EMOTION-II Model: A Theoretical Framework for Happy Emotion as a Self-Assessment Measure Indicating the Degree-of-Fit (Congruency) between the Expectancy in Subjective and Objective Realities in Autonomous Control Systems

David Tam*

Department of Biological Sciences, University of North Texas, Denton, Texas 76203, USA

Abstract: The second phase of the "*EMOTION-II*" model ("*Emotional Model Of the Theoretical Interpretations Of Neuroprocessing*") introduces the theoretical framework for the evolution of emotion as an internal measure of modeling errors (discrepancy signals) for assessing the degree-of-fit (congruency) between internal model and external world in autonomous control systems. It is derived based on the inevitable real-world consequence that modeling errors often occur in the internal model that represents the external world. When the contextual abstraction of the external world is compared with the internal world model, the discrepancy between the two models (objective reality and subjective reality) serves as a feedback for self-corrective actions. The assessment and recognition of these internally generated signals representing modeling errors of expectancy (and conversely, congruency between the two realities) form the basis for emotion formation in animals and other self-correcting autonomous control systems.

Keywords: Evolution of emotion, autonomous control system, model of emotion, neural processing, modeling error, model congruency, expectancy, subjective reality, objective reality.

INTRODUCTION

To further explore the theoretical frameworks of emotions introduced in the previous paper [1], we will introduce an additional model encapsulating the essence of emotions in autonomous systems called the "*EMOTION-II*" model ("*Emotional Model Of the Theoretical Interpretations Of Neuroprocessing*"). This model is based on the formation of internal model produced by an autonomous system to represent the external world, and how it handles modeling errors in relation to the real world. From this model, emotion emerges as the representation of internal signals indicating the congruency (or discrepancy) between the model expectancy and the reality.

ERROR CORRECTION IN CONTROL SYSTEMS

In control systems engineering, error signals are essential feedback signals to correct for systems errors. In any real-world systems, errors occur either in internal components of the system or introduced by external perturbations. Thus, errors are inevitable in the real world where the system's response is to correct for these errors, wherever the error source may be. Error-correction is one of the crucial components in control systems engineering.

ERROR RECOGNITION

Similarly, for a self-adaptive autonomous control system, error-correction also plays an important role in its operation. Intuitively, the better error-correction ability, the better the system is able to adapt to the environment. Thus, this paper focuses on the role played by error-recognition in defining what is known as emotion.

EMOTIONS AS INDICATORS FOR ASSESSING MODEL ACCURACY

We propose that internally generated error signals in a control system can be used as internal consistency-check for assessing the degree-of-fit between the expectancy of the internal model and external world. The degree-of-congruency between these two models represents the quantity known as emotion.

IMPLICIT INTERNAL MODEL REPRESENTATION

Implicit to any control system is the representation of an internal model of the external world. Even for a simple sensorimotor stimulus-response I/O function, such as reflex discussed in the previous paper [1], there is an implicit internal model represented by the I/O function. The implicit model is essentially the I/O mapping between the input space and the output space. Such mapping may take a form of a look-up table (LUT) for discrete input/output values, or an I/O function for continuous signals. The I/O mapping function may be stochastic [2] or deterministic. The I/O function can be a many-to-many mapping such as the mapping function implemented in most neural networks.

MODEL REPRESENTATION BY I/O FUNCTION

Regardless of the internal representation, the implicit internal model is embedded in the I/O function itself in which given an input, it can predict what the output would be, based on the neural mapping function for control and estimation [3-5].

*Address correspondence to this author at the Department of Biological Sciences, University of North Texas, Denton, Texas 76203, USA;
E-mail: dtam@unt.edu

1874-110X/07

MODEL PREDICTION

Furthermore, implicit in the definition of internal model is the prediction provided by the model itself. That is, given a set of inputs, the model can produce some outputs that predict the corresponding I/O relationship. In the example of a simple I/O function, such as reflex, the I/O function essentially predicts what the reflex output response would be, given a specific stimulus. This is essentially the definition of biological reflexes where a stereotypic response is always invoked, given the same stimulus.

IMPLICIT PREDICTIVE POWER OF REFLEXES

Although technically speaking, a reflex does not model the external world or predict its action based on its input; nonetheless, a reflex does encapsulate these predictive properties and modeling attributes implicitly. In other words, the consequence of its action produces the phenomenon of prediction and modeling. Thus, the predictive property and modeling property of such a simple system can be considered as emergent properties also.

CONTEXTUAL REPRESENTATION BY MODEL

As discussed in the previous paper [1], one of the advantages for creating an internal model of the external world is that it provides a contextual representation of the outside world as well as the prediction of how its future actions may have on the environment. This predictive power of the internal model provides the high-level processing ability for emotions to be evolved for assessing the accuracy of the system.

REALITY IN MODELING EXTERNAL WORLD

We propose that there are two representations of the external world – the "objective reality", which corresponds to the true representation of the world, and the "subjective reality", which corresponds to the internal representation of the external world created by the internal model of the autonomous system.

In any real-world system, errors always exist between the internal model and the external world. These errors can be introduced internally by errors in the internal model or externally by distortion/perturbation of sensory inputs outside the control of the autonomous system.

In either case, the error represents a discrepancy between the internal model and the real world. Conversely, if no error exists, it represents congruency between the internal model and the external world.

DEFINITIONS OF OBJECTIVE REALITY AND SUBJECTIVE REALITY

The contextual abstract representation of the external world is called the "reality". The objective reality represents the true representation of the external world as is. The subjective reality represents the internal representation of the external world through the filters of sensory signals, internal processing (such as signal enhancement) as perceived by the internal model, and the model itself that represents the reconstruction of the external world.

SOURCE OF MODELING ERRORS

The discrepancy between the objective and subjective reality is a major source of error for the internal model. These factors can contribute to the overall error of the internal model that represents the external world.

In addition to modeling errors, internal components are also susceptible to various errors, such as component failures and perturbations from external sources. These errors can contribute to the failure of the model to respond appropriately in the real world. Thus, in order for an organism to survive successfully in the real world, it has to address the error issue.

ERROR CORRECTION BY INTERNAL MODEL

Numerous errors can occur in the reconstruction process to re-represent the external world in which the autonomous system relies on so that it can correct its behavioral outputs. Thus, a system relies on the error signals as source of indicators for generating corrective actions. Most importantly, these error signals are essential for correcting the internal model if discrepancy occurs between the real world and the internal model.

EMOTION AS A MODEL CONGRUENCY MEASURE

We propose that emotions are measures of different congruency (or discrepancy) indicators for assessing the internal model such that self-correction of the internal model can be made. In other words, these internal feedback indicators provide the necessary measures for the internal model to assess its accuracy in predicting how it fits in the environment successfully.

HAPPINESS AS PERFECT PREDICTION

Ideally, in a perfect world with a perfect model, there would be a perfect congruency between the real world and the internal model as predicted. In such instance, an autonomous system would be perfectly happy metaphorically because everything is all right.

UNHAPPINESS AS IMPERFECT PREDICTION

If there were discrepancies between the real world and internal model, then the autonomous system would not be happy by the same token. Although this interpretation anthropomorphizes an autonomous system, it essentially captures the essence of happiness/unhappiness.

CLASSES OF EMOTIONS

Basically, emotion (in this context) can be subdivided into two basic classes: happy and unhappy emotions. They correspond to the assessment of how prefect or imperfect the model predictions are, with respect to the real world.

HAPPY EMOTION

Happy emotions can be subdivided into other subclasses, such as contentment and ecstasy, which correspond to how perfect the predictions are as well as whether they exceed the prediction.

