Modeling intrinsic motivation as reflective planning

Yang Xiang (yyx@g.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA 02138 USA

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology and Center for Brain Science, Harvard University
Cambridge, MA 02138 USA

Abstract

Why do people seek to improve themselves? One explanation is that improvement is intrinsically rewarding. This can be formalized in reinforcement learning models by augmenting the reward function with intrinsic rewards (e.g., internally-generated improvement signals). In this paper, we develop an alternative explanation: the drive for improvement arises from planning in a state space that includes internal states (e.g., competence). Planning is therefore reflective in the sense that it considers the value of future internal states (e.g., "What could I accomplish in the future if I improve my competence?"). We formalize this idea as a sequential decision problem which we dub the *reflective Markov Decision Process*. The model captures qualitative patterns of skill development better than a range of alternative models that lack some of its components. Importantly, it explains these patterns without appealing to intrinsic rewards.

Keywords: Intrinsic motivation; Planning; Reinforcement learning; Markov decision process; Competence; Skill development

Introduction

Reinforcement learning (RL) models formalize how agents take actions to maximize cumulative state-dependent rewards (Sutton & Barto, 2018). When modeling, for example, an agent navigating to goal locations in a grid-world, it's natural to conceptualize grid locations as the states and goal attainment as the reward. But how much of human life is like the grid-world? Yes, we navigate two-dimensional spaces to pursue explicit goals, but this characterization neglects deep questions about why we pursue these (and more abstract) goals in the first place. What exactly is rewarding about increasing competence (White, 1959) or pursuing growth and discovery (Vallerand et al., 1986)? It seems impossible to understand these phenomena without reference to a richer set of *internal* states that coin reward and drive intrinsic motivation (Fishbach & Woolley, 2022; Karayanni & Nelken, 2022).

Efforts to formalize intrinsic motivation have focused primarily on augmenting extrinsic rewards (e.g., food, money) with intrinsic rewards, such as goal achievement (Molinaro & Collins, 2023), curiosity (Burda et al., 2018; Schmidhuber, 1991; Still & Precup, 2012; Eysenbach et al., 2018), and skill mastery (Baranes & Oudeyer, 2013; Bougie & Ichise, 2020). These intrinsic rewards might have evolved to facilitate rapid learning in environments with sparse extrinsic rewards (Singh et al., 2010). While these approaches can explain learning in the absence of external reward, they still ground the state space purely in the external environment. In contrast, humans have a mental life that extends beyond the representation of external states. In this paper, we argue that many aspects of intrinsic motivation can be modeled by augmenting the state space with internal states, even in the absence of intrinsic rewards. The key idea is that intrinsic motivation arises from planning over these internal states—a form of self-reflection which we dub *reflective planning*.

We formalize this idea as a *Reflective Markov Decision Process* (rMDP). This framework represents aspects of the agent (e.g., competence) as part of the state space. It endogenizes intrinsic motivation through value functions defined over internal states. In the following sections, we formally define the rMDP, and then explore its application to a concrete example of skill development. We test its ability to explain several classical phenomena, comparing it to several lesioned versions (lacking particular aspects of the full model) and a model with intrinsic rewards inspired by Molinaro and Collins (2023). Finally, we discuss the implications of the model and outline experiments to test its validity.

Model

A Markov decision process (MDP) is a tuple $(\mathcal{S},\mathcal{A},T,R,\gamma)$ with the following components:

- A set of states $\mathcal{S}$.
- A set of actions $\mathcal{A}$, which can be state-dependent.
- A transition function $T : \mathcal{S}\times\mathcal{A}\to\mathcal{S}$, which we write as $s' = T(s,a)$. More generally, the transition function can be probabilistic, but for simplicity here we restrict our attention to deterministic dynamics.
- A reward function $R : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, which we write as $r = R(s,a)$. The reward function can also be probabilistic but here we restrict our attention to deterministic rewards.
- A discount factor $\gamma\in[0,1)$ that weighs short-term rewards more highly than long-term rewards.

