A cognitive basis for theories of intrinsic motivation

Nisheeth Srivastava
Dept of Computer Science
University of Minnesota
Minneapolis, MN 55455
Email: nsriva@cs.umn.edu

Komal Kapoor
Dept of Computer Science
University of Minnesota
Minneapolis, MN 55455
Email: kapoor@cs.umn.edu

Paul R Schrater
Dept of Psychology
University of Minnesota
Minneapolis, MN 55455
Email: schrater@umn.edu

***Abstract*—Since intelligent agents make choices based on both external rewards and intrinsic motivations, the structure of a realistic decision theory should also present as an indirect model of intrinsic motivation. We have recently proposed a model of sequential choice-making that is grounded in well-articulated cognitive principles. In this paper, we show how our model of choice selection predicts behavior that matches the predictions of state-of-the-art intrinsic motivation models, providing both a clear causal mechanism for explaining its effects and testable predictions for situations where its predictions differ from those of existing models. Our results provide a unified cognitively grounded explanation for phenomena that are currently explained using different theories of motivation, creativity and attention.**

I. Introduction

Standard models of reinforcement learning and control theory attempt to model an agent's evaluation of future action selection by assigning cardinal values to external incentives. As [1] recently point out, doing so necessarily forces the goals of the agents to be the same as the goals of the designers, which leads to an impoverishment of the range of behaviors such agents can display. Furthermore, assuming behavior to be driven purely by externally imposed motivations implies that agents will respond reliably to repeated encounters with external incentives. This is seldom the case in natural systems, necessitating research into models of intrinsic motivation.

Intrinsic motivation is typically viewed as motivation that is driven by an interest or enjoyment in the task itself, and exists within the individual rather than in external incentives. Traditional models of control and decision-making ignore the internal incentives that bias the decisions of real organisms away from normative expectations in ways that are predictably irrational [2]. We suggest that any realistic theory of decision-making must, therefore, take these intrinsic reward signals into account and thus be, indirectly, a theory of intrinsic motivation.

There are currently two dominant frameworks for modeling intrinsic motivation. The first, which is predicated fundamentally on the idea that the human brain encodes prediction error as a form of intrinsic reward, has been most extensively explored by Barto and colleagues [3], [4]. This line of research derives from the neuroscientific connections between reinforcement learning and neural dopaminergic activity first proposed by [5] It has been shown that dopamine signals also encode non-hedonic qualities, principally novelty [6], [7]. For example, [8] suggests that dopamine responses could be interpreted as reporting a generalized sense of error in prediction abstracted away from any notion of reward. The principal goal in this paradigm is to use prediction error to quantify intrinsic reward as a supplement to traditional external reward signals to optimize the performance of standard reinforcement learning algorithms. For example, [9] has shown how using such intrinsically motivated algorithms leads to better state-space exploration.

The second framework of intrinsic motivation modeling is related more closely to qualitative theories of psychology and sociology, beginning with Berlyne's [10] hypothesis that organisms are most intrinsically motivated by experiences of an intermediate level of novelty, surprise or complexity. This assertion has since been expanded upon by Czikszenthmihalyi [11] in his development of *flow* psychology. Schmidhuber has further expanded significantly [12] on this idea to create a genre of artificial curiosity algorithms by combining the novelty-seeking drive with another hypothesis that the human brain encodes its descriptions of the environment in information-theoretically efficient ways and derives pleasure or motivation in doing so [13]. Oudeyer's intelligent adaptive curiosity (IAC) [14] is yet another intrinsic motivation model that has found a formal convergence with active learning strategies derived from machine learning [15]. IAC 'loves to learn' i.e., is explicitly directed to attempt to learn as much about the environment as possible. This predisposes it to prefer exploring environmental options of intermediate complexity, a property of central significance in psychological theories of motivation.