UNHAPPY EMOTION

Unhappy emotions can also be subdivided into many subclasses, such as sad, anger and fear. These emotions correspond to the specific discrepancies between the model prediction and the real world.

DESIRABLE OUTCOME AS MODEL PREDICTION

Note that, "prediction" we refer to, in this context, corresponds to the "desirable outcome" as projected by the system. In other words, it is not merely the accuracy of the modeling function, but the accuracy of the projection of the model in predicting the desirable outcome.

In plain language, the desirable outcome is "what the system wants", and the accuracy of the model prediction is "how accurate it predicts to achieve" rather than "how accurate the modeling process is".

HAPPINESS AS PERFECT AS PREDICTED

There are many subclasses of happiness feeling, of which contentment is an emotion that corresponds to the perfect congruency between the predicted desirable outcome and the actual outcome. Ecstasy, on the other hand, is an emotion that corresponds to exceeding the model prediction.

EXCEEDING EXPECTATION

Exceeding expectation can be considered as a model mismatch, but it can also be considered as a perfect match if the "ideal" desirable end-goal is used as the criterion for matching. Thus, on the one hand, exceeding expectation could mean the model underestimated the prediction when the model could have predicted it more realistically (or more accurately). But, on the other hand, it could also mean that it is a perfect match between the "ideal" desirable outcome and reality when the model goal-state is considered as the criterion rather than the actual prediction of the model as the matching criterion.

ECSTASY AS REACHING IDEAL DESIRABLE GOAL

In the case of ecstasy, there is a presumed "ideal" desirable end-target even though that target may have been considered as unattainable or unrealistic by the model prior to that realization. When that ideal goal is attainable in reality, it becomes a perfect match, thus the feeling of ecstasy.

CONGRUENCY IN DIRECTION OF PREDICTION

More precisely, it is the direction of prediction that also affects the happy or unhappy emotion. A congruent prediction is that it accurately predicts not only the desirable outcome, but also that the prediction is in the right direction. In other words, the model is reaching its ideal target state closer and closer, thus becoming more congruent.

DISAPPOINTMENT AS FALLING SHORT OF TARGET

If the prediction falls short of the desirable outcome, it would become a disappointment (an unhappy emotion) rather than an ecstasy (a happy emotion).

In contrast, when the prediction exceeds the actual projection of the outcome, it is an ecstatic feeling (happy emotion). It is not considered as falling short of expectation, but approaching the expected desirable state. In fact, some people would call it more than perfect rather than imperfect prediction.

FUNCTIONAL DEFINITION OF MODEL PREDICTION ACCURACY

Depending on the definition of prediction used in this context, there are two interpretations of model accuracy. If the strict definition of congruency is used, then the interpretation is that the model fails to predict the modeled outcome accurately (in the case of exceeding expectation for ecstasy). But if the functional definition of prediction congruency is used, then there is congruency between the desirable outcome and the model prediction (for ecstasy).

"WANTED"/"UNWANTED" STATES AS MATCHING CRITERIA

This functional definition of reaching desirable goal state as accurate prediction implies what the system knows what it "wants". Achieving that end-result would contribute to a happy emotion, which is congruent with the model prediction.

By the same token, applying this functional definition of model prediction accuracy based on the desirable goal state as the criterion, then unhappiness is the emotion that represents the deviation from this desirable end-target (when arriving at the unwanted state).

UNHAPPINESS AS UNDESIRABLE STATE

Using this functional definition, if the system predicts what it does not want (undesirable outcome), and is actually happening, then even though the model is an accurate model and congruent in the strict sense, it fails the functional definition of system prediction in achieving what it wants (as the desirable goal).

Therefore, the emotion would still be unhappy (rather than happy) even though it predicts the reality accurately, yet it falls short of the expectation (model prediction). In this case, there is a discrepancy (gap) between the model prediction of desirable result and the actual outcome; therefore, an unhappy emotion in this context.

The specific unhappy emotion (such as angry, sad or fear) depends on the specific context of the discrepancy in the modeled prediction.

UNHAPPINESS AS MISSING THE TARGET (A MISMATCH)

Even though unhappiness can happen when the system perfectly predicts what it does not want (thus an accurate model prediction), yet if desirable target is used as the matching criterion for model prediction, then there is a mismatch. Reaching the unwanted state (no matter how accurate the model is) still misses the target, thus a discrepancy between the reality and the model prediction of what it should be.

MATCHING CRITERIA

Accuracy of model prediction (or expectancy), by this functional definition, requires satisfaction of two criteria:

1. absolute match: a match between the desirable goal state (as modeled) and actual state;
2. relative match: a match in the direction toward (or away from) that desirable goal (a time-derivative measure).

ABSOLUTE MATCH

Assuming there exists an ideal goal, absolute match means that there is a congruency between the projected outcome and the actual outcome. That is, there is an expectancy of what the ideal outcome should be. Whether that outcome is attainable is a different matter.

When reality and the expectancy of the desirable outcome match, it corresponds to the happy state. When that expectancy and reality do not match, it corresponds to the unhappy state.

RELATIVE MATCH

Even in a mismatch situation, if the mismatch is decreasing, it could lead to happiness. If the mismatch is increasing, it could lead to unhappy emotion. Thus, mismatch by itself does not necessarily represent unhappiness; if that mismatch is diminishing, it could still lead to happiness.

An intuitive example is that when something unfortunate (undesirable) event has occurred, but when things are getting better (leading toward desirable goal), it would be happy. When things are getting worse (leading away from desirable goal), it would be unhappy, even if the starting point is a happy state.

DESIRABLE GOALS

The ideal desirable goal may or may not be realistic. It can be a moving target. It is what the model projects as the ultimate state to achieve. In other words, this ideal goal may not be absolute; it can be relative and changeable.

Example of an absolute target is the innate state, such as the satiety state in food reward. Example of a relative target is the relative degree of satiation that is considered as satiated for different individuals.

INNATE GOAL STATE

There also exists innate desirable goal state such that the system will strive to achieve, this innate state usually exists without being overridden by the system.

For example, food satiety state is an innate desirable goal state such that animals will try to attain, which usually cannot be overridden as undesirable (for survival reasons). Nonetheless, the relative satiety state can be adjusted by the system as desirable (i.e., how full will the animal get before it considers that as satiated).

GOAL-DIRECTED SEARCH

The task of the system is to search for this goal such that there is a congruency between the modeled goal and the reality. Conversely, the task is to minimize the error (discrepancy) between the modeled goal and reality.

MODEL PERFORMANCE *VS* MODEL EXPECTANCY

Because errors can come from modeling errors (model inaccuracy and inaccurate perception of reality) and inaccurately modeled goal, the task of the system is to derive measures for self-consistency checking.

This accounts for the difference between perfect model performance (e.g., predicting something accurately but that may not be what it wants to happen) and perfect model expectancy (e.g., predicting what it wants accurately).

DESIRABLE OUTCOME AS CONTEXTUAL CONGRUENCY

The congruency between what the model wants and the actual outcome implies the contextual congruency. That is to say, when the system accurately predicts the environment within context such that it would fit in and operate perfectly in that environment (i.e., reaching the desirable state), then that is truly congruency.

Thus, congruency is a measure of how close it approximates the "ideal" goal that allows the organism to function most appropriately within context. (We will address how this ideal goal is formed in the context of innate response later in the model.)

DICTIONARY DEFINITION OF HAPPINESS

One of the dictionary definitions of happiness is defined as the "feeling satisfied that something is right or has been done right" [6]. Thus, the above description of an indicator measuring the congruency between the internal model and the external world satisfies this common-sense definition of happiness.