The rMDP is a form of MDP that represents aspects of the agent as part of the state space (see Figure 1). An agent that models the rMDP thus models itself (i.e., reflects upon itself). In particular, the rMDP decomposes the state and action into several components. **State**, $s = (x, c)$, includes *environmental*

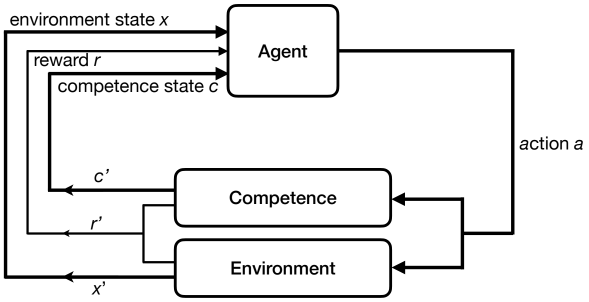

Figure 1: Illustration of the rMDP model.

state $x \in \mathcal{X}$ and the agent's *competence* $c \in \mathbb{R}^N$ which defines the agent's ability to engage in different activities when exerting maximum effort (N denotes the number of possible activities). **Action**, $a = (u, e)$, includes the agent's *activity choice* $u \in \mathcal{U}(x) \subseteq \{1, \ldots, N\}$, where the menu of possible activities depends on the current environment state, and the agent's *effort level* $e \in [0, 1]$, defined as a proportion of the agent's competence.

With these components, we can now define the transition dynamics and reward structure. The dynamics are specified by a factorized transition structure: $x' = T_X(x, a)$ and $c' = T_C(c, a)$. We assume the following properties:

- Competence is increasing in effort: $\partial c'/\partial e > 0$. You get better at activities when you exert effort on them.
- Competence is increasing in prior competence: $\partial c'/\partial c > 0$. You don't lose competence spontaneously.
- Reward is decreasing in effort: $\partial r/\partial e < 0$. Effort is aversive.

Following the convention in RL models, we assume that the agent seeks to maximize expected discounted cumulative future reward, or *value*, which can be expressed recursively using the Bellman equation:

$$Q(s, a) = R(s, a) + \gamma \max_{a'} Q(s', a'). \tag{1}$$

The value iteration algorithm uses the Bellman equation to compute $Q(s, a)$ for each action, and then sets the agent's policy to $\pi(s) = \text{argmax}_a Q(s, a)$. It iterates over these updates until convergence.

Example: Intuitive skill development

We illustrate the framework in an example of intuitive skill development. Consider the setting where the environment state is defined by a fixed set of tasks, $\mathcal{U}(x) = \{1, \ldots, N\}$, where x is always fixed (i.e., a "bandit" setting where the available tasks do not change across states). At each state, an agent can choose an effort level $e \in [0, 1]$ and a task u with difficulty d.

The reward is the utility of choosing an action (expected external reward minus cost). We make the assumption that the external reward of a task is proportional to its difficulty. The reward function is defined as:

$$r = dP(s = 1|c, e, d) - \alpha e, \tag{2}$$

where $\alpha > 0$ determines the cost of effort. $P(s = 1|c, e, d)$ is the probability of succeeding given the agent's competence, effort, and the task difficulty, which follows a logistic function:

$$P(s = 1|c, e, d) = \frac{1}{1 + e^{-k(ce-d)}}, \tag{3}$$

where k controls the steepness of the curve. Intuitively, this means that the probability of success is higher when the agent's output (competence multiplied by effort) is larger; when the output equals the difficulty of the task, the likelihood of succeeding is at chance (50%).

Past work has used a deterministic success function, where an agent succeeds if and only if they exert more force (ce) than the task difficulty (d) (Xiang, Vélez, & Gershman, 2024, 2023; Xiang, Landy, et al., 2023; Xiang et al., 2025) (but see Xiang, Gershman, & Gerstenberg, 2024). By contrast, Equation 3 offers more flexibility—it can be used to describe a variety of tasks, where a larger k captures more deterministic outcomes and a smaller k captures captures more variable outcomes (e.g., when luck plays a bigger factor).