While such models of intrinsic motivation are extremely useful in developmental robotics and other applications, they do not contribute much insight into the actual cognitive mechanisms involved in motivation and its effects on behavior. This is principally because all existing models of motivation are, in essence, *normative*. That is, they state how motivation *ought to be* within their own definition, and then directly attempt to achieve this standard algorithmically. Hence, no insights on the actual nature of motivation can be obtained. While the normative assumptions for each of the systems described above are quite reasonable and quite possibly reflect elements of the human motivational structure accurately, the contemporary spectrum of theories of motivation lacks a *positive* exploration of the nature of human motivation. This paper reports our effort at filling in this gap.

978-1-61284-990-4/11/$26.00

In this paper, we describe a model of sequential-decision making that is grounded in contemporaneous understanding of cognition and evolutionary theory. Specifically, we show that by assuming that the value of possible outcomes is encoded relative to all other outcomes, not with respect to some fixed psychological "no reward" condition, and by assuming that evolutionary selection has predisposed organisms to minimize their cognitive decision-making costs while selecting between different actions, we retrieve a formal model of sequential decision-making that improves upon existing decision theory in interesting ways.

We show further that, as expected, our model of decision-making naturally reproduces behavior that is currently hard-coded into theories of intrinsic motivation based on qualitative psychological theories. Our model of decision-making, therefore, provides an interesting implicit model of intrinsic motivation that qualitatively resembles state-of-the-art intrinsic motivation algorithms which explicitly encode an agent's motivational goals. The unforced emergence of such behavior from a formal choice model provides *positive*[1] insight into the nature of human motivation and a number of testable predictions about existing competing models of intrinsic motivation.

II. A SELF-MOTIVATED MODEL OF LEARNING

Any realistic model of decision-making must be consonant with the structure of human motivation, i.e. the intrinsic factors that affect an agent's choice behavior. One approach to developing a positive theory of human motivation would, therefore, be to develop a realistic model of decision-making and then analyze its functional aspects that appear most relevant to the emergence of motivated behavior. This is precisely the approach we adopt in this paper.

In this section, we describe a model of decision-making that we have recently developed [16]. Computational experiments show that its predictions match human behavioral data better than existing theories[17]. We begin by obtaining alternatives to two foundational assumptions that underpin traditional decision theories - the epistemological assumption of cardinal utility values and the teleological assumption of utility maximization.

A. Relative utility

Almost all existing decision theories presuppose the existence of state-specific quantified reward. This assumption ignores the failure of the persistent efforts made by early 20th century psychologists towards finding tractable mappings between physical stimuli and the value judgments of human subjects. In fact, it was not until von Neumann and Morgenstern [18] showed that it is possible, under a set of mathematical axioms (henceforth the VNM axioms) governing the nature of human preferences, to obtain consistent additive values of relative expected utility among various options that quantifications of human preferences could be meaningfully addressed. The von Neumann program is fundamental to the development of reward-based models of decision-making and planning in AI.

However, two major problems have arisen in the course of the adaptation of the VNM approach to computational decision models. First, it has been established by multiple empirical studies that the VNM axioms do not apply to human preferences. Second, in adapting the VNM approach to computational models, the idea that the additive utilities obtained are relative has been ignored, leading to absolute scalar values of reward unhesitatingly (and errantly!) being used in both the AI and reinforcement learning [19] literature.

In our formulation of self-motivated learning, we retreat to the pre-VNM state of understanding of preferences, by assuming that agents can only adopt preferences for particular outcomes relative to others observed in the same context.

B. Evolutionarily meaningful bounded rationality

Basic rational choice theory assumes that rational agents attempt to maximize the reward that they can obtain through their actions. However, this assumption has been shown by multiple behavioral studies to be unrealistic. The principal alternative to this assumption is the bounded rationality approach. However, traditional views of bounded rationality [20] continue to assume that agents attempt to maximize reward under computational constraints. In a subtle but important departure from conventional reward models, environmental phenomena are judged to be valuable only to the extent that they have been judged valuable in the past in our model. Judging utility by *whether* an option has been useful in the past as opposed to *how* useful it is removes the necessity to postulate Platonic rewards embedded in the environment. In our formulation, the relative value of possible outcomes must emerge from the process of sequential choice selection itself.