Conversely, the definition of unhappiness also satisfies the description of the indicator that assesses the discrepancy between the internal model's expectancy and external model of the real world.

CONGRUENCY WITH EMOTIONAL LABELS

The labels (terminologies) we use to describe the emotional attributes experienced by human also happened to describe the same phenomenon as defined in this context. Thus, regardless of whether these emotional terms are introspective constructs in psychology, they do fit the definition that corresponds to the measures of how accurate an action (or thought) is in functioning within the external world.

OPERATIONAL DEFINITION OF HAPPINESS

Essentially, happiness is a congruency measure used internally for assessing the accuracy of the model's

expectancy. Intuitively speaking, the more congruent the internal image of how it should be in its world, the happier it will be. That is to say, the more perfect the reality is that matches the internal image of how it fits in its world, the happier it usually is, by most objective observers. (We will address whether this emotion is true happiness or pseudo-happiness later in the model.)

Thus, the degree of perfect-fit between the modeled system and the actuality can serve as an indicator for this hypothetical quantity we called "happiness". This emotional indicator can be used as an internal guide for the system to self-correct any discrepancy between the model's expectancy and the reality.

EMOTION AS AN INTERNAL ERROR CHECK

Note that the key distinction in this definition is the *internal* representation of this congruency indicator. Any external representation of the error feedback signal would not be considered as emotion.

For instance, if the measure were represented externally and then presented to the individual, it would merely be an error feedback signal such as the reinforcement signal presented by an external "teacher" (as discussed in the previous paper) [1], a critic or an advice.

Thus, emotions have to be internally generated representing an internal state of a self-correcting autonomous control system (or an animal) in this definition.

HAPPINESS AS A CONGRUENCY BETWEEN SUBJECTIVE AND OBJECTIVE REALITIES

Based on this definition, happiness is defined by the internal representation of the state of degree-of-congruency between the modeled system's expectancy and the actual world. The representation within the modeled system corresponds to the so-called "subjective reality" while the representation of the actual world corresponds to the so-called "objective reality".

The degree-of-fit between these two entities represents a measurable quantity called "happiness". Conversely, the degree-of-mismatch between these two realities represent the measurable quantity called "unhappiness". These emotions form the two basic classes of emotions by definition.

DEGREE-OF-MATCH

We further hypothesize that the intensity of happiness is correlated with the degree-of-match in the models. That is, the intensity of the emotion can be quantified by the degree-of-fit between the internal model and external world (which can also be a model – the exact model). The better the degree-of-fit, the more intense the happy emotion will be, and vice versa, the greater the degree-of-discrepancy, the more intense the unhappy emotion will be.

QUANTIFYING THE INTENSITY OF EMOTION

Emotion, by this definition, is not just a qualitative term to describe an internal state of an autonomous system, but also a quantitative metric for assessing the intensity (magnitude of degree-of-match) of the emotional state. This provides the basic theoretical framework for the quantification of emotions by addressing the basic principles governing modeling-errors with minimal assumptions.

DEDUCTION OF TWO BASIC EMOTIONS – HAPPINESS AND UNHAPPINESS

Based on the proposed definition above, two distinct emotions (happiness and unhappiness) emerge as the emergent properties of a self-correcting autonomous system when the internal and external models are compared. The degree-of-fit between these models (the internal modeled system and the external world model) is a quantifiable measure that corresponds to the common definition of happiness.

CONGRUENCY INDICATOR

The greater the fit (or congruency) the more intense the happy emotion would be. In other words, the intensity of happiness is related to and quantified by the magnitude of degree-of-fit, in our hypothesis.

DISCREPANCY INDICATOR

Conversely, the degree-of-discrepancy between these two models represents the unhappy emotion. A corollary is that the intensity of unhappiness is also related to the magnitude of degree-of-discrepancy.

EMOTION AS A STATE

Based on this definition of emotion within this framework, it corresponds to the internal indicator for assessing the accuracy of the model and its actions. In a broader term, happy emotion can be defined as a "state" of congruency between the two models rather than a specific measure (parameter of the model).

Similarly, unhappy emotion represents a state of discrepancy between the two models, as indicated more specifically by the different subclasses of unhappy emotions, such as sad, fear and anger, to assess where the discrepancy lies.

FUNCTIONAL DEFINITION OF EMOTIONS

Emotions, as derived, are not necessarily unique to humans or animals, nor are they introspective constructs labeled/constructed by human to explain some psychological phenomena. Emotions are merely internal states or internal attributes used in the model to assess the accuracy of the modeled system's expectancy with respect to the true representation of the external world.

INTERNAL MODEL BY NEURAL NETWORKS

As reviewed in the previous paper [1], a generalized neural network essentially computes a nonlinear many-to-many I/O function mapping the input space into the output space. It is essentially a statistical model sampling the input parameter space by iterative search for a solution space for the system [7]. The generalized probabilistic I/O mapping function, $p[\cdot]$, for an individual neuron can be represented by the matrix equation:

$$\vec{Y}(t) = p\left[\vec{W}(t), \vec{X}(t)\right] \quad (1)$$

where $\vec{X}(t)$, $\vec{Y}(t)$ and $\vec{W}(t)$ are the input, output and connection weight matrices, respectively, although most often $\vec{X}(t)$ and $\vec{Y}(t)$ are vectors (one-dimensional matrices). Since the inputs change in time when the autonomous system is moving in time, the quantities above are functions of time, t, too.

NETWORK EQUATIONS

The I/O function of a generalized neuron takes on the weighted-sum function, thus expanding Eq. 1 gives:

$$y_j(t) = p\left[\sum_{i=1}^{n} w_{ij}(t)x_i(t)\right] \tag{2}$$

For a neuron located in the k-th layer, Eq. 2 can be re-written as:

$$y_j^k(t) = p\left[\sum_{i} w_{ij}^k(t)x_i^k(t)\right] \tag{3}$$

For a network of neurons with multiple layers, the I/O function of the network is given by:

$$\begin{aligned}\vec{Y}(t) &= \prod_k p\left[\vec{W}^k(t), \vec{X}^k(t)\right] \\ &= p\left[\vec{W}^k(t), \vec{X}^k(t)\right]\cdots p\left[\vec{W}^1(t), \vec{X}^1(t)\right]\end{aligned} \tag{4}$$

for a network with k layers. For a nonlinear network (with different nonlinear functions, $p^k[\cdot]$) at layer k, the output of the network is given by:

$$y_j^k(t) = \prod_k p^k\left(\sum_i w_{ij}^k(t)x_i^k(t)\right) \tag{5}$$

INTERNAL MODEL BY MAPPING FUNCTIONS

Thus, the I/O function of the entire network is a many-to-many mapping function, mapping from the inputs, $x_i^1(t)$, in the first layer to the outputs, $y_j^k(t)$, in the last (output) layer. Since the individual I/O functions for each neuron are nonlinear, the layers are not collapsible into a single layer for a typical network. Nonetheless, the multi-layered neural network described by Eq. 5 is a complex many-to-many I/O function, which essentially encapsulates an internal model of the input-output relationship for the network.

IMPLICIT MODEL BY I/O FUNCTIONS

This input-output relationship (represented by the I/O function) is an implicit model of the external environment if it has acquired the mapping (i.e., learned) such that given a specific set of inputs, it will respond with a set of outputs that is appropriate for the situation (similar to a reflex action discussed in the previous paper [1]). If the input-output relationship is appropriate for the circumstances, then it can be said that the system has acquired an implicit model of the environment in which it can act on appropriately.

