Affective-Cognitive Learning and Decision Making: A Motivational Reward Framework For Affective Agents

Hyungil Ahn and Rosalind W. Picard

MIT Media Lab, Cambridge MA 02139, USA
{hiahn, picard}@media.mit.edu

Abstract. In this paper we present a new computational framework of affective-cognitive learning and decision making for affective agents, inspired by human learning and recent neuroscience and psychology. In the proposed framework 'internal reward from cognition and emotion' and 'external reward from the external world' serve as motivation in learning and decision making. We construct this model, integrating affect and cognition, with the aim of enabling machines to make smarter and more human-like decisions for better human-machine interactions.

1 Introduction

Recent affective neuroscience and psychology have reported that human affect and emotional experience play a significant, and useful, role in human learning and decision making. In normal individuals, the covert biases related to previous emotional experience of comparable situations assist the reasoning process and facilitate the efficient processing of knowledge and logic necessary for conscious decisions [3]. It has been argued that positive affect increases intrinsic motivation [6]. Also, the neurotransmitter, dopamine serves as motivational reward and the dopamine (DA) circuit plays an integral part in the reinforcing or reward learning in the emotional brain. The DA circuit is significant in the emotional function of the brain - the basic impulse to search, investigate, and make sense of the environment [8]. Although a model of curiosity for self-motivated development has been researched [11] and a model of intrinsically motivated reinforcement learning based on the options framework has been proposed [2], the extension of cognitive theory to explain and exploit the role of affect in learning is in its infancy [10]. This paper presents a new computational framework of affective-cognitive learning and decision making, inspired by human learning and recent neuroscience and psychology.

2 Affective-Cognitive Learning and Decision Making

2.1 The Affective Agent

For humans or other animals, motivation is essential in their learning and decision making and 'wanting' is directly linked to motivation [4, 8]. Also, 'wanting'

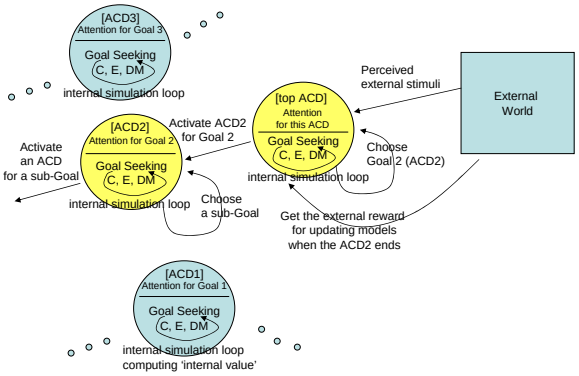

Fig. 1. The hierarchical network of affective-cognitive decisions for affective-cognitive learning and decision making: C, E and DM denote cognition-related models, emotion-related models, and the decision making model, respectively. For example, Goal 2 is chosen in the top ACD and then ACD2 is activated.

can be decomposed into 'implicit incentive salience wanting' and 'cognitive incentive goal-directed wanting' [4]. Inspired by recent affective neuroscience, our approach assumes that an *affective agent* has both cognition and emotion and its motivation can be modeled by reward and that 'internal reward from cognition and emotion' and 'external reward from the external world' can explain motivation in its learning and decision making. Further, we propose a mental or internal simulation loop which computes internal reward for each imagined decision. The internal simulation loop functions as 'expectation or imagination loop' for thinking and 'as-if-body-loop'[5] for feeling in an affective agent. This loop assesses the total weighted internal reward from cognition and emotion in the framework.

For example, the MIT Media Lab's Affective Computing Group in collaboration with Robotic Life Group is building a robotic computer (RoCo) that moves its monitor 'head' and 'neck'. RoCo is an affective agent that is being designed to interact with users in a natural way for applications such as learning, rapport-building, interactive teaching, and posture improvement.