In our formal model, described below, we postulate that humans are essentially searching for minimal-cost theories about how to choose high value options, where the cost is measured in terms of the complexity of encoding and storing the information needed to reliably make these decisions. We term it *cognitive processing cost*, which is essentially the cost of accessing past beliefs in the agent's memory. In brief, rather than view humans and other natural decision-making agents as reward maximizers, or even constrained reward maximizers, we view them as cognitively efficient need-satisficers. The expectation of efficiency allows us to pose the optimal control and decision problems using standard optimization techniques, with the relevant quantities to be optimized possessing internal as opposed to external ontological significance.

Informally, to make a sequence of decisions, the agent cycles between forming beliefs about the relative worth of options by accessing past experience, making choices, experiencing outcomes and updating these policies to minimize cognitive processing costs for future decisions.

[1]In this paper, we strongly differentiate between positive and normative theories of phenomena. In the interests of clarity, we briefly define positive theories as attempts to describe phenomena as they are. On the other hand, normative theories attempt to describe phenomena as they should be, according to some externally imposed idealized standard.

The formal structure of our model is homologous to the well-known minimum description length (MDL) principle, with the core premise that an agent tries to minimize its cognitive processing cost T while maintaining a 'satisficingly' high level of predictive confidence C in the quality of its choices. The self-motivated learning objective is to minimize a function of the form,

$$\begin{aligned} \underset{\mathbf{x}}{\text{argmin}} \quad & T \\ C_{\text{new}} \quad & \geq C_{\text{old}}. \end{aligned} \tag{1}$$

where T and C are quantified below in terms of policies.

Let the discrete probability distribution $x(s)$ represent an agent's policy, viz. belief about the relative quality of outcomes $s \in \mathcal{S}$ available to it. The surprise experienced by an agent operating with a policy x_a in comparison with policy x_b can be quantified with an information divergence [21] of the form,

$$R(\mathbf{x}_a, \mathbf{x}_b) = \sum_{j=1}^{n_a} \mathbf{x}_a^j(s) \log \frac{\mathbf{x}_a^j(s)}{\mathbf{x}_b^j(s)}. \tag{2}$$

In our model, cognitive processing costs are determined as the cost of accessing a belief in memory. Using information-theoretic arguments, we suggest that the access cost of a belief is determined by its predictive exceptionality, which in turn can be measured as a departure from the usual level of surprise that the agent experiences in making its predictions. We measure the informational exceptionality of a past policy x_{old} (and hence the ease with which it will be available for recall to the agent) as the deviation from the average surprise experienced by the agent R':

$$A(\mathbf{x}_{\text{old}}) = |R(\mathbf{x}, \mathbf{x}_{\text{old}}) - \overline{R}|, \tag{3}$$

where x is the agent's current policy.

Given this measure of ease of memory access for each past policy, a reasonable measure of the processing cost of selecting a subset M' out of the set M of all past policies is the inverse availability-weighted sum of the nominal cost of accessing all policies in M'. Assuming the nominal cost of accessing each policy to be unity, the total cost of memory access T becomes,

$$T = \sum_{\mathbf{x}_i \in \mathcal{M}'} A^{-1}(\mathbf{x}_i), \tag{4}$$

Our measure of the agent's confidence in its ability to predict its environment, $C : x \to [0, 1]$ captures the idea that confidence grows when the policies have low uncertainty and low surprise:

$$C = \frac{1}{C_{\text{max}}} \frac{\log |\mathbf{x}| - H(\mathbf{x})}{\sum_{\text{memory}} R(\mathbf{x}, \mathbf{x}_{\text{old}})}, \tag{5}$$

where the numerator decreases monotonically with respect to the Shannon entropy $H(x)$ of the policy. Note that C is normalized with respect to the greatest value it has previously been observed to achieve.