REFLEX AS A DYNAMAICAL MODEL

This implicit internal model is a dynamical model rather than a static model. An example of a static model is a static map of the external world. But, in this case, the model is dynamically generated with respect to the sensory inputs and output actions using the many-to-many I/O mapping function that is also a function of time. That is, the response of the system is time-dependent, which is dynamic rather than static (time-independent).

STRETCH REFLEX EXAMPLE

For the sake of discussion, we will use a physiological reflex as an example to illustrate the implicit dynamic model encapsulated by a simple reflex, such as the knee-jerk reflex (which is also called "stretch reflex"). The stretch reflex is a simple reflex with a well-known neurobiological neural circuitry in the spinal cord to maintain the upright postural position of the limbs [8]. In the legs, stretch reflexes are activated at the joints (such as hip, knee and ankle) to maintain upright posture.

REFLEX ARC CIRCUITRY

Stretch reflex in the knee is activated when the extensor muscle tendon is stretched, stimulating the monosynaptic reflex arc that triggers (activates) the contraction of the extensor muscle, causing the leg to extend (kick forward). At the same time, the same stimulus also activates the disynaptic reflex arc, which inhibits the flexor muscle *via* an inhibitory neuron (Renshaw cell) in the spinal cord. The inhibition of flexor muscle and the excitation of the extensor muscle together form the reflex commonly known as knee-jerk reflex (which is an example of the more generalized stretch reflex).

FUNCTIONAL READOUT OF REFLEX MODEL

This same stretch reflex is operating in the knee as well as in the Achilles tendon and other limb-joints whereby balancing is accomplished by maintaining an upright posture (especially in bipeds). The stretch reflex allows bipeds (such as humans) to stand upright even in the presence of external perturbations to the upright posture. Any gravitational perturbations that cause deviation from the centered upright position of the legs will trigger the stretch reflex to return the postural position back to a balanced position.

DYNAMICAL MODEL OF REFLEX

The stretch reflex essentially encapsulated an implicit internal dynamical model of how to maintain balance in an upright position. This model is not a static model, nor a static map, but a dynamical model that depends on the input stimuli such that it will respond appropriately based on the implicit dynamical model to maintain balance.

BEHAVIORAL READOUT OF MODEL PREDICTION

This illustrates the abstract model encapsulated by a neural network circuitry (in the stretch reflex case, encapsulated by a simple mono-synaptic and disynaptic neural circuitry in

the spinal cord) is a dynamical model of the external world (i.e., the physics model for maintaining upright posture for the legs controlled by the stretch reflex).

This model can be read out by the behavior involved in the interaction between the organism and the external world (in the stretch reflex case, it is read out by introducing perturbations to the upright posture), even though there is no explicit model of this interaction or mapping of the external world existed in the neural circuitry.

EMERGENT OF INTERNAL MODEL

What this means is that the internal model is another example of the "emergent property" of a neural network or a many-to-many nonlinear dynamical probabilistic mapping I/O function. Thus, we can safely assume that an internal model can exist (emerge) from the dynamical interactions with the environment if such network established (acquired or learned) the appropriate response functions such that it approximates the real-world physics model, physiological model or psychological model. These models can be behavioral model, cognitive model or emotional model in the abstract sense so long as they approximate the real-world functions.

MODELING ERRORS AND FAULTS

As discussed above, errors exist in any real-world models. The source of errors can come from input (sensory) errors, output (motor) errors, internal (modeling) errors and external (perturbation) errors. All these errors together can contribute to the inaccuracy of the final modeled predictions. These inaccuracies can lead to fault-conditions, failures, or inappropriate actions. Thus, for an autonomous system to operate effectively in the real world, these inevitable errors have to be taken into account in the model to produce corrective actions if such system were self-adaptive and autonomous (i.e., without the corrective actions being imposed or introduced from external sources).

EXCLUSION OF EXTERNAL ERROR MEASURES

Based on the above definition of emotion, any indicators for assessing the errors of the system can provide useful feedback for the internal model to self-adapt without any external guide. Error signals have to be derived from its own internal system if it were to self-correct without external guide. Thus, we exclude the possibility of externally generated error signals in our derivation of an emotional model with a minimal set of assumptions.

SELF-DERIVED ERROR SIGNALS

For a simple feedback control system, error-correction can be performed by a simple feedback loop with adjustable gain using the pre-determined error signal for adaptation. In contrast, for an autonomous system, the source of error may not be known in advance.

Thus, the system needs to adapt a different strategy in handling errors for self-corrective autonomous control. In other words, it has to derive its own error signals if the source of modeling error (or error of the model itself) were to identified for self-correction without external cues.

Emotion, in our proposed definition, serves the functional role in the auto-corrective paradigm for correcting any unforeseen errors in the internal model or the modeled expectation/prediction.

ERROR SIGNALS IN NEURAL NETWORKS

In neural network architecture and design, there are many different types of neural networks that use error-correction as its learning rule for adjusting their connection weights. The most well known neural network using error-correction signal for learning is the error back-propagation model [9]. It is a feedforward network that minimizes the error function using a gradient-decent method, and the error signal is provided by the external "teacher" to adjust its internal connection weights. It is essentially a "hand-holding" method where the error between desired output and the actual output of the system is used as the signal for correction, by minimizing the error.

EXCLUSION OF SUPERVISED LEARNING

Since the error-backpropagation learning algorithm belongs to the class of supervised learning that requires *a priori* knowledge of the "desired" output to generate the explicit corrective-error signals for the network to correct, it is not self-corrective, nor is it is helpful for autonomous systems where the end-state or desired output may not be known in advance. Thus, this class of supervised learning network architecture is not suited for auto-correction of modeling errors, which, by definition, not a likely candidate network model for emotion formation in this case.

ASSOCIATIVE REINFORCEMENT LEARNING WITHOUT EXPLICIT ERROR SIGNAL

In contrast, the associative reinforcement-learning model as discussed in the previous paper [1], the error signals are not explicitly presented in the system; rather they are deduced implicitly from its own actions. Rather than using the explicit error (discrepancy) signal between the desired and actual output for comparison to correct its action as in back-propagation neural network, associative reinforcement network uses the reinforcer signal for adjusting its connection weight, i.e., no explicit error signal is used for the self-adaptation. No external error-correction signal is presented or introduced into the system.

Thus, the reinforcer essentially serves as a guide to the direction (either positive or negative) of internal adaptation such that the network will adjust (and correct) itself accordingly. Therefore, the associative reinforcement learning network model is well suited for emotion formation in the current definition of emotion to assess the accuracy (or congruency) of the modeled system with respect to the real world.

PREDICTION WITHOUT *A PRIORI* KNOWLEDGE

In an associative reinforcement-learning network, the error signal is not explicitly presented or known in advance. In other words, the system does not have any *a priori* knowledge of the outcome of the system nor the external world itself. That is, it does not have a "model" of the external world prior to the acquisition phase of learning. The internal

model created is a consequence of the auto-association between input and output using (positive or negative) reinforcement as the cue to adapt its internal connection weights to produce its output. The error signals, if exist, are derived internally.

SELF-DERIVED PREDICTION

The ability to arrive at prediction (or arrive at a solution state) using reinforcement signal without any explicit error signal is essential to the emotion formation based on its own auto-corrective actions using self-derived error signals. Thus, this accomplishes two principles:

- auto-prediction without explicit end-targets, and
- auto-correction without explicit error-signals

in emotion formation.

INTERNALLY GENERATED CONSISTENCY-CHECK AND ERROR SIGNALS

The above model conforms to the proposed definition of emotion that error (discrepancy) signal is internally generated from within the model itself. If error signals are generated or introduced from external sources, it cannot be regarded as relating to emotion.