Competence evolves according to:

$$c' = c + \beta e \exp\left[-\frac{(c-d)^2}{2\sigma^2}\right], \tag{4}$$

where $\beta > 0$ is a coefficient determining the effects of effort. σ is a domain-specific parameter that determines how much a task's difficulty can deviate from current competence while still contributing to improvement. Larger σ allows for growth from a wider range of tasks, while smaller σ means only tasks near current competence are effective for learning. For example, running has a large σ (even short or slow runs contribute to endurance and fitness over time), whereas math has a small σ (solving very easy or very difficult problems doesn't typically improve skills).

Intuitively, more effort exertion leads to greater increase in competence. In addition, competence increases the most when the agent attempts tasks that are just at their current abilities (which demands near-maximal effort to succeed). These two features of Equation 4 align with the principle of deliberate practice, which was found to play an important role in the acquisition of expert performance (Ericsson, 2008; Ericsson & Pool, 2016; Ericsson et al., 1993).

Equation 4 extends the function postulated by past work (Xiang, Vélez, & Gershman, 2024), according to which competence increases based on the amount of effort exerted, and only when the agent succeeds. Our new formulation of competence dynamics captures the fact that competence can increase even when an agent attempts a task that is slightly beyond their competence.

Alternative models

To examine the necessity of including all the components of the rMDP, we consider two alternative models that each lesion one component. Additionally, we consider an intrinsic reward model inspired by past work (Molinaro & Collins, 2023).

Myopic model: $\gamma = 0$

The first alternative model lesions prospection. This means that the agent does not consider future rewards in their planning; they will always choose the action that yields the greatest immediate reward.

Fixed mindset model: $\beta = 0$

The second alternative model lesions expectations about competence change. This means that the agent does not factor competence increase into their planning, expecting to permanently stay in the same competence state. Past work on growth versus fixed mindsets lends plausibility to this model (Dweck & Leggett, 1988; Blackwell et al., 2007; Haimovitz & Dweck, 2017). The fixed mindset model might appear to be similar to the first alternative model; it also chooses the most immediately rewarding action, because (in the agent's mind) the current best and future best are the same. However, there is one key difference: Competence in the first alternative model gets updated after each action, and therefore the agent may take different actions at different time points. By contrast, competence in the second alternative model is expected to remain the same over time, and therefore the agent always takes the same action.

Intrinsic reward model

The intrinsic reward model includes an extra component in the reward function: an intrinsic reward $\tilde{r} = c' - c$. It postulates that the agent has a motivation to improve their competence independently of the external rewards. The reward function is therefore:

$$r = w\tilde{r} + (1-w)(dP(s=1|c,e,d) - \alpha e), \tag{5}$$

where $w \in [0,1]$ is a mixture parameter that controls the weighting of the intrinsic reward and expected utility. When $w = 1$, the agent only values competence increase, whereas when $w = 0$, the reward function is the same as for the other models. For our simulations, we set $w = 0.5$. Note that in the implementation, we additionally multiply the intrinsic reward $\tilde{r}$ by the maximum possible reward ($\max(r)$. This brings the intrinsic reward and extrinsic reward into a commensurable range. Without it, the intrinsic reward model makes almost the same predictions as the rMDP unless $w = 1$.

Implementation details

We simulated each model in a "planning phase" with a starting competence of $c_0 = 10$, over 50 time steps. At each time step, the agent chooses a task and an effort level, based on its beliefs about how competence changes. During the "execution phase" (when the agent carries out its planned actions), we computed the competence trajectory according to the chosen task difficulty and effort level at the corresponding time steps. During this phase, the agent's ground truth competence always changed according to Equation 4 for all models.

The parameter values used in the simulations below are summarized in Table 2. Note that the specific values don't affect the overall patterns as long as the parameters are within reasonable ranges; for example, if effort cost (α) is too high, the agent won't exert any effort. If the effect of effort on learning (β) is too small, the intrinsic reward term ($\tilde{r}$) in the intrinsic reward model will be negligible.