Any algorithmic solution of our agent's objective function must solve three problems - one, specify a memory update specifying how existing policies are integrated into the agent's current policy; two, specify an environmental update, which shows how the perceived goodness of various outcomes at the present moment, , which we call *reward-inference*[2], are integrated into the agent's current policy; three, specify a combinatorial optimization algorithm specifying which subset of existing policies the agent will *recall* to form its new policy, such that the objective function we have defined above is satisfied. In [16], we describe a direct policy search-based solution to all three problems for simple outcome spaces, resulting in a self-motivated learning algorithm for predicting choices made by agents in sequential settings. The resultant algorithm outputs beliefs corresponding to the relative preference for each of the possible outcomes in the agent's decision context.

III. Experiments

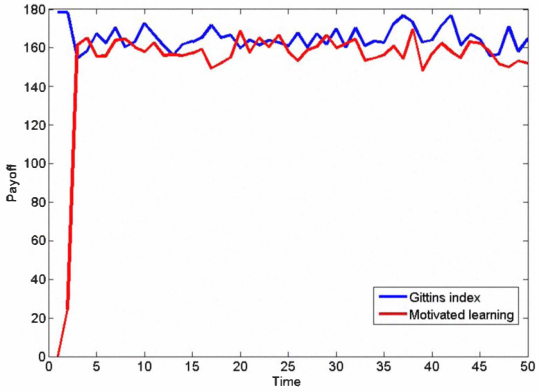
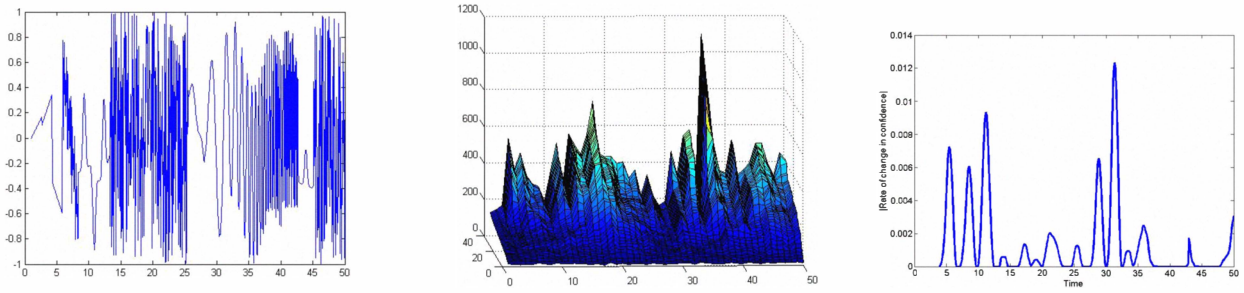


Fig. 1. Self-motivated learning tracks optimal Bayesian performance when natural dynamics transition model is given, i.e. for standard multi-arm bandit problems. The reward on each arm is assumed to be drawn from a Gaussian distribution of known mean and variance and results are reported by averaging over 10 trials of 50 time steps each. Since optimal GI has a model of all arms and their associated rewards, while our algorithm must visit each state individually to obtain a reward estimate, we allow our algorithm to simulate trials, viz. visit multiple (all) arms in each trial.

Self-motivated learning closely resembles reinforcement learning in its functionality, tracking exemplars of useful experiences and promulgating choice policies based on its recall of exceptional experiences. For statistically normal domains, where useful experiences are typical, the predictions of the self-motivated model closely track theoretically optimal values. Fig 1 shows that our algorithm performs close to optimality in the standard multi-arm bandit setting, where the Bayesian Gittins Index (GI) solution is known to be both optimal and analytically tractable. We are thus assured of a baseline match between our algorithm's predictions and those obtained using other reinforcement learning methods. We now demonstrate how the behavior of our algorithm also replicates the behavior of explicit models of intrinsic motivation as well as human-like behavior in interesting ways.

