin Shu and Yuandong Tian. M3rl: Mind-aware multi-agent management reinforcement learning. *ICLR*, 2019.

Herbert A Simon. Models of man; social and rational. 1957.

Christopher A Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3): 665–690, 2003.

Michael Spence. Job Market Signaling. *The Quarterly Journal of Economics*, 1973.

Klaus Spremann. Agent and Principal. *Agency Theory, Information, and Incentives*, 1987.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Alexander Trott, Sunil Srinivasa, Douwe van der Wal, Sebastien Haneuse, and Stephan Zheng. Building a Foundation for Data-Driven, Interpretable, and Robust Policy Design using the AI Economist. *arXiv preprint arXiv:2108.02904*, 2021.

Amos Tversky and Daniel Kahneman. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

Ying Wen, Yaodong Yang, Rui Luo, and Jun Wang. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. *arXiv preprint arXiv:1901.09216*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

Yaodong Yang, Lantao Yu, Yiwei Bai, Jun Wang, Weinan Zhang, Ying Wen, and Yong Yu. A Study of AI Population Dynamics with Million-agent Reinforcement Learning. In *AAMAS*, 2018.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI Economist: Optimal Economic Policy Design via Two-level Deep Reinforcement Learning. *arXiv preprint arXiv:2108.02755*, 2021.

A Additional Bandit Experiment Details and Results

A.1 Additional Details

Noise Structure. Recall in the bandit experiments, the output z is a noisy function of the hours action h chosen by agent. Specifically we use the following output distribution:

$$p(z|h) = \begin{cases} 0.7 & \text{if } z = h \\ \frac{0.7^{|z-h|}}{0.3 \cdot c} & \text{otherwise} \end{cases} \tag{12}$$

Where c is a numerically calculated factor such that $p(z|h)$ is a proper probability distribution satisfying $\sum_{z \in \mathcal{Z}} p(z|h) = 1$. This noise distribution assures high correlation between h and z with the probability of z levels decreasing exponentially as the difference between h and z grows. We plot the probability heatmap in Figure 6a.

Constant Relative Risk Aversion (CRRA) Function The constant relative risk aversion (`crra` function (Pratt, 1978) commonly used in economics has constant risk aversion which means decisions are invariant to scale. It is a concave function where the risk aversion parameter, $\rho \geq 0$, determines the concavity. This models diminishing returns with higher values. The function is given by:

$$\texttt{CRRA}(u, \rho) = \begin{cases} \frac{u^{(1-\rho)}-1}{1-\rho} & \text{if } \rho \neq 1, \rho \geq 0 \\ ln(u) & \rho = 1 \end{cases} \tag{13}$$

Inattention Model From Theory. The model studied in classical theoretical Principal-Agent literature we compare with (Mirrlees, 1976) utilizes the same utility functions for the principal and the agent. It has a slightly different model of principal inattention where it models costly principal inattention with a factor θ that controls the magnitude of the observation noise in the output signal. Specifically, for a true output level z^*, the principal observes output $z = z^* + \frac{1}{\theta}\epsilon$ where ϵ is noise. The principal incurs an attention cost that is a function of θ^2 and is subtracted from their utility function. While this attention model is different from our mutual information based model, it shares many similar properties. In both our models the inattention introduces noise which makes acting optimally difficult. Additionally in both models, an attention parameter can reduce this noise at a cost to the principal and the principal can choose the optimal, utility-maximizing value of this attention parameter. Such similarities explain why, even under different models, our results are similar to those from theoretical analysis.

Training Hyperparameters. We used learning rates of 1e-3 for the training the principal policy parameters and 5e-3 for the mutual information classifier. We used a batch size of 128 and trained the principal for a total of 100000 batches. During training we gradually annealed λ_{a_p} from 0 to the desired value at a rate of $4/10000$ per batch. We average all results across 5 random seeds and we set the random seed for pytorch, numpy, and python's internal random module for each run. We used seeds of $[0, 4]$. All experiments were run on 16CPU cloud compute machines with 54GB of memory. Given a pay schedule which consists of mean and standard deviation parameters $(\boldsymbol{\mu_z}, \boldsymbol{\sigma_z})$ for each output level $z \in \mathcal{Z}$, to calculate the principal policy we first sample 100 pay schedules and calculate the agent utility per output for each output level for each pay schedule. We then average to