Table 2: Model parameter values.

Parameter	Meaning	Value
$\mathcal{C}$	Competence range	[1, 100]
$\mathcal{E}$	Effort range	[0, 1]
$\mathcal{D}$	Task difficulty range	[0.5, 100.5]
k	Slope of the logistic function	2
α	Effort cost	0.2
β	Coefficient determining effects of effort	2
σ	Domain-specific variability	1
γ	Discount factor	0.9
w	Weight of intrinsic reward	0.5

Simulations

We simulated four phenomena in skill development based on the empirical evidence. These phenomena demonstrate the effects of certain constraints on learning. A summary of the phenomena, empirical evidence, description of constraint, and implementation details are summarized in Table 1. The code and simulated data are publicly available at `https://osf.io/6cyhv/`.

The results are visualized in Figure 2. Each column corresponds to one model. Each row corresponds to one phenomenon, where the coral curve shows the condition with constraints imposed by each phenomenon (as summarized in Table 1), and the turquoise curve shows the unconstrained condition (which is the same across phenomena). Overall, without constraints, the rMDP and the intrinsic reward models predicted that the agent would learn quickly and continue to improve. The intrinsic reward model predicted sightly more competence increase than the rMDP model due to the additional incentive the agent had for improving their competence ($\tilde{r}$ in Equation 5). The myopic model predicted less overall improvement compared to the rMDP model. The fixed mindset model did not predict much competence increase.

Below, we describe the four phenomena and simulation results in detail.

Phenomenon 1: Appropriate difficulty led to more mastery than overly difficulty tasks ("ZPD")

In two studies, Zou et al. (2019) and R. Baker et al. (2020) collected and analyzed data from an online learning platform, Learnta. Both studies found that students gained more mas-

Table 1: Constraints in each phenomenon.

Phenomenon label	Empirical evidence	Constraint	Implementation
ZPD	Zou et al. (2019); R. Baker et al. (2020)	Task always too difficult	d forced to c_0 + 1
Clustering	Pierce et al. (2011)	Task always too simple	d forced to c_0 - 1
Pygmalion	Rubie-Davies (2016)	Task starts right but doesn't change	d forced to c_0
Adaptation (Low)	Corbalan et al. (2008)	Yoked to less competent agent	d forced to agent starting with $c_0 - 3$
Adaptation (High)	Corbalan et al. (2008)	Yoked to more competent agent	d forced to agent starting with $c_0 + 3$

tery when they took on tasks that they were "ready to learn"—i.e., within their Zone of Proximal Development (Vygotsky, 1978)—compared to tasks that they were "unready to learn". These findings held across math and English learning domains and various success levels (excellent, normal, or, struggling).

We model the constraint under "unready to learn" scenarios as the task difficulty being fixed to $c_0 + 1$, where c_0 denotes the agent's starting competence. In other words, the agent is forced to take on a task that is more difficult than their starting competence. The agent could still choose the amount of effort to exert.

The rMDP model captured the qualitative pattern of the constrained curve growing more slowly than the unconstrained curve. The constrained curve showed a little bit of learning at the very beginning, but then stopped improving because the task was too hard. The intrinsic reward model showed similar patterns. The myopic model was able to capture the relative differences between constrained and unconstrained conditions for most of the time steps, but it predicted that the agent improves quicker in the constrained condition at first, and the relative differences overall were smaller compared to the rMDP and intrinsic reward models. The fixed mindset model predicted the opposite pattern—the constrained curve was slightly above the unconstrained curve.

Phenomenon 2: Gifted learners learn faster in cluster settings ("Clustering")

Pierce et al. (2011) studied the impact of using cluster grouping to support gifted learners' math achievement in urban elementary schools. They found that clustering enabled greater improvement in both gifted and comparison learners. Gifted learners enjoyed a differential rate of learning success when placed in cluster settings, as compared to gifted learners in noncluster classrooms.