Although this is intuitively obvious that external congruency indicators cannot be regarded as emotions or relating to emotions, nonetheless, it is an important distinction as far as autonomous control system is concerned because error signals in most control systems are derived from external sources rather than internally generated. Thus, only internally generated error signal for congruency measure of modeling errors can be considered to fit our proposed definition related to emotion.

MODEL PREDICTION WITHOUT EXPLICIT MODEL

Implicit in the characteristic of any model in general is the predictive property produced by the model. By definition, a model is an abstract representation of the actual phenomenon with the ability to predict the outcome of the phenomenon even with insufficient data or knowledge of the predicted (or modeled) system.

As discussed earlier, a model can be a dynamical model, such as the equations governing a physical process, even though the model does not necessarily have any physical static representation or corresponding physical map of the phenomenon. So long as the equations (in the above example) can predict the outcome of the phenomenon it describes, it can be considered as a model.

DYNAMICAL PREDICTION

By the same token, the internal model of an autonomous system can be a dynamical model so long as its equations governing the neural connectivity, architecture and learning rules (in our example) can predict how the outcome of its actions can interact with the environment it lives in (i.e., how it responds to the environment conditions in an appropriate way). Thus, if the output of an autonomous system can produce an action that approximates an appropriate response to the actual phenomenon in the real world without external instructions, it is essentially providing a prediction of the modeled process.

PREDICTION WITHOUT COGNITION

Note that the prediction we refer to does not need to have any cognitive connotation that the system is aware of this prediction, making conscious prediction or making explicit prediction; in just the same way that a set of equations can predict a physical phenomenon, it does not imply that the equations have a mind of its own in making those predictions.

PREDICTION WITH INCOMPLETE DATA

Model prediction is an important characteristic in an abstract model whereby the outcome of the modeled process is predictable by the model, which approximates the real phenomenon. Another important characteristic of a modeled system is that it can produce its response action such that even in a non-ideal condition of sufficient data (or missing data), it can produce a fair approximation of the response that fits the actual scenario. In other words, it can fill in the missing information to provide a fairly good prediction of the phenomenon it models.

CONTEXTUAL PREDICTION

Contextual prediction is one of the attributes of the modeling process that becomes important in autonomous system, since if the prediction fails to approximate the appropriate interactions with the real world, the autonomous system would fail to function as a functioning system. The greater the discrepancy between the modeled output and the actual appropriate presumed response, the greater the inefficiency of the system.

DESIRABLE OUTCOME PREDICTION

Thus, the ideal goal of the autonomous system (if such desired goal exists) is to produce the action that best predicts the appropriate interaction with the environment. In other words, if there is congruency between the modeled response and the actual scenario, the autonomous system is better prepared for the environment. In biological systems, if the organism can produce actions that are congruent with the external world, the better chance of survival will be.

REWARD EXPECTANCE AS MODEL PREDICTION

There are many forms of prediction in modeling the external world. One of the predictions is the expectation of the reward signal called reward expectancy. In the reinforcement-conditioning paradigm discussed in the previous paper [1], the resulting reinforced behavior is a form of prediction in which a reward is expected upon the presentation of the conditioned stimulus (CS) signal. That is, it implicitly modeled the predictive nature of the dynamics between the stimulus and response function.

REWARD EXPECTANCY BY NUCLEUS ACCUMBENS

Neurobiologically, there is evidence showing that nucleus accumbens is not only identified as the site of reward signal activation in the brain, but also as the site of predicting the reward magnitude [10] and predicting the time interval of the reward signal onset [11]. In fact, the neural activity of the nucleus accumbens is more correlated with the expectation of the reward, i.e., craving, than the actual reward, i.e., "rush" pleasurable sensation in fMRI studies of activation of emotions elicited by cocaine stimulus [12].

SUPERSTITIOUS BEHAVIOR AS MISGUIDED PREDICTION

If the system acquired the prediction such that the presentation of the CS signal leads to subsequent reward, then the internal model accurately predicts the reality. If the system has acquired a wrong signal as the CS signal (such as superstitious behavior), which is not always followed by a reward, then the internal model erred in predicting the outcome correctly.

SUPERSTITION AS ERRONEOUS MODEL PREDICTION

"Superstitious behavior" is a classic example of malformed learned behavior in which an animal acquired a pseudo-CS signal as the CS signal for the expectance of reward signal. For instance, in a Pavlovian dog example, a dog will learn by operant conditioning to acquire the cues that lead to petting by the dog-owner *via* the pairing between the CS signal and the petting reward. If it happened that the dog got excited and dance in a circle before getting petted, the dog will learn to associate dancing in a circle with being petted.

Because the dance is reinforced by a reward, even though it is unintentional, it will establish the association. As a result, the dog will dance in a circle (as the superstitious behavior) every time it wants to get petted. In this case, the internal model erred in accurately predicting the outcome (the reward) by dancing because the dog-dance is unrelated to the petting behavior by the owner.

This acquired behavior is a superstitious behavior that does not always lead to the prediction of the presumed outcome in reality. Thus, could lead to unhappiness (and frustration) when the misguided expectation fails to materialize in reality.

MECHANISMS FOR SUPERSTITION FORMATION

Therefore, even though the conditioning paradigm can establish association between the CS and CR (conditioned response), that association may or may not be an accurate prediction of the outcome in reality. Because conditioning is merely making association of between environment and the reinforcer, not all the environmental cues are relevant to the delivery of the reinforcer. Some environmental cues are neutral or unrelated to the presentation of the reinforcer. If these environmental cues are used as the CS signal to establish the CR, then the model acquired an inaccurate predictor of the behavioral output, which often leads to superstitious behavior.

EXPECTANCY AS A BELIEF SYSTEM

Thus, reinforcement conditioning will establish any association between the environmental cues and the reinforcer, even though the correlation may or may not lead to a causal-relationship between the environmental stimulus and the prediction of the reinforcer. Nonetheless, the established correlation creates what is called the "expectancy" signal within the modeling system that attempts to predict the outcome of a phenomenon, even though the prediction may or may not become a reality (as in superstitious behavior).

FORMATION OF BELIEF SYSTEM

In fact, this expectancy often forms the so-called "belief system" in human, when we expect the outcome to be true regardless of whether there is any causal relationship between the expectancy and the outcome in reality. The belief system is established because there is a correlation relationship between the two associated events, but because correlation does not always imply causality, that prediction often fails since the correlation is established by coincidence, but not by causality.

ASSESSING BELIEF SYSTEM

Thus, in order for an autonomous system to form a correct model of the external world, it needs to assess the correct CS signal for CR in establishing the association. Failure to establish the correct CS signal will often lead to superstitious behavior or false belief system in the internal model that does not always allow the system to function appropriately in the real world.

INTERNAL CONSISTENCE CHECK

It is essential for an autonomous system to assess its internal model such that it verifies the accuracy of the prediction to confirm if the prediction matches the actual outcome. That is to say, there needs to be some internal consistence check in which the subjective reality is assessed against the objective reality for validation.

If there is a discrepancy between the two realities, then it generates an internal flag as an indicator to locate the source of the error. This internal consistence check as a flag that indicates the state of consistence within the model is what we propose as the state of emotion within an autonomous neural system. This emotional state is an indicator for internal consistence of the model.

EMOTIONAL STATES

If the internal states are consistent, it represents the state of happiness. If the internal states are inconsistent, then it sets the flag for the state of unhappiness. If error or inconsistence exists, the source of error or the source of inconsistence is yet to be determined at this stage. Other subclass of unhappy emotions, such sad, anger and fear, will serve to identify the specific cause (or source) of inconsistence or error signals.

