TR-NIS- 0005

Machine Psychodynamics: Toward Emergent Thought

Andrzej BULLER

2006.3.17

国際電気通信基礎技術研究所　ネットワーク情報学研究所
〒619-0288 「けいはんな学研都市」光台二丁目2番地2
Tel: 0774-95-2641　Fax: 0774-95-2647

Advanced Telecommunications Research Institute International (ATR)
Network Informatics Laboratories
2-2-2, Hikaridai, " Keihanna Science City", Kyoto 619-0288, Japan
Tel: +81-774-95-1111　Fax: +81-774-95-2647

MACHINE PSYCHODYNAMICS
TOWARD EMERGENT THOUGHT

Andrzej BULLER

Abstract—This report outlines machine psychodynamics – a paradigm of building brains for robots intended to achieve a human-level intelligence. Machine psychodynamics admits that a target robot is to develop its cognitive structure by itself. What is novelty is that the self-development is to be reinforced by pleasure understood as a measurable quantity that rises when a bodily or psychic tension plummets. Machine psychodynamics also proposes that some ambivalence may accelerate a robot's cognitive growth. Mechanisms for pleasure generation and ambivalence jointly make a psychodynamic robot an adventurous creature. A quest of machine intentionality is also addressed and the notion of proto-intentionality proposed.

Keywords**:** Human-level intelligence, machine psychodynamics, pleasure, adventurousness, ambivalence, proto-intentionality.

1. Introduction

Machines' total integration into human life seems to be the ultimate aim of robotics (Coiffet 2005). Accordingly, the race for human-level machine intelligence is going on. The list of seven issues concerning a target humanoid robot, proposed ten years ago by Rodney Brooks (1996), may serve as a definition of the discipline. The list, let us call it the Big Seven, contains: *bodily form*, *motivation*, *coherence*, *self adaptation*, *development*, *historical contingencies*, and *inspiration from the brain*. On the way to human-level machine intelligence none from the list should be neglected and, furthermore, none from the list should be developed in isolation. Let us call this statement the principle of the Big Seven integration. Obeying the principle seemingly gives us the best chance for demonstrating one day a real C3P0[a] or a real Sonny[b]. On the other hand, one can hardly admit that to any of the seven issues there is an approach that is the only correct one. In this report I concentrate on motivation and coherence. I discuss the issues from the so called psychodynamic perspective that emphasizes the perpetual conflict between

[a] The name of a humanoid robot who was a character in the *Star Wars* film-epopee.

[b] The name of a character of the *I, robot* movie who was a human-shaped robot demonstrating a variety of human-like feelings and motivations.

competing feelings, judgments, or goals, and pays special attention to the intrinsic dynamics of mental events, to which pleasure serves as a universal reinforcer and motivator (cf. Westen 1999, p. 15). I propose to treat pleasure as a measurable quantity induced by certain mental processes in a set of dedicated generators. I argue that mechanisms built based on psychodynamic ideas may substantially boost a robot's cognitive self-development. The considerations outline a new discipline I call *machine psychodynamics*.

The human being strives after pleasure and seeks to avoid unpleasantness, and so shall a target robot. That is the basic tenet of machine psychodynamics. What, in the case of an artifact, may the notion of pleasure mean and where can pleasure come from? That is the basic question of machine psychodynamics. Let us note that the notion of pleasure (sometimes called joy or happiness) is used in a description of several agents developed without any reference to the psychodynamic perspective. Let us mention *Cathexix* (Velasquez 1997), *Kismet* (Breazeal 2002), *ModSAT* (Henninger et al. 2003), *Aryo* (Halavati et al. 2004), or *Max* (Becker et al. 2004). Indeed, the agents can hardly be called psychodynamic since in their case pleasure is a component of a defined state or a function of such a state to be achieved when certain arbitrarily designed conditions are satisfied. Although pleasure defined in this way can be used as a reward reinforcing desired behaviors, it is actually not the agent's desire, but still their designer's desire. What then should be done to get closer to an implementation of the challenging idea of a robot's own desire? The tips can be found in the writings of, e.g., Aristotle, St. Augustine, and Sigmund Freud (related quotations will be provided in the next section). Inspired by the thinkers, machine psychodynamics proposes that pleasure dynamics is closely related to the dynamics of certain bodily or psychic tensions. A bodily tension relates to the degree to which a part of a robot's body deviates from its state of resting. A psychic tension relates to the degree to which a drive (such as boredom, anxiety, fear, or expected pain) deviates from its homeostatic equilibrium (Buller 2006a). And, what is most essential is that not the state of a low tension, but a *move toward* such a state induces pleasure. And the speed of the move correlates with the power of the pleasure signal. Therefore, pleasure can be treated as a measurable quantity that rises when a bodily or psychic tension plummets (Buller 2005). Disregarding the question of how well the above idea matches the still unknown truth about the nature of the human psyche, such a view of pleasure is supposed to work in an artificial mind and to dramatically enhance its ability to self-develop. The supposition is backed by some theoretical considerations and experimental results, among which perhaps the most remarkable was an emergence of a communication behavior in an artificial agent (Liu et al. 2006).