We model the constraint that gifted students have to deal with in nonscluster classroom as fixed task difficulty of $c_0 - 1$. In other words, the agent has no choice but to take on a task that is too easy for them. The agent could still choose the amount of effort to exert.

Both the rMDP and the intrinsic reward models captured this qualitative pattern. The myopic model also did, but the difference between conditions was small. The fixed mindset model predicted the same pattern between constrained and unconstrained conditions.

Phenomenon 3: Students taught by high-expectation teachers learn faster ("Pygmalion")

Through a three-year Teacher Expectation Project (Rubie-Davies, 2014), Rubie-Davies (2016) found that students in the classes of high-expectation teachers (intervention group) gained more marks on their standardized tests (equivalent to almost three additional months of learning) by the end of the year, compared to students in the classes of control teachers. The high-expectation teachers set individual learning goals with their students that were continuously challenging for all students. They monitored student progress regularly and moved students to higher levels individually as they were ready.

We model the constraint that students face with control teachers as task difficulty fixed at c_0. This means that the student starts with the task difficulty just right for them $d = c_0$, but it doesn't change as learning goes on.

Both the rMDP and the intrinsic reward models captured this qualitative pattern. The myopic model predicted faster progress at start under constraints, and the difference between conditions was small. The fixed mindset model predicted almost the same pattern between constrained and unconstrained conditions.

Phenomenon 4: Adaptation leads to better learning outcomes compared to yoked ("Adaptation (Low)" and "Adaptation (High)")

Corbalan et al. (2008) studied students' learning in a Web application. They found that learners who took on tasks that were matched to their competence learned more effectively and more efficiently, compared to their yoked counterparts, who received exactly the same sequence of tasks. The finding shows that adaptation (whether chosen by the computer or the learner) leads to better learning outcomes than yoked controls.

We modeled the constraints that yoked learners had by simulating a counterpart whose competence is either lower (starting with $c_0 - 3$) or higher (starting with $c_0 + 3$) than the agent. The agent is yoked to their counterpart in terms of difficulty. However, the agent still has freedom to choose their effort allocation and we assume that they plan their effort "as if" they have control.

When the agent was yoked to a less competent counterpart ("Adaptation (Low)"), the rMDP, intrinsic reward, and myopic models predicted that the agent's learning would be lagged compared to the unconstrained condition, but they