[2]Our model assumes that sensory data is encoded into the space of possible outcomes as a relative preference by existing neuronal processes. Thus, our usage of the term *reward-inference* accentuates the fact that it is obtained after perceptual processing of environmental stimuli.



(a) Sample environment with varying function complexity

(b) Total exploration over time with R-IAC algorithm

(c) Rate of change in confidence for self-motivated learning

Fig. 2. Results for complexity affinity experiments

A. *Affinty for intermediate complexity*

Baranes and Oudeyer [22] have recently proposed Robust-IAC as an improved version of the original IAC algorithm and demonstrated its functionality on a set of experimental test examples. Here, we compare the functionality of self-motivated learning against RIAC[3] on the simplest of the test examples provided by [22]. The dataset is a simple 1-dimensional environment containing two noisy parts [0.25 0.5] and [0.8 1] and an 'increasing difficulty part' [0.5 0.8], as shown in Figure 2(a). Unsurprisingly, as seen in Figure 2(b), since doing so is its explicit algorithmic goal, R-IAC spends more time exploring the region of increasing complexity and ignores the noisy and quiet regions.

Whereas IAC uses time spent exploring a region as a proxy for motivation levels, we use the rate of change in predictive confidence as our metric for measuring motivation. The intuition behind this mapping is that the agent's self-perception of ability to engage with the environment is contingent on it being able to predict future choices accurately. Doing so causes an increase in confidence. Thus, an active learning strategy manifests itself in our cognitive model as a confidence-improving heuristic. Thus, the absolute value of the rate of change of confidence takes the place of exploration time as our measurable variable. Note in Figure 2(c) that our algorithm is biased towards simplicity and therefore, along with the region of increasing complexity identified by IAC, identifies the initial low complexity regions as interesting just as well. The motivation predictions of both algorithms, however, present as fairly similar.

B. *Prisoners' dilemma and human motivation*

While informational complexity provides an interesting abstract framework for understanding the causal structure of human motivation, it is necessary for the predictions of positive theories of intrinsic motivation to resemble the behavior of human subjects at a behavioral level as well. Of particular interest are known scenarios where the predictions of normative models of behavior diverge from actual behavior.

[3]The RIAC and testbed code for this experiment was used as is from http://flowers.inria.fr/riac-software.zip

TABLE I
BASIC SETUP FOR A PRISONERS' DILEMMA TASK. NOTE THAT $T > R > P > S$.

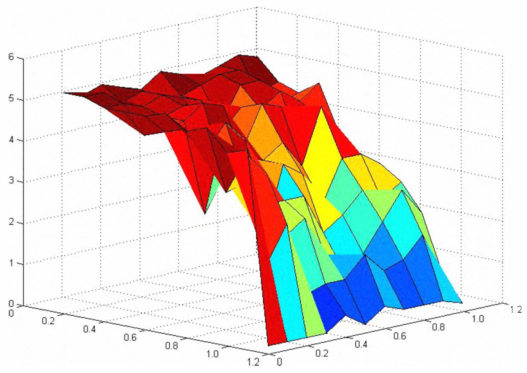
	Cooperate	Defect
Cooperate	R,R	S,T
Defect	T,S	P,P

Here, we demonstrate the emergence of super-rational [23] behavior in our model's predictions in an iterated prisoner's dilemma (PD) setting.