Also, RoCo can learn and decide how to behave to maximize external rewards from the user. In a simulation of RoCo, we adopted a user model where a user tended to give a reward ('show pleasure') under these decisions: RoCo decides to 'track' when the user state is 'attentive', RoCo decides to 'entertain' when the user state is 'distracted', and RoCo decides to 'stretch' when the user state is 'slumped'. From the perspective of RoCo, the user is the external world. We assume that RoCo can perceive the external world (the user) and by means of an *attention model*, can recognize one of a user's five different states - 'attentive', 'distracted', 'slumped', 'showing pleasure' (as a reward for RoCo), 'showing displeasure' (as a punishment for RoCo) - with an accuracy of 90%. A user's state recognized by RoCo can be viewed as part of the *cognitive state* of RoCo, since once it is recognized it is merely a label to RoCo.

We construct RoCo to have two goals in the simulation: to try to make the user 'attentive' or to 'show pleasure'. RoCo's internal cognitive rewards are always given toward encouraging itself to make its goals. Thus, RoCo has a positive internal cognitive reward for decisions that result in the user being 'attentive' or 'showing pleasure'. We call such a state a *'cognitively wanted state'*.

RoCo also has two internal *emotional states* labeled 'feeling good' and 'feeling bad', and we bias RoCo to want to feel good rather than bad. In other words, 'feeling good' is RoCo's *'emotionally wanted state'*. When the user 'shows pleasure' the 'feeling good' state becomes more likely, while when the user 'shows displeasure' RoCo's 'feeling bad' state becomes more probable.

Each of RoCo's decisions – 'tracking', 'entertaining', etc., may be composed of lower-level decisions. Thus, we propose that an affective agent should have a hierarchical network of decisions, and each decision should have an internal affective-cognitive decision-making process that can decide which lower-level decision is appropriate given the current cognitive and emotional states.

2.2 The Hierarchical Network of Affective-Cognitive Decisions

In the previous section we proposed a hierarchical network of decisions for an affective agent. However, we prefer to use a more special and meaningful term *'affective-cognitive decision (ACD)'* rather than to use a general term 'decision'. According to this terminology, an affective agent has a hierarchical network of ACDs (Figure 1) and an ACD is composed of temporally sequential lower-level ACDs or lower-level primitive skills. An *affective-cognitive decision (ACD)* has a certain goal and is able to start (activate) or stop (deactivate) other ACDs for achieving the goal. An *episode* is completed when a cognitive state of the goal is achieved. An ACD is activated when a higher-level ACD sends an activation signal to it, and deactivated when a higher-level ACD sends a deactivation signal to it or it completes an episode.

Our model is framed probabilistically. Let c, e, $\hat{e}$, d, c' and e', respectively, denote the current cognitive state (c), the current emotional state (e), the probability distribution of the current emotional state ($\hat{e}$), the current lower-level ACD (d) being chosen, the next cognitive state (c') after the current lower-level ACD, and the next emotional state (e'). An ACD maintains an attention model, a cognition model $\Pr(c' \mid c, d)$, an emotion model $\Pr(e' \mid c', e)$, a decision-making model $\Pr(d \mid \hat{e}, c)$, a cognitive reward model $Q_{CR}(c, d)$ and an emotional reward model $Q_{ER}(\hat{e}, c, d)$.

Whenever a lower-level ACD d has been made and acted upon, the corresponding *external reward* from the external world can be assessed and be used for updating the models of the ACD with the self-motivating *internal reward*.

An ACD also maintains an *attention model* to extract useful features from the perceived multi-modal external stimuli (from the external world, e.g. recognizing postural and facial features of the user). Each ACD can have its own attention model for its goal. Thus, an attention model can lower the dimensionality of the cognitive state space for each ACD. In this paper we are not focusing on how an attention model for an ACD works, rather we assume that each ACD has a given

attention model for its goal. A *cognitive state* $c \in C = \{1, \cdots, |C|\}$ for an ACD results from the attention model. In addition, we assume that in each time step or time window an ACD can have only one cognitive state although it might be false in the case that the external world is partially observable. Actually in this case, we think that it might be preferable to extend the framework to employ a probability distribution of cognitive states rather than the cognitive state itself. The *cognition model* $\Pr(c' \mid c, d)$ is a cognitive state transition model or a world simulation model. The cognition model $\Pr(c' \mid c, d)$ together with the *model of the cognitive reward* $R_{CR}(c')$ will allow us to represent a kind of 'cognitive self-motivation' in choosing the ACD d.