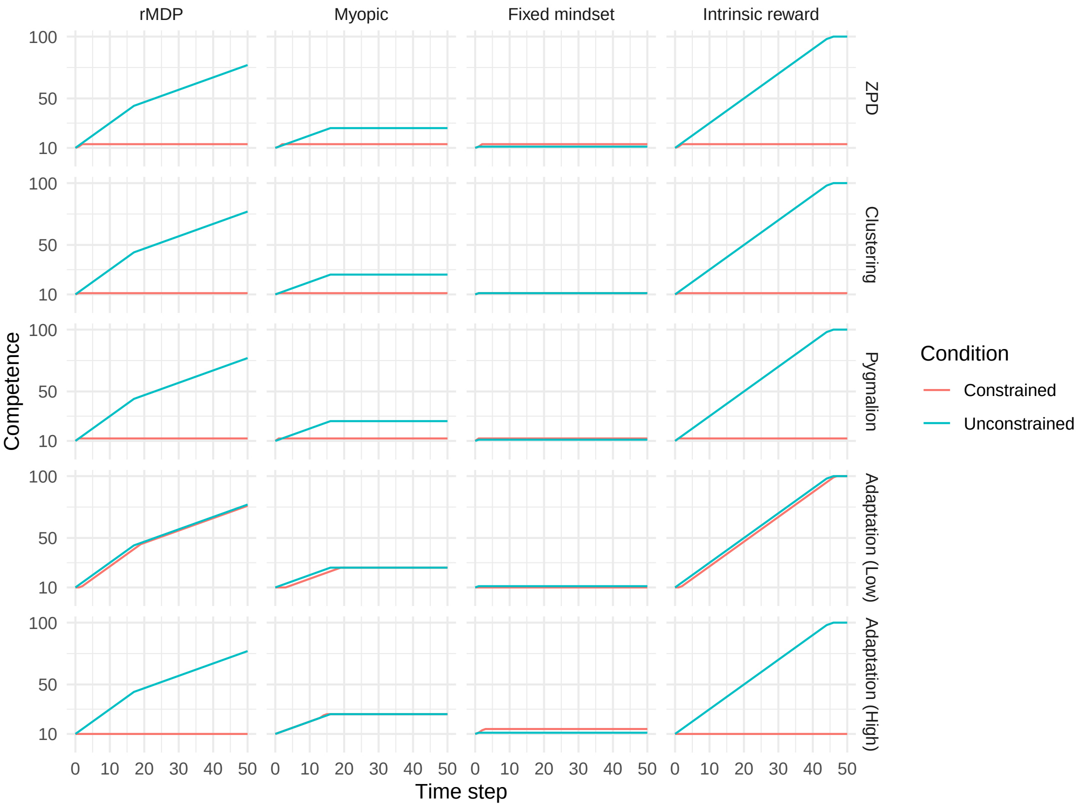

Figure 2: Simulation results. Constrained curves show the competence trajectories under constraints imposed by the specific phenomenon. Overall, the rMDP and intrinsic reward models were able to capture the qualitative patterns of each phenomenon. The myopic and fixed mindset models were unable to do so.

would still benefit from their counterpart once their counterpart reached their competence level. The fixed mindset model predicted almost the same pattern between constrained and unconstrained conditions.

When the agent was yoked to a more competent counterpart ("Adaptation (High)"), the rMDP and intrinsic reward models predicted that the tasks appropriate for their counterpart would be too difficult for the agent, and with time the tasks get even harder. The myopic model predicted the same pattern between constrained and unconstrained conditions, with the constrained condition plateauing a bit faster than the unconstrained condition. Finally, the fixed mindset model predicted that the agent would improve faster when they were yoked, which is the opposite of what was empirically observed.

Teasing apart predictions of the rMDP and intrinsic reward models

In the results above, we found that the intrinsic reward model showed similar patterns to the rMDP model. In this section, we simulate two studies that disentangle them.

Phenomenon 5: Demotivating effects of getting awards

In a field experiment, Robinson et al. (2021) found that students who received awards for perfect attendance attended less school in the following month. Importantly, students were told that these were one-time awards. We modeled this using extrinsic rewards that only existed for tasks that were harder than a threshold, so that getting the reward required improving competence. These rewards also remained the same for all tasks beyond the difficulty threshold. As shown in Figure 3 (bottom panel), for both easy and hard goals where reward is only available for tasks with $d \geq 15$ or 50, the rMDP model predicted that the agent would improve until they are capable of getting the reward, but they are not motivated to improve their competence further. This prediction aligned with the empirical finding. By contrast, the intrinsic reward model predicted that the agent would continue to increase their competence regardless of the threshold.

Phenomenon 6: Demotivating effects of negative feedback and motivating effects of goal alignment reminders

Anand et al. (2023) found that frequent unfavorable feedback about goal progress demotivated participants. In addition,