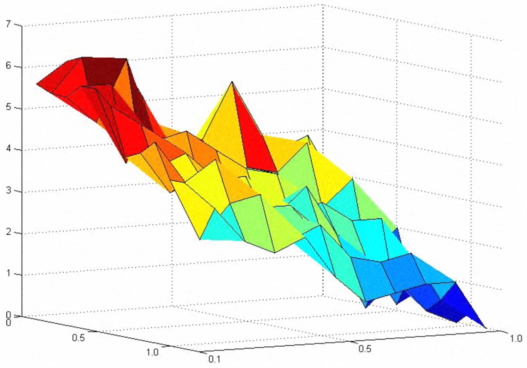
The prisoner's dilemma problem has been extensively studied in the game theory and sociology literature [24]. In the basic PD setting shown in Table I, the utilitarian rational strategy for either player is to defect, which, sadly, leads to a poor outcome for both players. Blindly cooperating, however, is even worse, which means that a player must have a responsive strategy for dealing with an opponent. Playing multiple PD games with the same parameters affords the opportunity to discover the opponent's strategy and adjust one's own. This makes iterated PD an interesting problem domain for testing agents that purport to behave in a manner consonant with autonomously motivated humans.

It is possible to describe the space of strategies in two player PD games as a set $\mathcal{S}$ of tuples $S(p1, p2)$, where $p1$ is the probability with which the agent defects if the opponent cooperated and $p2$ the probability that the agent defects if the opponent defected on the previous turn. We evaluate the performance of two different agents: one using a hard-coded tit-for-tat (TFT) strategy (cooperate if the other player cooperated last turn, defect otherwise), the other beginning agnostically and learning an appropriate strategy for dealing with the agents it encountered using self-motivated learning. We assessed the performance of both agents against a grid of mixed strategies obtained by varying $p1$ and $p2$ between 0 and 1 in increments of 0.1 along two axes.

Figure 3(a) shows the payoff obtained by TFT. To briefly orient the reader, the rightmost corner represents the case where the opponent's strategy is to always defect, i.e. $S(1, 1)$ which causes both players to continually defect (tit-for-tat) and obtain low payoff. On the other extreme, the *saintly* strategy $S(0, 0)$ lives in the leftmost corner. Here, both players



(a) Payoff surface for a theoretically optimal tit-for-tat strategy



(b) Mean payoff surface for a strategy learned by a self-motivated agent over 50 trials

Fig. 3. Results for prisoners' dilemma experiment

continually cooperate and receive the intermediate payoff.

Figure 3(b) shows the average payoff obtained by our model over 50 iterations on each grid point. Our model learns, for the most part, a strategy that closely resembles TFT. The TFT strategy is well-known in the PD literature both for having been postulated as a model of human behavior as a theory of reciprocal altruism [25] and for being exceptionally robust as a game-theoretic strategy against other strategies in iterated PD games [24]. Unlike existing adaptive agents [26], [27], our agent does not possess an explicit model of its adversary's choices. Its preference to cooperate, therefore is intrinsic, not game-theoretically planned. The motivation to select the cooperative option instead of the defect option inspite of a lower extrinsic payoff is explained by the additional intrinsic payoff obtained by the agent not having to continually change its prediction (and incurring a cognitive cost) once it has begun cooperating.

This finding suggests cognitive efficiency as a causal mechanism for learned altruistic behavior, i.e., the cost of predicting other agents' behavior rises under antagonistic choices, causing altruistic choices to be preferred. This simple explanation presents a parsimonious mechanism for the development of learned altruism [25]. Reciprocal altruism, as seen in tit-for-tat repeated PD games, is widely acknowledged as a powerful ultimate explanation for human altruism in small and stable groups [28]. One of the principal criticisms levied against direct reciprocity theories is that they have heretofore assumed that agents cooperate in anticipation of future reward. Such a utilitarian mechanism cannot explain the emergence of strong reciprocity - cooperative actions performed in the absence of external reward. Our experimental results demonstrate that intrinsic payoffs support the development of approximately reciprocal strategies in two-player games and potentially resolve this criticism of reciprocal altruism models. This view is further supported by evidence from neurobiology that suggests that individuals experience particular subjective rewards from mutual cooperation [29].