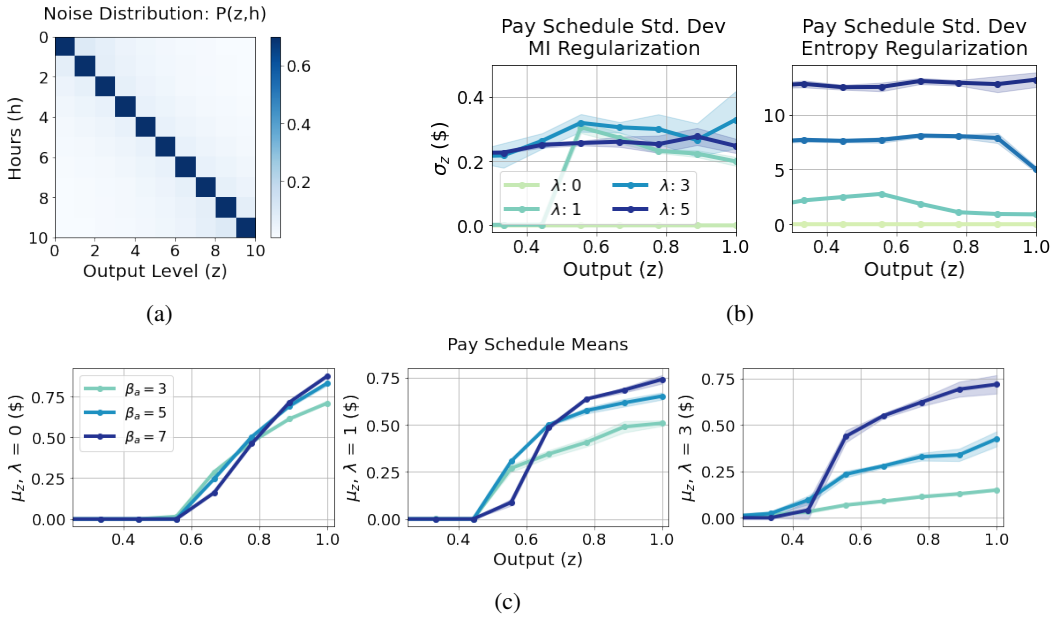

Figure 6: Additional bandit experiment results. All results were averaged across 5 random seeds and have 95% confidence regions shaded. (a) The noise distribution between hours and output ($P(z|h)$). (b) The pay schedule standard deviation plots for MI and Entropy Regularization. (c) A comparison of pay schedule means across different Agent policy β values.

calculate the average agent utility for each output level. We use the noise structure to calculate the average utility for each action and use the soft-q formulation over the utilities given in the main text to obtain the agent's stochastic policy.

A.2 Additional Results

We provide a few additional figures for the bandit experiment. Recall the principal's policy was the pay schedule parameterized by mean and standard deviation values for each output level $(\boldsymbol{\mu_z}, \boldsymbol{\sigma_z})$. Figure 6b shows the standard deviation parameters $\boldsymbol{\sigma_z}$ learned under MI and Entropy regularization for different cost factors λ. Increasing entropy regularization results in very large increases to the standard deviation parameters while MI regularization leads to much smaller standard deviation increases. The pay schedule means for Entropy regularization become sharper to overcome this increase in standard deviation.

Figure 6c compares pay schedule means ($\boldsymbol{\mu_z}$) across different Agent policy temperature (β) values. For low inattention costs, the pay schedules are similar. However the pay schedules for higher β (more deterministic agents) decrease slower with increasing principal attention cost. Given the pay schedule parameters, the agent has an optimal h action. More deterministic agents take this optimal action with higher probability so it is more valuable for the principal to incentivize them to higher optimal actions, which in turn lead to higher outputs. This explains why Agent and Principal utility are higher for higher β values.