BRAIN REGIONS FOR REWARD EXPECTANCY

Neurobiologically, there is evidence showing different parts of the brain are involved in the prediction and expectancy of outcome in human. In particular, the midbrain dopamine system (which includes the nucleus accumbens) is ascribed roles in reward expectancy, error detection, prediction, and memory [13]. The nucleus accumbens, the sublenticular extended amygdala (SLEA) and orbital gyrus are shown to be involved in the anticipation or expectancy of monetary gains or losses in fMRI studies [14]. Thus, there are evidence for the neurobiological basis for neural processing that are related to the expectancy of the outcome of the model within the brain, which is congruent with our hypothesis.

MODEL CONGRUENCY WITHOUT EMOTIONAL ASSESSMENT

Note, as explained before, the congruency between the modeled process and actual phenomenon may not necessarily imply that the organism is consciously predict its action or aware of the external world consciously. Using the example illustrated before, stretch reflex provides a fairly good predictor of the righting-reflex process by taking the physics of muscle contraction and gravity into account using a simple neural circuitry to implement its feedback control equations.

Nonetheless, this modeled behavior (or response) is achieved without any implied cognitive attributes of what the reflex is attempting to accomplish, and without any emotional attributes even when the prediction (the reflex) fails. This is because no internally derived error signals or congruency measures are incorporated into the internal model as potential guides to correct the system response autonomously.

INDICATORS FOR MODEL ASSESSMENT

Whereas, if the internal model generates an internal state such that the congruency of the modeled response can be evaluated and assessed as an internal feedback for future corrective action, then it would allow the autonomous system to self-correct, self-adjust, and update the internal model autonomously.

If update of the internal model can be done by the assessment of the internally generated congruency measures without any external instruction (or control), then the autonomous system is self-correcting its internal model predictions based on the environment context.

INNATE RESPONSE FORMATION

The feedforward model described in the previous paper [1] is essentially producing a prediction of the outcome of the system. Using reflex as an example, it is a feedforward model whereby the system responds with a feedforward prediction of the projected response outcome when it is confronted with a given set of stimuli. The innate response in a reflex often has gone through the evolutionary survival fitness-test such that the feedforward prediction is often fairly accurate in most circumstances under simplified ideal conditions. The prediction is essentially the dynamical model of the equations governing the physiological reflex.

HARDWIRING OF RESPONSES BY FEEDFORWARD PREDICTION

When simple reflex (implemented by simple hardwired neural circuitry) can produce rather robust prediction of the system's response behavior, a nonlinear system with multi-layered neural network architecture can produce similar dynamical predictions of the system response function when the innate response is established through similar evolutionary survival-of-the-fittest paradigm. As derived from the previous paper [1], the connectivity between neurons in the $k^{\text{-th}}$ layer are consolidated (hardwired) into a feedforward system such that the connection weights are given by:

$$\Delta w_{ij}^{k}(t) = l(t) \cdot r \cdot x_{m}^{k}(t) \cdot y_{j}^{k}(t) \cdot x_{i}^{k}(t) \qquad \forall i, j, k, m \tag{6}$$

for

$$l(t) = p'\left[\left|\Delta \overline{w}\right|_{t}\right] \tag{7}$$

and

$$\left|\Delta \overline{w}\right|_{t} = \frac{\sum_{q=0}^{s-1}\left|\Delta w_{ij}^{k}(t - q\Delta t)\right|}{s} \qquad \forall t \tag{8}$$

where $\Delta w_{ij}^{k}(t)$ is the weight-change between the $i^{\text{-th}}$ input and $j^{\text{-th}}$ output at the $k^{\text{-th}}$ layer at time t, $l(t)$ is the learning-rate, which is dependent on a function of the moving-average of the weight-change, $\left|\Delta \overline{w}\right|_{t}$, and s is the moving-average time-window.

ELIMINATION OF INCORRECT PREDICTIONS

As discussed before, the learning-rate decreases as the system stabilizes to a candidate solution, and the weight-change in Eq. 6 approaches zero; thus the circuitry is consolidated automatically, and become hardwired without being modified over time. This process essentially establishes the feedforward prediction of the outcome of the system through the establishment of innate hardwired response. Whether the projected, feedforward solution predicted by the system is a local minimum or global minimum will be "weeded out" by the evolutionary survival-of-the-fitness process.

EXPECTATION AS PREDICTION OF PREDICTIONS

Given that the above neural mechanism can establish innate response such as reflex action (which is a modeled prediction of the reflex process), a higher-level cascaded prediction can be formed. That is, prediction of the prediction can be accomplished by cascading its actions at multiple neural network levels.

CASCADING META-NETWORKS AND META-MODELS

A meta-network can be formed from incorporating many subsets of networks such that a meta-model emerges by encapsulating its component models. Thus, the formation of

meta-networks and meta-models can provide the theoretical framework for prediction of predictions (i.e., prediction of the accuracy of the sub-systems).

PREDICTION OF PREDICTIONS IN EMOTIONS

In a self-corrective autonomous system, prediction of the accuracy of the expected systems response forms the basis for emotion formation. In other words, the expectation of the modeled response to be accurately produced such that it corresponds to the actual response in actuality in the real-world, then this forms the foundation for emotion formation.

EFFERENT COPY AS A MECHANISM FOR COMPARING MODELED AND ACTUAL OUTPUTS

Once the model predictions have established, the system can produce "expectation" of its output response *via* axon-collaterals (branches of its axonal output) called "efferent copies" in neurobiology. Efferents are the axons of the output neurons, and efferent copies are the branches of same output fed back into the system for comparison of the accuracy of the modeled and actual outputs [15-17]. Thus, many examples of comparison between modeled and actual neural outputs *via* efferent copies are found in biological neural systems, although most of these examples are found in motor systems such as cerebellum, brainstem and spinal cord.

EXPECTANCY OF DERSIRABLE GOAL

Generalizing the concept of efferent copy in neural network, the branching of output for comparison can be considered as introducing the concept of expectation by separating out an efferent copy of prediction for expectation. Although expectation and prediction are synonymous, expectation implies the anticipation of the predicted results, whereas prediction is merely the instantiation of the forthcoming result.

More specifically, expectancy implies the targeting of an end-goal whereas prediction simply projects the response outcome without necessarily having any desired end-goal or target. Thus, expectancy implies the existence of a desired end-goal to be targeted, whereas prediction implies the execution of the command signal to produce the system's output.

EXPECTANCY OF IDEAL TARGET

More precisely, prediction is a modeled response generated by the dynamical equations of the system, whereas expectancy is the set-point of the system for the dynamical equations to follow. Efferent copy allows the system to compare its modeled response (prediction) with the actual output to assess whether it achieves the expected output (end-goal) in actuality.

SEPARATING OUT ERROR SOURCES

If there is a discrepancy between the expected and actual outputs, then the system is in error. As discussed earlier, the source of error may come from modeling error (i.e., imprecise model), model expectancy error (i.e., incorrect end-goal), system output error (i.e., production error), sensory error (i.e., perceptual error), or external perturbation error (i.e., unforeseen circumstance outside the control of the system). These errors, as a whole, contribute to the discrepancy between the expected and actual outputs.

ERROR MINIMIZATION

For a self-adaptive autonomous system, minimization of the discrepancy between the expected and actual output can lead to a more accurate model of the external world. In order to minimize the discrepancy between the internal model and the external world, error signals can be used to quantify the incongruence.