We note, in passing that our model's learned behavior deviates from TFT in interesting ways. For example, in cases where the opponent is too saintly and does not retaliate, our strategy learns to exploit it by electing to defect continually. Further analysis of these and other deviations of our model's predictions from TFT presents an interesting direction for future work.

IV. Discussion

Both the Barto approach (exemplified in [9]) and the Oudeyer [14] computational models of motivation share an underlying expectation with our model of choice-selection in desiring to improve the ability to learn, and seeking to associate intrinsic motivation with an agent's self-perception of its predictive ability. However, by disregarding the cognitive machinery involved in the process of learning and tying the intrinsic motivation directly to the statistically measured learning rate, the Barto approach fails to replicate the dyadic nature of natural agents' motivations, viz. they stop trying to learn both when the environment becomes too predictable and when it becomes too unpredictable, an insight that the Oudeyer model captures. Since such dyadic behavior is characteristic across multiple empirically supported theories of motivation [10], [13], it is difficult to consider models that do not retrieve this property to be realistic.

The Oudeyer model encodes this dyadic pattern explicitly based on the intuition that agents seek out experiences with intermediate novelty, which, on the surface, directly contradicts our model which assigns greater significance to both extremely unpredictable and extremely predictable events. However, it it important to remember that while Oudeyer et al are directly trying to model motivation, our model is one of memory access, which operationally creates a model of motivation. A closer examination of our model reveals that situations with high predictability, if sustained, cause the agent to stop updating its beliefs, since no further increases in confidence are possible once certainty is assured. On the other hand, highly unpredictable decision instances lead to low confidence, preventing further updates of similarly confidence-reducing novel instances. In both cases, the rate of change in confidence, and hence the motivation level is

decreased, retrieving precisely the behavior we expect from a realistic model. Our model differs from IAC in that the need to learn is not encoded explicitly as the algorithm's objective function. Hence, as we describe in [17], our model can behave suboptimally in a manner that reflects learned helplessness through past experience with highly unpredictable or adversarial environments. In such cases, our model believes that the best choice is to stick with its existing strategy, even though a different strategy might lead to better learning in the new environment. Thus, a testable prediction contrasting IAC with our model is that IAC will not show confirmation biases in its learning strategy, unlike realistic biological agents and the self-motivated learning model.

As we see above, the comparison between IAC and our model leads to the conclusion that motivation is proportional to the rate of change in predictive confidence. This is very interesting, since our definition of predictive confidence is very similar to Schmidhuber's notion of descriptive complexity reduction. If we disregard the denominator in our confidence term, we retrieve entropy reduction as a goal for both Schmidhuber's algorithm and ours. Our approach differs from Schmidhuber's in that we predicate cognitive cost reduction as the primary objective of the organism. A simple example to accentuate this difference is that teenaged Schmidhuber agents would enjoy cleaning their rooms. On the other hand, self-motivated agents will weigh potential environmental simplicity against the cognitive cost of putting things in their right places and demur if the latter appears too high. We leave it as an exercise for the reader to determine which model of motivation appears more realistic.

Our model also provides novel insight into a potential underlying mechanism for *flow*, as defined by Czikszenthmihalyi. We hypothesize that the state of flow reported in [11], [30] is simply a state of low cognitive cost utilization, as defined in our model. An important testable prediction arising from this hypothesis is that, whereas it is suggested in [30] that flow arises only in situations where high competence meets a high level of challenge, according to our theory, a state of flow should be achievable irrespective of the complexity of the task at hand. All that is necessary is that the agent have high confidence in its ability to navigate its current environment.

V. Conclusion

In place of existing normative accounts of decision-making, we have introduced a new model of self-motivated behavior that makes predictions in concordance with those obtained from state-of-the-art models of intrinsic motivation. Simple experiments show that our model behaves like a classical reinforcement learning algorithm for standard domains, replicates behavior predicted by existing motivation models, and demonstrates human-like behavior in simple two-player game-theoretic simulations. Our work provides a theoretical bridge between choice models and theories of motivation, and presents a unified and cognitively grounded explanation for phenomena currently explained by different qualitative psychological theories of motivation, attention and curiosity.