An *emotional state* e is modeled as a discrete state in the current framework: $e \in E = \{1, \cdots, |E|\}$. In the current RoCo simulation, E is the simple binary set {'feeling bad', 'feeling good'}. However, in each time step we assume that e is not explicitly known and only the probability distribution $\hat{e}$ of the current emotional state e is known: $\hat{e} = (\hat{e}_1, \cdots, \hat{e}_{|E|})$ where $\hat{e}_j$ is the probability that the current emotional state e is $j \in E = \{1, \cdots, |E|\}$. The *emotion model* $\Pr(e' \mid c', e)$ can be considered as the cognitive appraisal model for emotional states. For instance, Roseman's cognitive appraisal model can give rise to seventeen emotions - surprise, fear, hope, joy, relief, sadness, distress, frustration, disgust, liking, dislike, anger, contempt, pride, regret, guilt, shame - by a small number of appraisals such as unexpectedness, motivational state and situational state, probability, control potential, problem type and agency [9]. By combining the *emotion transition model* $\Pr(e' \mid \hat{e}, c, d) = \sum_{i=1}^{|C|} \Pr(e' \mid c' = i, \hat{e}\,) \Pr(c' = i \mid c, d) = \sum_{i=1}^{|C|} \sum_{k=1}^{|E|} \hat{e}_k \Pr(e' \mid c' = i, e = k) \Pr(c' = i \mid c, d)$ and the *model of the emotional reward* $R_{ER}(e')$, we will construct a form of 'emotional self-motivation' for choosing an ACD d. Moreover, the *decision-making model* $\Pr(d \mid \hat{e}, c)$ can be thought of as a seeking model of sub-goals or lower-level ACDs.

2.3 The Algorithm

Each ACD runs the following algorithm which has two main parts: the actor (choosing a lower-level ACD through the internal simulation loop while the ACD is activated) and the critic (updating models of the ACD whenever the previously chosen lower-level ACD is finished).

In the mental or internal simulation loop the agent can simulate an imagined lower-level ACD according to the decision-making model $\Pr(d \mid \hat{e}, c)$ which results from the Boltzmann selection using the state-ACD value $Q_{DM}(\hat{e}, c, d)$, and obtain an internal cognitive value $v_{int,C}$ and an internal affective value $v_{int,E}$. In order to compute the total internal value v_{int}, the internal cognitive and affective values are weighted by a willpower factor $\lambda_C(d)$ and a feeling factor $\lambda_E(d)$, respectively. Cognition and emotion influence the decision-making process through changing willpower and feeling factors as well as through internal values. The agent simulates different lower-level ACDs until it finds out a lower-level ACD

which has positive internal value ($v_{int} > 0$). When such a lower-level ACD is located, the agent actually performs it and receives the external reward r_{ext}.

Whenever the previously chosen lower-level ACD is finished, the reward models and the decision-making model are updated in the reinforcement framework. The computation of the internal cognitive reward and the internal emotional reward suggests that we also construct models related with cognitively wanted states $R_{CR}(c')$ and with emotionally wanted states $R_{ER}(e')$, respectively. The value of $R_{CR}(c')$ models how much the next cognitive state c' is cognitively wanted to achieve the goal of the ACD after acting upon the current lower-level ACD d at the current cognitive state c. Similarly the value of $R_{ER}(e')$ models how much the next emotional state e' is emotionally wanted: for instance, if e' has positive valence, it is more emotionally wanted. The internal cognitive reward and the internal emotional reward, respectively, are used for updating the cognitive reward model $Q_{CR}(c, d)$ and the emotional reward model $Q_{ER}(\hat{e}, c, d)$.

The following is the algorithm for any ACD. However, the top ACD is special and different from other ACDs in that it is always activated.

The algorithm for each affective-cognitive decision (ACD)

Loop forever

if (the activation signal is received from a higher-level ACD) Activate this ACD and start a new episode and get the probability distribution of the current emotional state $\hat{e}$ from the higher-level ACD

if (the deactivation signal is received from a higher-level ACD) Send the deactivation signal to current lower-level ACD and deactivate this ACD

if (an episode of this ACD is completed) Send the completion signal to the higher-level ACD and deactivate this ACD

The *Attention model* extracts useful features from the perceived multi-modal external stimuli for the goal of this ACD and outputs the features as a cognitive state $c \in C = \{1, \cdots, |C|\}$.

