Emergent rewards in open-ended systems

Richard M. Bailey

Oxford University Centre for the Environment, University of Oxford

richard.bailey@ouce.ox.ac.uk

Abstract

Unambiguous identification of the rewards driving behaviours of entities operating in complex open-ended real-world environments is typically not possible. Nonetheless, goals and associated behaviours do emerge and are dynamically updated. Reproducing such dynamics in models would be highly desirable in many domains. Simulation experiments described here assess a candidate mechanism for dynamic reward updating through learning and inheritance, and successfully demonstrate the abandonment of an initially rewarded but ultimately detrimental behaviour.

Introduction and problem statement

If we reject the idea of mind as a separate entity, the 'ghost in the machine', and view the 'ghost' as an emergent property of physical processes, we must look to computational explanations to explain the purpose, intentions and behaviours of agents. In machine-learning, this reduces to the reward function definition, but leaves open the question of how reward functions emerge. In well-defined explicit tasks (e.g. 'put object A in to box B'), human-designed shaped reward functions yield success. However, 'real-world' agents encountering open-ended problems in potentially open-ended environments (e.g. plants/animals, institutions) face a more implicit goal, to survive by undertaking a range of tasks (implicit because the agent's reward function may be unknown to it, and does not necessarily contain explicit terms related to survival). As environments change, so do the challenges faced by such agents in solving their survival problem. For living entities, (longer-term) evolution provides some aspects of adaptation, and (shorter-term) learning provides others (potentially over multiple lifetimes through 'cultural inheritance'). It is this interplay, and possible co-evolution, of the entity and its behaviour (through changes in its reward structure), under changing boundary conditions, that is the focus of this paper: whether evolution of the basic properties of agents, plus an evolution of rewards (associated with reinforcement learning), can produce simulated populations which dynamically adjust to changing environments and solve their survival problem with no external (modeller) input beyond their initial conditions. Learning is formulated as follows: the true state of the world is only partially observable by each agent and a Partially-Observable Markov Decision Process is described by the tuple $\langle S, A, O, D, R, \gamma \rangle$, where $S = \{s_1, s_2, \ldots, s_n\}$ is a set of partially observable states, $A = \{a_1, a_2, \ldots, a_m\}$ is a set of actions, and $O = \{o_1, o_2, \ldots, o_k\}$ is a set of observations. $\boldsymbol{D} : S \times A \times S \rightarrow [0, 1]$ models the unknown system dynamics as a transition distribution. The immediate reward, $R(\theta, s, a) : S \times A \rightarrow \mathbb{R}$, where $\boldsymbol{\theta} \in \Theta$ are the reward coefficients, is then discounted at a rate $\gamma \in [0, 1]$. Reinforcement Learning (RL) is an attempt to find a policy π_R^* that maximizes the expected cumulative discounted reward R of trajectories $\mathcal{T}$ drawn from a set of initial conditions $N \subset O$, such that $\pi_R^* = \arg\max_\pi \mathbb{E}_{\mathcal{T}(N,\boldsymbol{D},\boldsymbol{\pi})} \left[\sum_{i=0}^{|\mathcal{T}|} \gamma^i R\left(\boldsymbol{\theta}, \boldsymbol{s}_i, \boldsymbol{a}_i\right) \right]$. The behaviour of the agent is therefore dependent on the reward function R, which is typically shaped through human selection of $\boldsymbol{\theta}$ to result in the desired outcomes. The success of any policy can be quantified as $\mathcal{J}(\pi) = \mathbb{E}_{\mathcal{T}(N,\boldsymbol{D},\boldsymbol{\pi})} [\mathcal{G}(\boldsymbol{\mathcal{T}})]$, where $\mathcal{G}$ is the metric for trajectory success. In unambiguous RL tasks, the relationship between R and candidate (human-created) evaluation functions $\mathcal{G}$ can be straightforward, and the chosen $\mathcal{G}$ can even be used as a fitness metric against which to evolve successful reward functions (Faust et al. (2019)). Defining success for an agent (or population of agents) in real-world open-ended systems is not so straightforward, as multiple related tasks are underway simultaneously and both the properties of agents and the challenges they face may change over time. Further, the agent may face both behavioural and evolutionary trade-offs which are not explicitly known, and maximizing lower-level (LL) rewards may have impacts on achieving higher-level (HL) success. For example, maximizing LL rewards for consumption might produce longer-lived entities and therefore yield larger more successful populations (HL success) even though population size/persistence is not part of the reward definition. Such HL success could in principle be measured by a suitable metric, but such functions ($\tilde{G}$) are typically practically unknowable in real-world open-ended systems. The present work is not therefore an attempt to find an optimally-shaped reward function (i.e. $\arg\max_\theta \mathcal{J}(\hat{\pi})$)

that maximizes an explicit success metric. Instead, the goal is to produce an emergent process which endogenously updates $\boldsymbol{\theta}$ in order to maximise the unobservable success metric ($\tilde{G}$). Ultimately, the process must explore and adjust both the reward coefficients ($\boldsymbol{\theta}$) and the basic properties of the agents ($\boldsymbol{P}$) to prevailing conditions. Given that HL objectives emerge dynamically from contingent (survival) challenges faced by populations of entities, it is not guaranteed that endogenous processes (through selection pressures) acting on $\boldsymbol{\theta}$ will result in convergence to optimal solutions, $\tilde{\pi} = \arg\max_{\boldsymbol{\theta}} \tilde{\mathcal{J}}\left(\tilde{\pi}\right)$, where $\tilde{\mathcal{J}}\left(\tilde{\pi}\right) = \mathbb{E}_{\mathcal{T}(N,\boldsymbol{D},\tilde{\pi})}\left[\tilde{\mathcal{G}}\left(\boldsymbol{\mathcal{T}}\right)\right]$. Indeed, in real-world complex adaptive systems there is ample evidence of entities occupying local physiological and behavioural minima due to inescapable inheritance or slow behaviour updating (e.g. Wedel (2011)).