References

[1] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," in *IEEE Transactions on Autonomous Mental Development*, 2010.

[2] D. Ariely, *Predictably irrational: The Hidden Forces That Shape Our Decisions*. Harper Collins, 2009.

[3] S. Singh, A. Barto, and N. Chentanez, "Intrinsically motivated reinforcement learning," in *Proceedings of Advances in Neural information processing systems (NIPS) 17*, 2005.

[4] J. Sorg, S. Singh, and R. L. Lewis, "Internal rewards mitigate agent boundedness," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[5] P. Montague, P. Dayan, and T. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive hebbian learning," *Journal of Neuroscience*, vol. 16, pp. 1936–47, 1996.

[6] P. Dayan and W. Belleine, "Reward, motivation and reinforcement learning," *Neuron*, vol. 36, pp. 285–298, 2002.

[7] S. Kakade and P. Dayan, "Dopamine: Generalization and bonuses," *Neural Networks*, vol. 15, pp. 549–559, 2002.

[8] J. Horvitz, "Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events," *Neuroscience*, vol. 96, no. 4, pp. 651–656, 2000.

[9] O. Simsek and A. Barto, "An intrinsic reward mechanism for efficient exploration," in *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML)*, 2006.

[10] D. Berlyne, *Conflict, Arousal and Curiosity*. McGraw-Hill, 1960.

[11] M. Czikszenthmihalyi, *Flow - the Psychology of Optimal Experience*. Harper Perennial, 1991.

[12] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990-2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[13] ——, "Curious model-building control systems," in *Proceedings of the International Joint Conference on Neural Networks*, 1991, pp. 1458–1463.

[14] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 2, pp. 265 –286, 2007.

[15] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 590–604, 1992.

[16] N. Srivastava and P. Schrater, "An evolutionarily motivated model of decision-making under uncertainty," Available at SSRN: http://ssrn.com/abstract=1687205, 2010.

[17] ——, "A value-relativistic decision theory predicts known biases in human preferences," in *Proceedings of the* 33^{rd} *Annual Meeting of the Cognitive Sciences Society (to appear)*, 2011.

[18] J. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[19] A. Barto and R. Sutton, *Reinforcement Learning: an introduction*. Univesity of Cambridge Press, 1998.

[20] A. Rubinstein, *Modeling bounded rationality*. Prentice-Hall, 2003.

[21] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[22] A. Baranes and P.-Y. Oudeyer, "R-IAC: Robust intrinsically motivated exploration and active learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 3, pp. 155–169, 2009.

[23] D. R. Hofstadter, *Metamagical Themas: questing for the essence of mind and pattern*. Bantan Dell, 1985.

[24] R. Axelrod, *The Evolution of Cooperation*. Basic Books, 1984.

[25] R. Trivers, "The evolution of reciprocal altruism," *Quarterly Review of Biology*, vol. 46, pp. 35–57, 1971.

[26] D. Kraines and V. Kraines, "Learning to cooperate with pavlov: an adaptive strategy for the iterated prisoner's dilemma with noise," *Theory and Decision*, vol. 35, pp. 107–150, 1993.

[27] M. Nowak and K. Sigmund, "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game," *Nature*, vol. 364, pp. 56–58, 1993.

[28] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, no. 6960, pp. 785–791, 2003.

[29] J. K. Rilling, D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts, "A neural basis for social cooperation," *Neuron*, vol. 35, no. 2, pp. 395 – 405, 2002.

[30] M. Czikszenthmihalyi, *Creativity-Flow and the Psychology of Discovery and Invention*. Harper Perennial, 1996.