Update the probability distribution of the current emotional state $\hat{e} = (\hat{e}_1, \cdots, \hat{e}_{|E|})$ where $\hat{e}_j$ is the probability with which the current emotional state $e \in E = \{1, \cdots, |E|\}$ is j: $\hat{e}_{old}$ is the probability distribution of the previous emotional state and assume that the emotion model $\Pr(e' \mid c', e)$ is given, then

$$\begin{aligned}\hat{e}_j &= \Pr(e = j|c, \hat{e}_{old}) = \sum_{k=1}^{|E|} \Pr(e = j, e_{old} = k|c, \hat{e}_{old}) \\ &= \sum_{k=1}^{|E|} \Pr(e = j|e_{old} = k, c, \hat{e}_{old}) \Pr(e_{old} = k|c, \hat{e}_{old}) \\ &= \sum_{k=1}^{|E|} \hat{e}_{old,k} \Pr(e = j|c, e_{old} = k) \qquad \text{for all } j \in E\end{aligned}$$

if (the cognitive state c is different from the previous cognitive state) flagCogStateChanged = 1
else flagCognitiveStateChanged = 0

// the actor part

if (this ACD is activated and (previously chosen lower-level ACD is completed or flagCogStateChanged == 1))

α is the discount factor, $0 < \alpha < 1$.

β is the inverse temperature for the Boltzmann selection, γ is the learning rate, $0 < \gamma < 1$.

The number of steps in the internal simulation loop nInternalStep = 0, and the maximum number of steps nMaxInternalStep = 100.

For each lower-level ACD $k \in D = \{1, \cdots, |D|\}$, willpower factor $\lambda_C(k) = 1$, feeling factor $\lambda_E(k) = 1$, the discount factors for willpower and feeling factors are $0 < \varepsilon_C < 1$ and

$0 < \varepsilon_E < 1$.

while ($v_{int} < 0$ and nInternalStep < nMaxInternalStep) // the internal loop

Choose a lower-level ACD $d \in D = \{1, \cdots, |D|\}$ from

$$\Pr(d = k|\hat{e}, c) = \frac{\exp(\beta Q_{DM}(\hat{e}, c, k))}{\sum_{l=1}^{|D|} \exp(\beta Q_{DM}(\hat{e}, c, l))} = \frac{\exp(\beta \sum_{j=1}^{|E|} \hat{e}_j Q_{DM}(e = j, c, k))}{\sum_{l=1}^{|D|} \exp(\beta \sum_{j=1}^{|E|} \hat{e}_j Q_{DM}(e = j, c, l))}$$

Get the internal cognitive value $v_{int,C} = Q_{CR}(c, d)$

Get the internal affective value $v_{int,E} = Q_{ER}(\hat{e}, c, d) = \sum_{j=1}^{|E|} \hat{e}_j Q_{ER}(j, c, d)$

Get the internal value $v_{int} = \lambda_C(d) \times v_{int,C} + \lambda_E(d) \times v_{int,E}$

Update willpower and feeling factors when cognition and emotion conflict with each other (Note: The following is an example model which assumes that factors are exponentially decreasing):

if ($v_{int,C} > 0$ *and* $v_{int,E} < 0$)
  if ($v_{int} > 0$) $\lambda_C(d) = \varepsilon_C \lambda_C(d)$
  else if ($v_{int} < 0$) $\lambda_E(d) = \varepsilon_E \lambda_E(d)$
else if ($v_{int,C} < 0$ *and* $v_{int,E} > 0$)
  if ($v_{int} > 0$) $\lambda_E(d) = \varepsilon_E \lambda_E(d)$
  else if ($v_{int} < 0$) $\lambda_C(d) = \varepsilon_C \lambda_C(d)$

nInternalStep = nInternalStep + 1

end // end of the internal loop

if (flagCogStateChanged == 1) {
 if (the lower-level ACD d is different from the current running lower-level ACD)
  Send the deactivation signal to the current running lower-level ACD and activate the lower-level ACD d
}
Send the activation signal to the lower-level ACD d

end // end of the actor part

// the critic part

if (the completion signal is received from the lower-level ACD d)