B Additional Multi-Agent Multi-Timestep Experiment Details and Results

B.1 Additional Details

Training Hyperparameters. We used learning rates of 1e-4 for the Principal and Agent's policy parameters and 1e-3 for all the mutual information classifiers. We used a batch size of 512 episodes and train the principal and agent through 60000 batches. We train a single RIRL-actor for the

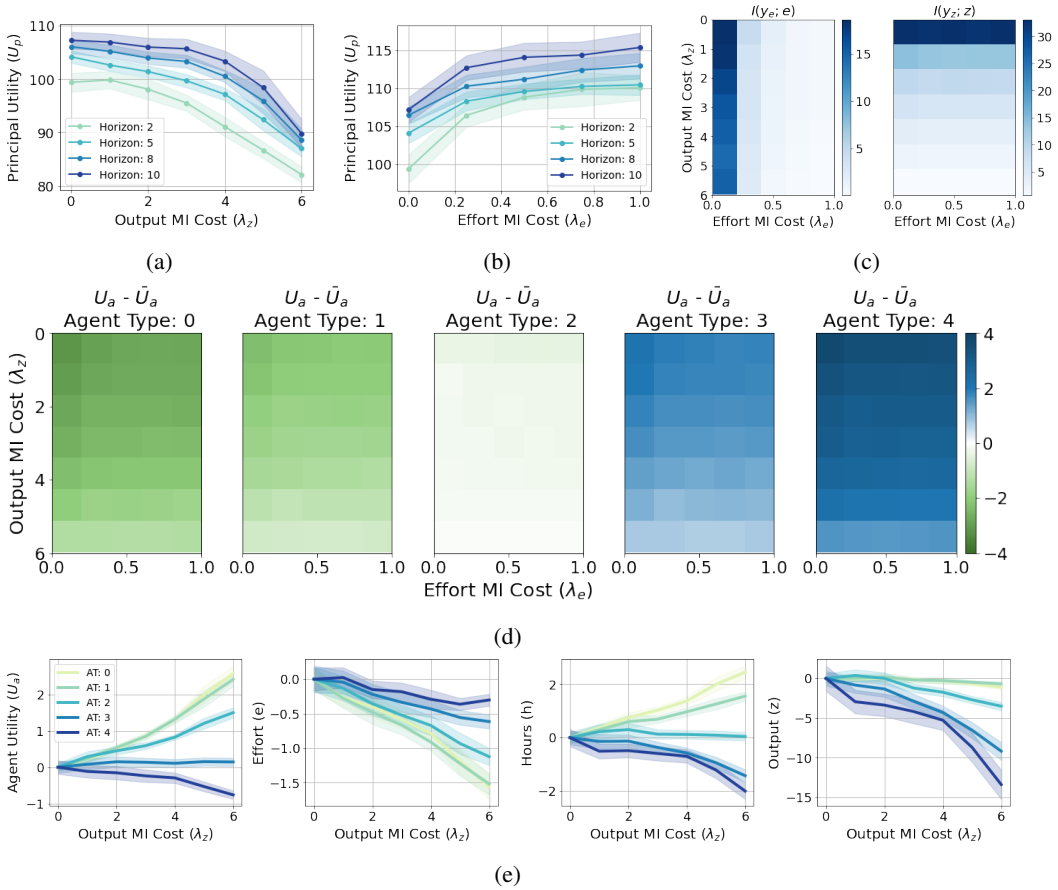

Figure 7: Additional sequential, Multi-Agent experiment results. All results are averaged over 20 random seeds and all confidence regions depict 95% confidence intervals. (a, b) Principal Utility with Output (λ^z) and Effort (λ^e) MI Costs. (c) The unscaled Output and Effort MI with λ^z and λ^e. (d) The different between each agent type's utility and the mean utility across all types with λ^z and λ^e. (e) How agent utility, effort, hours, and output changes with respect to Output MI Cost λ^z for each agent type. All values are plotted as the change amount from the $\lambda^z = 0$ value.

Agents and concatenate the experiences of all n_a Agents when updating the policy. The Agent policy therefore effectively has a batch size of $512n_a$. To avoid vastly different total episode returns, we scaled the rewards by the horizon during training. We run all experiments on 8CPU cloud computing machines with 26GB of memory. We average all results across 20 random seeds and we set the random seed for pytorch, numpy, and python's internal random module for each run. We used seeds of $[0, 19]$.

B.2 Additional Results

We present some additional results for the sequential multi-agent experiments. Figures 7a,7b show Principal utility with output MI cost (λ^z) and effort MI cost (λ^e) for different horizons. We see the trend of Principal utility decreasing with λ^z and increasing with λ^e occurs across all horizons.

Figure 7c shows unscaled MI cost for output and effort with λ^z and λ^e. We see that $I(y_e|e)$ decreases with both λ^z and λ^e. $I(y_z|z)$ decreases with λ^z but is not affected greatly by λ^e.

