Psychoanalytic concepts for the control of emotion in pet-like robots

Stephane Zrehen
Center for Neuromorphic Systems Engineering
California Institute of Technology, 193-36, Pasadena, CA 91125
Email: zrehen@caltech.edu

Abstract

This paper argues the introduction of psychoanalytic concepts for modeling emotion control in a pet robot. I propose a model of the Freudian Ego as part of a neurally modeled "mind", and illustrate its function in the problem of learning to wait. This model is intended for children, dogs and pet-like robots.

Introduction

Children have in common with young dogs a remarkable impatience. When they want something, they expect to obtain it immediately. If they don't, their become intensely frustrated. They express it loudly, through cries or barks. However, they eventually learn to accept waiting. One observes that the intensity of their frustration in these situations decreases with time and education. They are thus able to modify their emotional state, and for the least the conditions in which to express their emotions.

Certainly, a similar capacity is desirable in a pet-like robot. This type of learning is among the first things we have to do when educating children and dogs. If a robot could display the same change of behavior through its interactions with us, it would surely increase its believability as a creature endowed with a psychology.

The ability to control emotions and to decide the right timing for satisfying internal drives is attributed by psychoanalytic theory to one of the Mind's agencies: the Ego.

Psychoanalysis offers a theory of human personality development, based on the dynamical interactions of affects with cognition and behaviors. Ethological theories of animal behavior such as socio-biology account for regularities of behavior observed across all individuals of a species. In contrast, psychoanalytic theory describes individual behavior on a more precise scale. It offers an explanation for the causes of an individual's actions and feelings.

Freudian psychoanalysis is mostly concerned with a person's history. It accounts for particular individual traits of character in terms of past experiences and general rules of mental functioning. It also adds the essential psychological component of individual instincts, desires and personal meaning given to events. Furthermore, it provides insights about the nature of these desires and how they are involved, together with past memories–including phylogenetic memories– in the choice of behaviors taking place at any instant in a human being. It offers a theory of memories and of the psychological apparatus in general. Psychoanalysis is based on the idea that personality traits and tastes observed in adults are linked to childhood events, some due to unconscious processes of repressed–and transformed–childhood memories.

In the project of building a pet robot that behaves like a real dog, affective behavior is crucial, especially in the dynamical aspects of its interaction with people, other robots or real animals around it. In that respect, psychoanalysis is an attractive framework for the design of animal-like robots endowed with a believable psychology.

Psychoanalytic theory is a vast domain, which has undergone several changes since its conception. Its computational modeling can thus be expected to take several years of research. In this paper, I concentrate on the design of a simple model of a "mind" containing an Ego that accounts for learning how to wait and limit the emotional reaction to a state of need.

A selective review of psychoanalytic concepts

I believe the fundamental concepts of importance in psychoanalytic theory for AI research are the following:

1. **The pleasure principle**

   The core principle of psychoanalysis is that humans try to maximize [psychological] pleasure and minimize unpleasure. This is a mere extension of the principle at play in the reinforcement learning paradigm, but applied to psychical agencies instead of reflex mechanisms. An important aspect of this principle is that unpleasure needs not be real. Its evocation can be sufficient. In addition, what causes pleasure and unpleasure can change through life. It cannot be reduced to a pre-defined set of circumstances such as being hit or petted. Conflicts may arise in response to either internal or external events. For instance, a conflict can exist when one desires a person with whom

social rules prohibit relationships. In such cases, the function of the Ego is to displace the representation of desire (see item 3).

In the case of an artificial machine, implementing the pleasure-principle amounts to designing drives or tendencies to look for pleasure, and allowing the appropriate displacements in the face of conflict.

2. **The reality principle**

   The reality principle is the principle by which the person learns that there are limits to his/her pursuit of pleasure. The discovery that not all forms of pleasure are permissible is painful and forces the organism to "invent" new ways to deal with the world, through modification of behavior, and through a shift in the object of desire, that is, displacements. The function of the Ego is to find compromises between the pleasure principle and the reality principle. According to Freud (0, page 3):