EXCLUSION OF EXTERNAL ERROR CORRECTION

If the error signals were provided by external sources, the system is not auto-corrective autonomously, thus this framework does not fall into the definition for emotion. On the other hand, if the error signals were internally derived, and internally generated, then these measures can serve as candidate emotional components of the system.

ABSOLUTE DIFFERENCE IN ERROR

As discussed before, two different sets of discrepancies exist between the modeled world and external world in terms of emotional responses. One set of discrepancies corresponds to the absolute difference while the other set corresponds to the relative difference between the modeled and external worlds. That is, the absolute difference the modeled and external world assumes a true objective accurate representation of the real world is compared with the internal model.

RELATIVE DIFFERENCE IN ERROR

On the other hand, such true objective model of the real world may not be known or accessible to the autonomous system. In presence of missing or incomplete information about the real world, the autonomous system can only rely on incomplete information for its internal model formation. One of the strategies is to form a tentative desirable goal as the target in face of incomplete information. Thus, relative difference between the true desire and the faulty desire exists as a result.

For instance, accumulating money may be a tentative desirable goal for the system to achieve when no other better desirable goals are available. This may be a faulty desire if other more innate desirable goals are found, in which case, most people often discover that acquiring money does not necessarily lead them into true happiness.

PSEUDO-CONGRUENCY FOR RELATIVE ERROR

In the minimization process of discrepancies between the two worlds, if the external world model as perceived by the model does not correspond to the actual world in reality, the autonomous system can bring the presumed external world and internal world into congruency, but still fails to produce an accurate prediction of the ideal target for the system to achieve.

PSEUDO-HAPPINESS: FALSE SENSE OF EMOTION

In terms of emotions, if the congruency measures are based on relative difference rather than absolute difference between the modeled and external real world, then a false

sense of happiness can be produced. The relative difference is based on the subjective model (or the subjective reality) of the external world when there is incomplete or inaccurate information about the real world.

SOURCE OF EMOTIONAL BIAS

The incomplete (or inaccurate) information may be due to sensory distortion (faulty sensory input), encoding errors (faulty system), or errors in internal model (faulty model). These errors contribute to the subjective model (subjective reality) for the system to respond to, creating the phenomenon of false sense of happiness – a form of emotional bias. Thus, true happiness depends on the accurate assessment of the true objective reality by comparing the absolute differences.

OPTIMISM AS EMOTIONAL BIAS

Optimism is a form of false sense of happiness in which the prediction of the future is much more positive than the reality. The prediction tends to be an overestimation of the future. This bias is called "optimism bias" in psychology. On the other hand, pessimism is a form of false sense of unhappiness when the prediction of the future is much more negative than the reality. These predictions tend to be overestimating or underestimating the outcomes of the future events. Because of the difference in the relative difference between the objective and subjective realities, it can lead to incongruence of the model expectancy.

NEUROBIOLOGICAL CORRELATES

There is neurobiological evidence supporting this hypothesis in the overestimation or underestimation of expectancy of the future in fMRI studies [18]. It was reported recently that the rostral anterior cingulate cortex is correlated with the optimism bias traits, while depressed patients showed a decrease in activity in the rostral anterior cingulate cortex when they projected pessimistic view of future events.

INNATE IDEAL DESIRABLE GOAL FORMATION

In absence of an external objective assessor to create an accurate model of the real world presenting to the autonomous system to evaluate its internal model, it will rely on its best practice. The most reliable information of best practice that stands the test of time may come from the innate source since it survived the evolutionary survival-of-fittest test. Thus, the innate response can serve as the next-to-the-best beacon for congruency comparison.

This innate response is, in fact, formed automatically (as discussed in the preceding paper [1]) without any *a priori* assumption about the fitness of the response or *ad hoc* selection criteria for comparison in the congruency test.

REFLEX AS INNATE DESIRABLE RESPONSE

Because once the innate response is formed, the subsystem is hardwired to produce a stereotypical response (such as a reflex). This implies that a fixed predictable response of the subsystem is always generated as a result. This leads to the predictable expected outcome (or expectancy) of the system in the meta-network circuitry that may be considered as the end-goal or end-target for the innate response. Thus, an autonomous system that developed an innate response will always have a highly predictable response that can be encapsulated by the meta-system as the expected state.

PLEASANT SENSATION AS INNATE IDEAL FEEL

In terms of emotion formation, based on the contextual emotional feel in sensation derived in the preceding paper [1], the innate emotional feel of pleasant (or unpleasant) sensation can serve as the fixed end-target for the system's response. Thus, the innate response becomes the quasi-absolute target (desirable goal) for comparison in the congruency test.

INNATE (ABSOLUTE) DESIRABLE GOAL

What this means, in common sense term, is that there is an innate state of happiness where individuals will thrive to attain. This corresponds to the absolute difference comparison in the congruency measure for happy emotion, which most people refer to as true happiness. Ecstasy is achieved when the actual outcome reaches the innate (absolute) target even though it is unexpected as modeled by the system as unattainable prior to the realization. This is the "dream comes true" phenomenon where the dream is the ideal desirable goal.

RELATIVE DESIRABLE GOAL AS MOVING TARGET

On the other hand, relative difference comparison in the congruency measure does exist, which leads to the commonly known notion of false sense of happiness. Since relative difference is made, it becomes a moving target; thus leading to the common phenomenon of endless chase of pseudo-happiness metaphorically speaking.

ROBUSTNESS OF EMOTION

Because there exists an innate absolute target as the desirable goal state (such as food satiety state), one cannot merely pretend that he/she does not need food, and be extremely happy when food is available. That is, one cannot simply change the absolute goal (lower the standard) in order to experience ecstatic feeling when the absolute state is reached. As an example, a person cannot fool himself/herself by pretending that he/she does not need food (or money), and then be ecstatically happy when food (or money) is available.

In other words, there is an absolute standard of desirable goal state from the innate source that is built into the system that the system cannot be artificially inflating its emotions for the sake of experiencing it, by this definition,

CONCLUSION

We derived a set of conditions under which an autonomous system needs to deal with in order to survive successfully in an environment within context. This set of conditions includes the ability to assess its own internal model for accuracy of prediction of the expected outcomes. If the expected outcomes are congruent with the real world within context, that prediction of ideal desirable goal is congruent with real-

ity. This corresponds to model accuracy and congruency with reality.

Emotions, in this context, are the internal measures for assessing the internal consistency check of the model predictions (or more accurately, model expectation or model desire). If the expectancy is consistent with reality, and is desirable based on the innate desired goal as a guide, then the emotion corresponds to happiness. Conversely, if the expected outcome is inconsistent with the reality, and is undesirable, then the unhappiness emotion serves as a guide for the system to correct its model, and minimize its subsequent modeling errors, if happiness were to be achieved.

There are also relative and absolute differences between the projected outcomes and desirable outcomes, which leads to the differences between subjective reality and objective reality. These differences can contribute to the difference between true happiness and pseudo-happiness. The absolute difference is also dependent on the existence of an innate ideal goal state. The innate ideal desirable goal is established by pre-existing conditions passed on by previous generations that filtered out most of the "undesirable" states by the survival-of-fittest test. Thus, these innate states can serve as the *de facto* standard of desirable states for the system to achieve in the self-consistency check for modeling errors.

Happiness emotion is the emergent state in which the internal model is self-consistent and congruent with the reality within the context for best survival. From this, we derived the origin of emotions without relying on subjective introspection or retrospection of human perception of what emotions are philosophically and anthropologically.