Figure 7d plots the agent utility for all agent types. It expands the plot of Figure 4d. The plots show the difference between the utility for each agent type and the average utility.

Figure 7e demonstrates how different values for different agent types change with output MI cost (λ^z). We plot each result as the change from the $\lambda^z = 0$ values. As mentioned previously, utility

increases with λ^z for lower ability agents while it decreases for the highest ability agent. We observe similar trends for hours h. We additionally notice effort e generally decreases which leads to z staying constant or decreasing for all agent types, even though some agent types have increased hours. This lower output along with higher average wages explains the decrease in Principal utility.

C RIRL IMPLEMENTATION DETAILS

We use this section of the appendix to cover important implementation details and tips. This is intended for practical guidance.

Stochastic Encoder Module Configuration and Initialization. Section 4 describes the architecture of our RIRL-actor policy class. As described, we learn an encoder $f^m(y_t^m|o_t^m, \psi_t)$ for each observation channel m. Encoder f^m takes observation o_t^m and recurrent state ψ_t as inputs and outputs the parameters (means and standard deviations) of a stochastic encoding y_t^m.

We find that, in practice, learning does not progress if each y^m contains very little information about o^m at the start of training. To address this, we recommend two implementation choices. First, implement f^m as a residual-style module:

$$\mu_t^m, \sigma_t^m = f^m(o_t^m, \psi_t), \quad y_t^m = o_t^m + \mu_t^m + \sigma_t^m \cdot \epsilon_t^m, \tag{14}$$

This simply requires setting the output y^m to have the same size as observation o^m and adding o_t^m to the mean μ_t^m. Second, initialize the output layer of f^m such that σ^m is consistently very small. We perform this by adding a constant negative offset to the bias units associated with the $\log \sigma^m$ outputs, which we exponentiate to get σ^m. As a result of this strategy, y_t^m closely follows o_t^m at the start of training.

Hidden State as an Encoder Input. We emphasize that the inputs to encoder f^m is the concatenation of the observation o_t^m and the hidden state ψ_t. Similarly, when using the discriminator $d^m(y_t^m, [o_t^m, \psi_t])$ to estimate $\tilde{I}_{f^m}$ and when training the discriminator, we also apply this concatenation. In other words, d^m regards $[o_t^m, \psi_t]$ as the observation, such that the $\tilde{I}_{f^m}$ captures the MI between y_t^m and $[o_t^m, \psi_t]$.

This is an important detail for ensuring that MI regularization works as expected in the multi-step setting. For instance, we observed that, if the discriminator does not see the hidden state ψ, f^m learns to encode o^m such that $I(y^m; o^m)$ is minimal but where o_t^m can still be easily recovered given y_t^m *and* ψ_t.

Optimization During training, we found that learning was most stable if separate learning rates were used for the policy modules $\{f^1, \ldots, f^M, \texttt{LSTM}, \omega\}$ and for the discriminator modules $\{d^1, \ldots, d^M, d^\omega\}$. Importantly, *the discriminator modules use a 10x higher learning rate.* Concretely, we use learning rates of 0.0001 and 0.001 for the policy and discriminator modules, respectively. Configuring learning rates this way helps to ensure that discriminator d^m can adjust to changes in encoder f^m faster than the encoder can adapt to changes in the discriminator. Intuitively, this improves the quality of the MI estimates during training.

Another important optimization detail concerns gradient flow. Gradients from $\nabla \log \omega(a_t|\cdot)$ need to backpropagate through the encodings $\left[y_t^1, \ldots y_t^M\right]$ in order for the encoder modules to receive meaningful gradients. In Pytorch, which we use for this implementation, ensuring that this gradient flow occurs requires some attention. Given the (learned) mean and standard deviation parameters (which are functions of the encoder input), Pytorch constructs the sampling distribution as `m = Normal(`μ, σ`)`. The *output* of the encoder module y_t^m must be sampled via the reparameterization trick: `y = m.rsample()`, which allows gradients to flow through y_t^m. Finally, this output should be detached from the backpropagation graph when calculating its *log probability*: `encoder_log_prob = m.log_prob( y.detach() )`, which is needed for computing the policy gradients. Similarly, care must be taken to ensure that inputs to the discriminators are detached in the same way.