   "In consequence of the pre-established connection between sense perception and muscular action, the ego has voluntary movement at its command. It has the task of self-preservation. As regards *external* events, it performs that task by becoming aware of stimuli, by storing up experiences about them (in the memory), by avoiding excessively strong stimuli (through flight), by dealing with moderate stimuli (through adaptation) and finally by learning to bring about expedient changes in the external world to its own advantage (through activity). As regards *internal* events, in relation to the id, it performs that task by gaining control over the demands of the instincts, by deciding whether they are to be allowed satisfaction, by postponing that satisfaction to times and circumstances favourable in the external world or by suppressing their excitations entirely [...] The raising of these tensions is in general felt as *unpleasure* and their lowering as *pleasure* [...] The ego strives after pleasure and seeks to avoid unpleasure. An increase in unpleasure that is expected and foreseen is met by a *signal of anxiety*; the occasion of such an increase, whether it threatens from without or within, is known as a *danger*"

   The reality principle is an internalization of the rules of the external reality. Far from being present at birth, it is acquired through the interactions of the infant with his parents. Since a major element of this principle is to learn how to differ the satisfaction of needs, the infant needs to face a world in which he gets a chance to predict the future to some extent in order to build some confidence that his needs will be taken care of. In doing so, he learns the reality-principle and discovers there are non-immediate ways to obtain pleasure.

   An infant raised in chaotic conditions, with no schedules or stability of emotional responses from his caretakers, is likely to become psychotic. One particular aspect of psychotic patients is their very personal view of the world: they do not share a common interpretation of the world with their peers. In particular, they are likely to attribute intentions to objects that have none. They do not separate the internal from the external. They also expect to obtain instant gratification.

   Freud's definition of the function of the Ego can apply to dogs, for they can learn to delay the satisfaction of their needs.

3. **Displacement**

   In psychoanalytic theory, displacements are a displacement of representation: if an event X is normally represented by a symbol B in the mind, it happens at times that another symbol, B' will be used to represent the same event X. Memories of events can be transformed: the recollection a person has of a particular event may turn out to be completely false. In dreams, it is common that persons or objects are represented by other persons or objects.

   These phenomena are interpreted in psychoanalysis as not being due to noise but as having a cause: they are the expression of the Unconscious. The Unconscious is mostly a repository of desires and repressed memories which remain active although unconscious. They find a way of expressing themselves through transformations which hide their nature.

4. **Psychical energy**

   The concept of psychical energy is one of the most controversial of psychoanalytic theory. Nevertheless, it is very useful for modeling. The idea is that affects are represented by a quantity of energy, which can bind itsef to particular representations, or be displaced to other representations. For example, when we get angry and are forbidden to express our anger, it is not unlikely that we will see things in a negative fashion and will express our anger on any new object, like in hitting the wall. In psychoanalytic terms, we displace the energy of the initial anger to other targets.

Consequences of psychoanalytic concepts on cognitive modeling

In the last section, we saw that it is through the predictability of the environment's response to his personal needs that the child learns to accept to delay fulfillment of his needs. The ability to predict the response to internal needs can be modeled with a hebbian learning paradigm. In addition this explanation answers essential questions about artificial learning models:

1. what is the value of the correct learning rate? In other words, how much of the association must be learned at every presentation?
2. When should learning be stopped?
3. If learning must be stopped at some point, that is, if the network must be stabilized, how can one ensure that enough representative situations have been encountered by the organism? This is especially crucial

when such models are applied to the control of robots and the very network undergoing adaptation is at the same time used to control the robot's behavior. In this case, the actions taken by the robot are chosen without supervision, so there is always a risk that an under-representative set of data will be encountered during the learning period.

The account of psychoanalysis shifts the emphasis from learning objective, complete and affect-less facts about the world to the satisfaction of internal needs and the regulation of emotions. The fact that there is a critical period in early childhood for psychosis to settle–probably between zero and five years–implies that there is a time limit on this primary learning process. Thus, the "right" moment to stop learning depends on age alone. But since biology creates needs at regular intervals, the child gets a chance to learn the response to these needs often enough during this period. The baby will typically get hungry, sleepy, cold or hot a few times a day, which amounts to about 300 "trials" over three months.

In modeling this stage of learning in a robot, it is then possible to pre-program internal drives such as hunger and tiredness to be at their peak at various times of the day. One must also design a system capable of recognizing the appropriate response by the human caretaker to the robot's needs, as well as inappropriate responses. The needs must be implemented in a fashion that requires external intervention such as the presence of food for their extinction. The learning algorithm must be able to learn predictions within the pre-computed number of trials associated with the satisfaction of needs, which specifies unambiguously the right learning rate.