Thus, emotion is derived based on the operational definition for a self-consistency check for the internal congruency measure of its model expectancy of the desirable goal state within the context of the environment in reality. In other words, it is an error indicator for model correction, without which self-correction of the internal model may not be accomplished. Thus, emotion, in this context, is a necessary attribute for a self-correcting autonomous system to operate successfully in the real world.

Thus, emotion can exist for autonomous systems that are capable of assessing its own model expectation and self-correcting it without pre-programmed external guides. This is consistent with the fact that autonomous robots mimicking human emotions do not necessarily possess true emotions per se (according to this definition).

Similarly, animals that are not capable of evaluating its own expected prediction of the real world or correcting its modeling errors in perception may not possess emotion (by this current definition) even though they may possess reflexes that enable them to respond to aversive or non-aversive stimuli. Thus, a cockroach that can escape from predators by means of reflex action within context may not be endowed with true emotion of fear, or that it experiences happiness when it finds food, unless it can assess and correct its model errors and expectancy internally by self-consistency-check.

In other words, it is the self-assessment, self-recognition and self-correction of the modeling discrepancy with the reality within context for survival that defines emotions in the *EMOTION-II* model. Whereas in the *EMOTION-I* model, the contextual abstraction of sensory inputs within the context of the environment is what defines the sensational feel (the emotional sense as opposed to physical sense), a pre-processing function for the *EMOTION-II* model.

With this definition of emotion, and applying this model of emotions, we could better evaluate our emotions for productive functions so that it allows us to pinpoint the source of discrepancy between our expectation and the reality, and correct the modeling errors (which include our unrealistic assumptions about the real world, our subjective perception errors or other sources of errors).

Experimental validation of this hypothesis of models of emotion will be provided in subsequent papers. Derivation of other subclasses of emotions (sad, angry and fear) will be given in subsequent papers.

SUMMARY

An emotional model is proposed to account for the emergence and subsequent formation of emotion based on the internally generated error signals for assessing the congruency between the modeled world and external world with a minimal set of assumption or any *a priori* knowledge of the desired end-goal state. It is a self-bootstrapped model for self-adaptive auto-correcting autonomous system without any external guide or externally presented error signals for error-correction, self-organization or self-learning to produce the congruency measure for emotion formation. Based on the proposed definition of emotion, happy emotion is a congruency measure between the modeled world and the real world, and unhappy emotion is a discrepancy measure between the modeled and real worlds accordingly.

The origin of innate response and innate target-goal state for congruency comparison in emotion formation is derived based on an associative reinforcement-learning neural network model combined with the evolutionary survival fitness-test. The resulting autonomous system is a meta-network formed from many subnets, each contributing to the meta-model for the formation of expectation in emotion whereby prediction of the predictions can be assessed by efferent copies for comparison.

The innate target-goal state serves as the best-practice, quasi-absolute target for comparison between the objective and subjective realities. Comparison with the absolute difference leads to phenomenon called true happiness while comparison with the relative difference leads to a false sense of happiness (pseudo-happiness) in common sense term.

Emotion, by this definition, can form in autonomous systems that can generate internal error signals for self-consistency check and assess the congruency between the modeled expectation and external reality without external guidance. Similarly, robots and other species of animals can possess true emotions if they can self-assess and self-correct their internal model prediction errors and expectancy by

comparing the congruency between objective and subjective realities.

Thus, emotion can be derived based on first principles with minimal assumptions, without any *a priori* assumptions about the biological/psychological phenomenon that may not necessarily be unique to human, animal kingdom or robots.

REFERENCES

[1] D. Tam, "EMOTION-I Model: A Biologically-Based Theoretical Framework for Deriving Emotional Context of Sensation in Autonomous Control Systems", *The Open Cybernetics & Systemics Journal,* vol. 1, pp. 28-46, Dec 2007. [Online] Available: hrttp://bentham.org/open/tocsj/.

[2] A. Destexhe, and D. Contreras, "Neuronal computations with stochastic network states", *Science*, vol. 314, pp. 85-90, Dec 2006.

[3] B. M. Angstkesson and H. T. Toivonen, "A neural network model predictive controller", *J. Process Control*, vol. 16, pp. 937-946, Oct 2006.

[4] J. Dunn, "An investigation into neural network assisted model predictive control for nonlinear systems", Ph.D. thesis, Brunel University, West London, UK, 2001.

[5] J. Q. Huang, and F. L. Lewis, "Control and estimation - Neural-network predictive control for nonlinear dynamic systems with time-delay", *IEEE Trans. Neural Netw.,* vol. 14, pp. 377, Mar 2003.

[6] http://encarta.msn.com/dictionaryV1861616536/happy.html

[7] D. Faraggi, and R. Simon, "The maximum likelihood neural network as a statistical classification model", *J. Stat. Plan. Inference*, vol. 46, pp. 93-104, Jul 1995.

[8] H. C. Diener, J. Dichgans, and F. Bootz, "Functional plasticity of spinal and supraspinal reflexes in maintaining upright stance", *Adv. Otorhinolaryngol.,* vol. 30, pp. 288-290, Jun 1983.

[9] D. E. Rumelhart, J. McClelland, and P. R. Group, *Parallel distributed processing explorations in the microstructure of cognition. Volume I, Foundations*. Cambridge, Ma.; London: MIT Press, 1986.

[10] C. Giertler, I. Bohn, and W. Hauber, "The rat nucleus accumbens is involved in guiding of instrumental responses by stimuli predicting reward magnitude", *Eur. J. Neurosci.*, vol. 18, pp. 1993-1996, Oct 2003.

[11] T. Kalenscher, B. Diekamp, and O. Guíuntuíurkuíun, "Neural architecture of choice behaviour in a concurrent interval schedule", *Eur. J. Neurosci.*, vol. 18, pp. 2627-2637, Nov 2003.

[12] H. C. Breiter, R. L.; Gollub, R. M. Weisskoff, D. N. Kennedy, N. Makris, J. D. Berke, J. M. Goodman, H. L. Kantor, D. R. Gastfriend, J. P. Riorden, R. T. B. R. MathewRosen, and S. E. Hyman, "Acute effects of cocaine on human brain activity and emotion", *Neuron*, vol. 19, p. 591-611, Sept 1997.

[13] M. R. Stefani, and B. Moghaddam, "Rule learning and reward contingency are associated with dissociable patterns of dopamine activation in the rat prefrontal cortex, nucleus accumbens, and dorsal striatum", *J. Neurosci.*, vol. 26, pp. 8810-8818, Aug 2006.

[14] H. C. Breiter, I. Aharon, D. Kahneman, A. Dale, and P. Shizgal, "Functional imaging of neural responses to expectancy and experience of monetary gains and losses", *Neuron*, vol. 30, p. 619-639, May 2001.

[15] C. C. Bell, "Duration of plastic change in a modifiable efference copy", *Brain Res.,* vol. 369, pp. 1-2, Mar 1986.

[16] A. G. Feldman, and M. L. Latash, "Afferent and efferent components of joint position sense; interpretation of kinaesthetic illusion", *Biol. Cybern.,* vol, 42, pp. 205-214, Jan 1982.

[17] A. G. Feldman, and M. L. Latash, "Interaction of afferent and efferent signals underlying joint position sense: empirical and theoretical approaches", *J. Mot. Behav.*, vol. 14, pp. 174-193, May 1982.

[18] T. Sharot, A. M. Riccardi, C. M. Raio, and E. A. Phelps, "Neural mechanisms mediating optimism bias", *Nature*, vol. 450, pp. 102-105, Oct 2007.

Received: November 5, 2007 Revised: November 19, 2007 Accepted: December 10, 2007