Reread the probability distribution of the current emotional state $\hat{e}$ (because $\hat{e}$ may have changed in the lower-level ACD d)

Perceive the new cognitive state c_{new} and the new cognitive reward $r_{CR,new}$

Get the external reward r_{ext}

Update the model of the cognitive reward: $R_{CR}(c_{new}) \leftarrow (1 - \gamma) R_{CR}(c_{new}) + \gamma r_{CR,new}$

Update the model of the external reward: $R_{ext}(c, d) \leftarrow (1 - \gamma) R_{ext}(c, d) + \gamma r_{ext}$

Update the cognition model (world simulation model):
$Q_C(c, d, c_{new}) \leftarrow Q_C(c, d, c_{new}) + 1$

$$\Pr(c' = i|c, d) = \frac{Q_C(c, d, c' = i)}{\sum_{k=1}^{|C|} Q_C(c, d, c' = k)} \quad \text{for all } i \in C$$

Update the cognitive reward model:

$$Q_{CR}(c, d) = \sum_{j=1}^{|C|} \left\{ R_{CR}(c' = j) + \alpha \max_{d'} Q_{CR}(j, d') \right\} \Pr(c' = j|c, d)$$

Update the emotional reward model: for all $j \in E$

$$Q_{ER}(j,c,d) = \sum_{k=1}^{|C|}\sum_{l=1}^{|E|}\left\{R_{ER}(e'=l) + \alpha \max_{d'} Q_{ER}(l,k,d')\right\}\Pr(e'=l|c'=k,e=j)\Pr(c'=k|c,d)$$

$$\left(\text{Note: } \Pr(e'=l|c'=k,e=j)\Pr(c'=k|c,d) \ = \Pr(e'=l,c'=k|e=j,c,d)\right)$$

Update the decision-making model:

$$Q_{DM}(j,c,d) = R_{ext}(c,d) + \sum_{k=1}^{|C|} R_{CR}(c'=k)\Pr(c'=k|c,d)$$
$$+ \sum_{k=1}^{|C|}\sum_{l=1}^{|E|} R_{ER}(e'=l)\ \Pr(e'=l|c'=k,e=j)\Pr(c'=k|c,d)$$
$$+ \alpha \sum_{k=1}^{|C|}\sum_{l=1}^{|E|} \max_{d'} Q_{DM}(l,k,d')\Pr(e'=l|c'=k,e=j)\Pr(c'=k|c,d) \qquad \text{for all } j \in E$$

end // end of the critic part

2.4 Preliminary Experimental Results

Now we describe preliminary experimental results showing the performance of the algorithm. We applied the algorithm to a simulation of RoCo, a physically animate computer described in Section 2.1. The simulation of RoCo and detailed results can be seen in [1].

RoCo's goals in this simulation are to make the user 'attentive' and to 'show pleasure' (give a reward). We use the top ACD for this joint goal and RoCo has three kinds of lower-level ACDs for the top ACD: $D = \{$ 'tracking', 'stretching', 'entertaining' $\}$. RoCo's cognitively wanted states are the user's 'attentive' or 'showing pleasure' states: $r_{CR} = 1$ for c' = 'attentive (to user's task)' or 'showing pleasure', otherwise r_{CR} = -1. Moreover, we assume that RoCo's emotionally wanted state is 'feeling good': $R_{ER}(e'$ = 'feeling good') $= 1, R_{ER}(e'$ = 'feeling bad') $= -1$. RoCo obtains an external reward $r_{ext} = 1$ or -1 whenever the user shows pleasure or displeasure, respectively. We also assume that the 'entertaining' ACD has three kinds of lower-level ACDs: for the 'entertaining' ACD, $D = \{$'ent1', 'ent2', 'ent3'$\}$. The 'entertaining' ACD is assumed to use the same kinds of cognitive states and emotional states as in the top ACD. However, differently from the top ACD, the cognitively wanted state of the 'entertaining' ACD is the user's 'showing pleasure' state only.