The nature of associations to be learned is of the following kind: "after my meal I get taken out", "After my master comes home and takes off his hat and coat, he feeds me", "No matter how loud I bark, I never get fed in the morning", "When I climb on the couch, I get shouted at", "When I empty my bowels in the living room, I get shouted at", "If I empty my bowels in the street, I get patted on the head". At the end of the training period, the robot has had a chance to learn how to delay the satisfaction of its needs by acquiring knowledge that they are indeed going to be satisfied anyway, after an acceptable delay. What has been learned is of the type: "I *will* experience pleasure in a short while, provided I do not seek to obtain it immediately".

A neural model of the Ego for emotion control in waiting

In this section, I propose a model of learning to wait in the face of an important need such as hunger. As was explained in the introduction, children and dogs will learn not to express intense frustration when they feel hungry, but this takes time. Typically several months. According to psychoanalytic theory, children learn to build confidence that they are going to be satisfied eventually, in a time that is acceptable physiologically. I propose to extend this notion to robots with the following paradigm:

The robot feels "hungry", so it starts barking and expressing impatience through typical movements. All of these reactions are hard-wired in advance. Then the master arrives at the sound of the barks, goes to the kitchen then comes back with a dish of "food" (whose recognition and meaning are also hardwired), shows it to the robot and goes with it to the kitchen where the food is placed in front of the robot. The robot can then "eat". Through eating, the blood sugar level raises. This reduces hunger and pain. After a few similar experiences, the robot learns to follow the master at the minute it feels hungry, and not to express its emotional state anymore. What it has learned is to predict that the master will come and feed it, and that it will soon stop feeling pain.

Predicting the future of the robot's sensory state, including its emotional state, is Ego's function. The model of Ego development is part of a model of the development of a complex sensory-motor architecture represented on Figure 1. Its elements are the following:

- the homeostatic state of the body (H). H measures for instance the sugar level in the blood.
- the drives resulting from H activity (D). D's activity is the resulting psychical representation of the drive, such as hunger.
- the present external sensory state (PS). It recognizes and learns events.
- the expected external sensory state (ES).
- the present emotional state (PE). I limit here the possible emotional states to pain and pleasure, but there could be more.
- the expected emotional state (EE).
- the actions (A).

In this model, the mind's agencies are seen as neuronal structures emerging from the learning properties of neurons, and of the interactions of an organism with its environment. The Ego is modeled as a group of neurons with a priori connections to and from D,ES, PE, EE and A. It receives information about sensory-motor events sequences, and outputs predictions about sensory states.

Learning takes place as follows: the activity pattern in D activates a cell in Ego. At the beginning, the Ego is still blank, so the first winner is chosen randomly. Activation of D only occurs when there is a change in H, so the first Ego winner is expected to learn the sequence of events following the onset of hunger. The sequence ends with the end of the hunger. Ego cells activate themselves, so the current winner remains active until the next change in D.

Cells in PS recognize sensory events, represented in a high-level fashion. For instance "Master entered the room", "Master left", "Master placed food in front of me". I assume that efficient sensory pre-processing is

available for this type of representation and recognition. PS is also a Winner-Take-All, so only one cell is active at a time. However, the activity propagated along the links to Ego decreases with the number of events, in an exponential fashion. It's usual to model such decrease as a function of time (0). But in the present case, it makes more sense for the dynamics to depend more on the succession of sensory events recognized by the robot than on time.

When the final event of the sequence occurs and the robot can "eat", the activity pattern in D changes. This sends a learning signal to the whole network. As a result, the Ego cell learns through the modification of its input links the sequence of sensory events that occurred since the onset of hunger. It also learns through its output links that this sequence leads to the cessation of pain. The learning mechanism mostly amounts to all cells equating their input links to their input activity. The decrease in activity through the links allows to translate the analog value into distance in the sequence. Synaptic links implement the Probabilistic Conditioning Rule paradigm (PCR) (0). This learning algorithm was designed to allow online learning with delayed reinforcement. In the original version of the PCR, where the algorithm was used to learn binary associations, links were represented by two parameters. In the present case, since the analog value of the link–due to the activity decrease of its input cell–is important, three parameters are required. One is the actual link's analog value. The second is a binary parameter corresponding to the link's existence: It is 1 if the link should propagate signals and 0 otherwise. This parameter is interpreted as the use of the hypothesis "this link is partly responsible for the outcome of the situation". Credit assignment is performed through that parameter. The third parameter is the confidence associated to that link. This confidence is assessed and modified either when the robot gets satisfaction fast or when it doesn't get satisfaction after a long time. In the first case, confidence is increased, according to a predefined learning rate. In the second case, it is decreased. Then, the existence parameter is reversed probabilistically, with a probability proportional to the link's confidence value. The interest of such a type of coding and learning scheme is to rule out noise and learn with delayed reinforcement online.