Simulated environment

Entities ('Ents') inhabit the surface of a simulated bounded 3D physical environment, which they interact with through various actions (each of which entails an energy cost): moving, picking-up/dropping objects, exchanging currencies and signals, eating, processing raw materials and sexually reproducing (once mature, with exchange of 'genes'). They possess sensors and make observations equivalent to sight, hearing, smell and touch. Each Ent has a finite lifetime and their approximately spherical bodies grow as they consume food (they are also capable of storing finite amounts of excess energy from consumption). Incident light from an external source provides energy for the system, which is first converted by primary producers (PPs) in to biomass (plant mass and fruits). Ents can be damaged through impact, predation or mass loss due to starvation and can therefore die before their maximum potential lifetime is reached. In the present experiments, a single (non-evolving) Ent type and single (non-evolving) PP type are used. Ent behaviour (choice of action) is learnt using reinforcement learning (Proximal Policy Optimisation, Schulman et al. (2017)), and reward coefficients (ϕ) are updated endogenously in response to individual experience. The reward obtained in timestep t is calculated as $R_t = \sum_{i=1}^{N_C} \Delta c_i \theta_{\text{currency}\,i} + f\theta_{\text{consumption}} + b\theta_{\text{reproduction}} + p\theta_{\text{pain}}$, where the first term accounts for net changes in the ($N_c \in [0, 4]$) currencies (currency c_4 are 'coins' which are dropped in to the environment periodically, and can be collected, dropped or given to other Ents; uncollected coins vanish after a set duration); f depends on the mass and palatability of consumed food; $b > 0$ upon successful mating and depends on genetic distance of mate; $p > 0$ when the Ent experiences predation, impact damage or death. Currencies bestow no direct advantage (e.g. increased energy), and incur a transaction cost, but can be valued nonetheless.

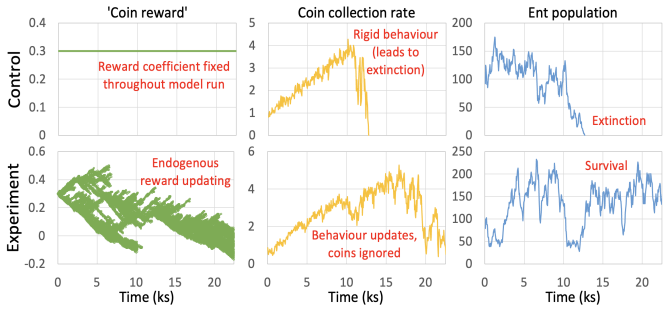

Figure 1: **Control:** Reward coefficient fixed as coin deposition rate linearly increased, causing population extinction due to opportunity costs associated with coin collection. **Experiment:** Reward coefficient allowed to endogenously updateGrey points indicate reward values at birth for individual Ents. Under detrimental increases in coin availability, coin rewards endogenously reduce over time. By comparison, reproduction rewards remain high.

Experiment and results overview

The purpose of the experiment was to demonstrate the possibility of endogenous updating of rewards in response to changing environmental pressures. The world was set up with constant benign conditions and initial reward coefficients ($\boldsymbol{\theta}$) were hand-chosen to yield learnt behaviour conducive to a persistent (multi-generational) population. Given the lack of direct benefit, collecting coins is an intentionally frivolous activity, providing no immediate benefit and entailing a time/energy (therefore opportunity) cost. During the experiment, the availability of coins was progressively increased. In the control condition (absence of reward adaptation), eventually so much time/effort is spent on coin collection that the population collapses (through starvation and a lack of reproduction). However, with reward adaptation (based on the experience and expectations of individual Ents), rewards associated with coins endogenously reduce (Fig.1), meaning Ents increasingly ignore coins, and the population is ultimately able to survive conditions that would otherwise cause its collapse.

Conclusions

The results show it is possible for populations of Ents to adapt their own rewards in response to their experience, in a way which is beneficial to their survival; this emerges without external direction. In the experiment, a rewarding but harmful behaviour was 'switched off', allowing survival under conditions which would otherwise have collapsed the population. Incorporation of multiple evolving Ent and PP types leads to complex behaviours and interactions, which will be reported.

References

Faust, A., Francis, A., and Mehta, D. (2019). Evolving rewards to automate reinforcement learning. *arXiv preprint arXiv:1905.07628*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Wedel, M. J. (2011). A monument of inefficiency: The presumed course of the recurrent laryngeal nerve in sauropod dinosaurs. *Acta Palaeontologica Polonica*, 57(2):251–256.