For $E = \{$'feeling bad (1)','feeling good (2)'$\}$, we use an emotion model $\Pr(e' \mid c', e)$ as follows. If c' is cognitively wanted, $\Pr(e'=2 \mid c', e=1) = 0.8$, $\Pr(e'=1 \mid c', e=1) = 0.2$, $\Pr(e'=2 \mid c', e=2) = 1.0$, $\Pr(e'=1 \mid c', e=2) = 0.0$. If c' is cognitively unwanted, $\Pr(e'=2 \mid c', e=1) = 0.0$, $\Pr(e'=1 \mid c', e=1) = 1.0$, $\Pr(e'=2 \mid c', e=2) = 0.2$, $\Pr(e'=1 \mid c', e=2) = 0.8$. If c' is neither cognitively wanted nor unwanted, $\Pr(e'=2 \mid c', e=1) = 0.3$, $\Pr(e'=1 \mid c', e=1) = 0.7$, $\Pr(e'=2 \mid c', e=2) = 0.7$, $\Pr(e'=1 \mid c', e=2) = 0.3$. This emotion model functions in such a way that the probability of feeling good increases if c' is cognitively wanted, the probability of feeling bad increases if c' is cognitively unwanted, and the probability of each emotional state goes exponentially to the neutral value (0.5), if c' is neither cognitively wanted nor unwanted.

Given a user's state transition model $\Pr_{user}(c' \mid c, d)$ where the user tends to give rewards ('showing pleasure') when RoCo takes 'tracking' at his or her 'attentive' state, 'ent2' at 'distracted' state, and 'stretching' at 'slumped' state, we confirmed that RoCo learned quite desirable decision-making models for both ACDs to maximize external rewards. Of course, RoCo made very good world simulation models for both ACDs that approximated the user's state transition model very well, in particular, for state-ACD pairs (c, d) related with the optimal policy. We also found out that the internal simulation loop was helpful in fast learning of the user's desire. With the internal rewards from cognition and emotion, the internal simulation loop utilized the trade-off between exploration and exploitation well.

3 Conclusions and Future Work

In this paper, we proposed a new framework of affective-cognitive learning and decision making where 'internal reward from cognition and emotion' and 'external reward from the external world' serve as motivation in learning and decision making. We also described a hierarchical network of ACDs and an algorithm that implements them. Future work includes more careful proofs of the efficacy of the internal simulation loop and internal rewards, detailed models of various factors in the algorithm, an emotion model for cognitive appraisal of emotional states, a model of curiosity as 'the exploratory drive', a model of empathy, and applications of the algorithm for more complex situations. We expect that the framework should ultimately enable affective agents to have much smoother human-computer interaction.

References

1. `http://web.media.mit.edu/~hiahn/roco.html`
2. Singh, S., Barto, A. G., and Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In Advances in Neural Information Processing 18.
3. Bechara, A., Damasio, H., Tranel, D., and Damasio, A. (1997). Deciding Advantageously Before Knowing the Advantageous Strategy. Science. 275, p. 1293-1295.
4. Berridge, KC. and Robinson, TE. (2003). Parsing reward. Trends in Neurosciences. Vol. 26 No. 9.
5. Damasio, A. (1995). Descartes Error. Emotion, Reason and the Human Brain. Quill.
6. Estrada, C., Isen, A.M. and Young, M.J. (1994). Positive Affect Influences Creative Problem Solving Reported Source of Practice Satisfaction in Physicians. Motivation and Emotion, 18, p. 285-299.
7. Mowrer, OH. (1960). Learning and behavior. John Wiley.
8. Panksepp, J. (1998). Affective Neuroscience. Chap 8 and 9. Oxford university press.
9. Picard, R. W. (1997). Affective Computing. MIT Press, Cambridge, MA.
10. Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, D., Resnick, M., Roy, D.and Strohecker, C. (2004). Affective learning - a manifesto. BT Technology Journal, Vol 22 No 4.
11. Schmidhuber, J., Self-Motivated Development Through Rewards for Predictor Errors / Improvements. Developmental Robotics 2005 AAAI Spring Symposium.