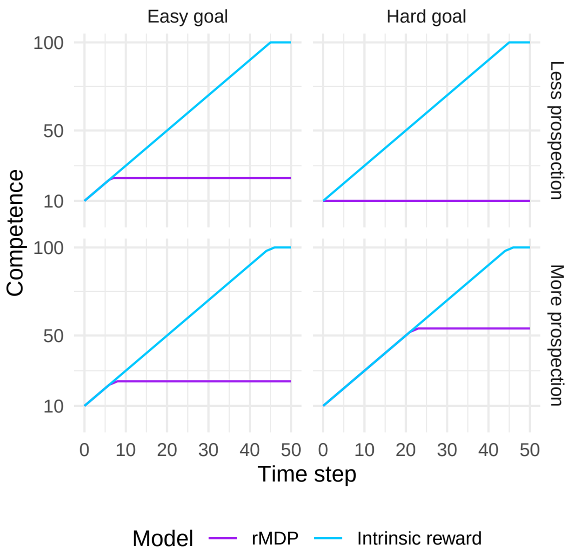

Figure 3: Simulations of two studies that tease apart the rMDP and intrinsic reward models (Robinson et al., 2021; Anand et al., 2023). Less and more prospection corresponded to $\gamma = 0.5$ or 0.9, respectively. Easy and hard goals manipulate the easiest task that yields reward, corresponding to $d =$ 15 and 50, respectively. The agent does not get any reward if they succeed at a task with difficulty below the task difficulty threshold. The agent gets a reward of 100 if they succeed at a task with difficulty equal to or above the threshold.

they found that a reminder about goal attainability mitigated the effect. We modeled sustained unfavorable feedback as a conceptualization of a harder goal. Similar to our simulation of Study 1, goal difficulty corresponded to the minimum task difficulty that yielded reward. We modeled the goal attainability reminder as changes in prospection (controlled by γ). In other words, the reminder reduced the discounting of future rewards. Figure 3 shows that, with less prospection, the rMDP model predicted that the agent would increase their competence to attain goals that they thought were easy, but not goals that they thought were hard (as indicated by the frequent unfavorable feedback). More importantly, this demotivating effect was predicted to mitigate when the agent increased their prospection (because of the reminder). Both these patterns were in line with the empirical finding. However, the intrinsic reward model did not predict these patterns.

Discussion

Our simulations lend plausibility to the claim that central aspects of intrinsic motivation can be formalized as reflective planning over internal states. We focused on competence as a paradigmatic example of an internal state. In the rMDP model, the agent's only goal is to maximize cumulative extrinsic rewards, but it nonetheless places value on competence improvement because this leads to larger extrinsic rewards in the future. A myopic model that ignores these future rewards is insensitive to competence improvement and therefore does not exhibit signatures of intrinsic motivation. Similarly, a "fixed mindset" model that ignores the dynamics of competence will not appear intrinsically motivated, because it operates on the belief that competence cannot change. Further, an intrinsic reward model that assumes an independent drive to increase one's competence fails to explain demotivation.

While our focus was on potential drivers of intrinsic motivation, the rMDP may shed light on related phenomena. For example, why do people sometimes choose difficult tasks over easy ones, despite generally disliking effort (Inzlicht et al., 2018)? As in many models of cognitive effort avoidance, we assume an effort penalty in the utility function (Kool & Botvinick, 2018). However, long-term value can be assigned to effort if it predicts future extrinsic reward (Eisenberger, 1992). Importantly, this effect is mediated by beliefs about the dynamics of competence. Effort may be valued less if its effect on future reward is ephemeral. The promise of increased competence through effort exertion has a potent motivational effect because of its enduring yields. From an evolutionary point of view, this logic may explain why people are motivated to improve themselves even in the absence of obvious extrinsic rewards.

Future directions

When teachers are assigning tasks to students (e.g., in the "Pygmalion" phenomenon), the teacher's choices and the student's choices are entangled. An extension of the framework could consider modeling the teacher as planning in a second-person rMDP, where they anticipate how students' competence would increase as they decide what tasks to assign them. Indeed, in the real world, the students themselves can sometimes find it difficult to judge which tasks are most helpful in improving their competence. This aligns with past work showing that teachers reason about learners' minds when deciding what to teach (Shafto et al., 2014; Vélez et al., 2023).

In the current work, we assumed that the agent's competence could only increase and that they had perfect knowledge about their competence. These assumptions are not core to the framework. The framework can straightforwardly incorporate competence decay (e.g., fatigue, or a skill getting rusty) into the transition function. One future direction is to incorporate the agent's uncertainty about their own competence into the framework, perhaps through a Partially Observable Markov Decision Process (POMDP). Another future direction is to extend the framework to multi-agent settings, where optimal collaboration requires recursive theory of mind (Xiang, Vélez, & Gershman, 2023; C. L. Baker et al., 2017) or some form of heuristics to approximate the process. A multi-agent version of the rMDP has the potential to explain the motivation behind classical phenomena such as comparative advantage, division of labor, and social conformity.