Each cell in ES is connected to its cell in register in PS through a fixed value link. If a sensory event X is coded on a cell N in PS, the cell in register with N in ES is also activated. This group also receives activation from Ego. PCR learning between ES and Ego allows associations between the expected events and the Ego winner cell.

This afference can be interpreted as a prediction of the next sensory event. A cell receiving activation from PS and ES checks the coherence between the two activity patterns. In case of incoherence, activation is sent to the "Pain" cell. Pain is felt, and an expression of this emotional state is performed. If the robot has learned that usually when it feels hungry a given sequence of events occurs and leads to the presence of food, it will bark when faced with an unexpected event.

The special role attributed by Freud to expected unpleasure suggests that a complete model should contain a module coding the expected emotional state. Besides, it is more coherent to code the outputs of Ego as predictions of sensory states, both internal and external. This module should have special links to the module coding the actual present emotional state experienced by the robot. In the present model, links between EE and PE are of fixed value, like the links between ES and PS. In addition, there are reverse inhibitory connections from EE to PE. The prediction of pleasure a priori inhibits pain and vice versa. Information is sent from PE to Ego. The Ego winner learns to associate the sequence with Pain. But when pain ends, the same cell learns to predict the end of pain. When the Pleasure cell in PE is activated by Ego, it inhibits Pain in EE.

In the ideal situation, at the end of the training period, the following sequence of activation takes place:

1. Hunger gets active in Drives. This activates pain and the robot barks.
2. This sends a signal to the Ego cell already associated to Hunger
3. The Ego winner sends activation to EE, which in turn inhibits Pain in PE. The robot stops, or reduces, barking.
4. The Ego winner sends activation to ES cells.
5. The coherence between the activation patterns in PS and ES is checked. In case of incoherence, a signal is sent to Pain. This increases the barking.

Thus, the robot has learned to wait for the food to be brought by the master, and to control it emotional state of frustration.

Discussion

In the learning paradigm described above, the caretaker plays a major role, not unlike that of a parent in psychoanalytic theory. It turns out that the modeling of learning through the Probabilistic Conditioning Rule can explain fairly complex psychological phenomena.

If the master responds in the same fashion everyday, the confidence associated to Ego's input and output links converge to a very high value after the pre-defined number of trials. On the other hand, if the response to the robot's emotional expression changes everyday, these values never get a chance to converge. As a result, the robot could become "psychotic". It could end up with the links corresponding to a rare event having high confidence, and the good links having low confidence. The robot would then expect something that only happened once, and would prove incapable of expecting the right event. As a consequence, it would constantly be frustrated, and would never learn to wait in an acceptable emotional state.

In the present model, the Ego is coded as a Winner-Take-All with afferences from the Drives module. In adult life, the Ego is in particular in charge of finding compromises between the satisfaction of various drives that could be active at the same time. If training is done properly, with one drive taken care of at a time, the Ego gets a chance to learn the sequence that leads to the satisfaction of each drive independently. If the caretaker's behavior is chaotic, it will be hard for the Ego to learn what is the correct sequence to satisfy one given drive. It is therefore be even harder to find compromises since the basic necessary knowledge won't have been acquired. This mechanism translates psychoanalysis's idea that it's the mother who gives meaning to her baby's cries by responding in a prototypical fashion.

It could be argued that a similar architecture without an Ego module could learn the same things. It is true that direct connections from the module coding the expected sensory events to the module coding the emotional state could in principle inhibit pain at the onset of hunger. But it would be an "automatic" connection. The interest of coding an Ego like this is that it allows the computational construction of a kind of homunculus. Indeed, this module oversees several modalities. Its nodes acquire explicit knowledge of the consequences of a given situation. However, such a homunculus is not fraught with the infinite recursion problems traditionally associated with this very notion: All of its properties are explicit, constructed, and of the same type as the rest of the artificial brain.