Acknowledgments

We thank Natalia Vélez, Arthur Prat-Carrabin, and Fiery Cushman for helpful discussions. This work was supported by the Kempner Institute for Natural and Artificial Intelligence, and a Polymath Award from Schmidt Sciences.

References

Anand, V., Webb, A., & Wong, C. (2023). Mitigating the demotivating effects of frequent unfavorable feedback about goal progress. *Journal of Management Accounting Research*, *35*(2), 5–32.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, R., Ma, W., Zhao, Y., Wang, S., & Ma, Z. (2020). The results of implementing zone of proximal development on learning outcomes. *International Educational Data Mining Society*.

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1), 49–73.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development*, *78*(1), 246–263.

Bougie, N., & Ichise, R. (2020). Skill-based curiosity for intrinsically motivated reinforcement learning. *Machine Learning*, *109*, 493–512.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology*, *33*(4), 733–756.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review*, *95*(2), 256.

Eisenberger, R. (1992). Learned industriousness. *Psychological review*, *99*(2), 248.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine*, *15*(11), 988–994.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Ericsson, K. A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*. Bodley Head London.

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Fishbach, A., & Woolley, K. (2022). The structure of intrinsic motivation. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(1), 339–363.

Haimovitz, K., & Dweck, C. S. (2017). The origins of children's growth and fixed mindsets: New research and a new proposal. *Child development*, *88*(6), 1849–1859.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in cognitive sciences*, *22*(4), 337–349.

Karayanni, M., & Nelken, I. (2022). Extrinsic rewards, intrinsic rewards, and non-optimal behavior. *Journal of Computational Neuroscience*, *50*(2), 139–143.

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*, 899–908.

Molinaro, G., & Collins, A. G. (2023). Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS Biology*, *21*(7), e3002201.

Pierce, R. L., Cassady, J. C., Adams, C. M., Neumeister, K. L. S., Dixon, F. A., & Cross, T. L. (2011). The effects of clustering and curriculum on the development of gifted learners' math achievement. *Journal for the Education of the Gifted*, *34*(4), 569–594.

Robinson, C. D., Gallus, J., Lee, M. G., & Rogers, T. (2021). The demotivating effect (and unintended message) of awards. *Organizational Behavior and Human Decision Processes*, *163*, 51–64.

Rubie-Davies, C. (2014). *Becoming a high expectation teacher: Raising the bar*. Routledge.

Rubie-Davies, C. (2016). High and low expectation teachers: The importance of the teacher factor. In *Interpersonal and intrapersonal expectancies* (pp. 145–156). Routledge.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the international conference on simulation of adaptive behavior: From animals to animats* (pp. 222–227).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*, 70–82.

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, *131*, 139–148.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vallerand, R. J., Gauvin, L. I., & Halliwell, W. R. (1986). Negative effects of competition on children's intrinsic motivation. *The Journal of Social Psychology*, *126*(5), 649–656.

Vélez, N., Chen, A. M., Burke, T., Cushman, F. A., & Gershman, S. J. (2023). Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences*, *120*(22), e2215015120.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychological review*, *66*(5), 297.

Xiang, Y., Gershman, S. J., & Gerstenberg, T. (2024). A signaling theory of self-handicapping. *PsyArxiv*.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, 105609.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2025). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, *116*, 104699.

Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.

Xiang, Y., Vélez, N., & Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology*, *150*, 101653.

Zou, X., Ma, W., Ma, Z., & Baker, R. S. (2019). Towards helping teachers select optimal content for students. In *Artificial intelligence in education: 20th international conference, aied 2019, chicago, il, usa, june 25-29, 2019, proceedings, part ii 20* (pp. 413–417).