In the pursuit of modeling consciousness, modeling the Ego in this fashion is attractive. It allows one to account for the construction of knowledge about one's own actions or feelings, or in general, about oneself. Loosely speaking, a robot equipped with such a module could in principle reply to the question: "Why aren't you barking or suffering when you are hungry?" With direct connections between sensory modules alone, it would by definition be impossible since there would be no module with that type of knowledge about the whole system.

Conclusion

In this paper, I argued the use of Freudian psychoanalytic concepts in the design of artificial brains for believable pet robots. In particular, I proposed a learning model for waiting and controlling emotions. This model is based on Freud's definition of the Ego as a mind agency in charge of finding compromises between internal drives and reality's exigencies. It turns out that complex capacities like those attributed to the Ego can be modeled by very simple neural networks. Future work should add other key psychoanalytical concepts to the existing model of the Ego, in order to model all the elements necessary to the development of an artificial creature endowed with a psychology.

References

Sigmund Freud. *An outline of psychoanalysis.* Norton, 1969.

Philippe Gaussier, Andre Joulain, Arnaud Revel, and Stephane Zrehen. Living in a partially structured world: bypassing the limitations of traditional reinforcement learning techniques. *Robotics and Autonomous Systems*, 32(20):225–250, 1997.

Stephane Zrehen and Michael A. Arbib. Understanding Jokes: A Neural Approach to Content-Based Information Retrieval. In *Second International Conference on Autonomous Agents*, pages 343–351. ACM Press, 1997.

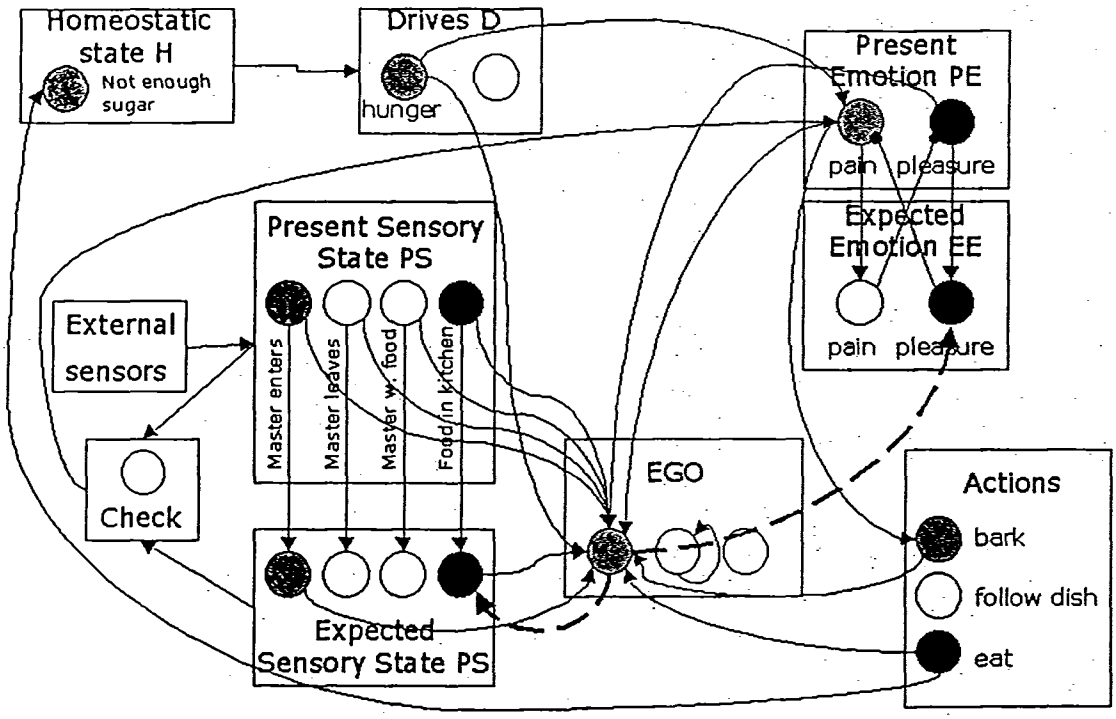

Figure 1: The neural architecture around the Ego. Grey color indicates the cells active at the onset of hunger. Black cells are activated only when the robot starts to eat.