The issue of coherence concerns mechanisms owing to which a humanoid robot is to cope with contradictory feelings, judgments, or goals. In mainstream AI/robotics, in the case of the simultaneous appearance of conflicting ideas about what to do a robot is required to work out, possibly quickly, a rational decision about which of the ideas to implement. To make robots behave so, their designers employ algorithms for arbitration, action selection, or calculating superpositions of related forces or gradients (Arkin, 1998, pp. 111-119). Yet such a machine "self-confidence" looks not too life-like. Let us also note that social psychologists provided empirical evidence that human subjects may abruptly switch from a highly positive evaluation to a highly negative one, or reversely, even if no new data about the object of interest could cause such a switch (Nowak & Vallacher, 1998, pp. 97-98). Therefore, machine psychodynamics proposes a mechanism for conflict resolution that allows contradictory ideas to fight against one another in the literal meaning; accordingly, a psychodynamic robot may sometimes hesitate about what to do or abruptly change its mind (Buller 2006b). In a further discussion I will argue that, owing to this mechanism and to the specific way of pleasure generation, a psychodynamic robot becomes an adventurous creature to which a relatively strong potential for some sort of intentionality I call proto-intentionality could be attributed.

Sadly, adventurousness and proto-intentionality can hardly attract contemporary corporate investors who usually demand "killer applications" on hand and right now. Hence, one can notice the dramatic growth of a zoo of artificial "cleaning guys", "speaking" mascots that "interact" with children, or human-shaped "reception-desk staff" that can welcome guests and even answer their questions. Yet a robot "speaking" prerecorded sentences is only a masquerade – good to impress unsophisticated folks, but a blind alley as for the dream about human-level machine cognition. The path to this level does not lead through hand-crafted speech generators. Furthermore, using such generators simply means killing the chance for a robot's cognitive self-development. On the other hand, one may note that, despite the decades of research on a perceptually-grounded language acquisition, still no machine can demonstrate a two-year-old baby's linguistic competence. Why did the AI community fall so short in such a key sector? One of the causes may be a sin against the principle of the Big Seven integration. Another cause may be yet deeper. Making a machine learn is only a half-success. What is the true point is to make a machine actually want to learn. As for machine psychodynamics, it has up its sleeve a key to the machine's will to learn. The key is a mechanism for an active striving after pleasure.

2. Pleasure Principle

If our goal is to build a robot whose ultimate human-level intelligence is to self-develop in a pleasure-driven manner, we should devote some time to studying pleasure-related phenomena. On the other hand, we had better keep away from the endless debate on the deep nature of pleasure that, for centuries, has been pending on the border of philosophy and psychology. What we need is only an operational definition of this notion. Fortunately, some renowned thinkers provide tips for such a definition. For example, Aristotle (350 BC) wrote:

> We may lay it down that Pleasure is a movement, a movement by which the soul as a whole is consciously brought into its normal state of being; . . . If this is what pleasure is, it is clear that the pleasant is what tends to produce this condition . . . It must therefore be pleasant as a rule to move towards a natural state of being, particularly when a natural process has achieved the complete recovery of that natural state.

Saint Augustine of Hippo (AD 397) wrote:

> Indeed, the very pleasures of human life – not only those which rush upon us unexpectedly and involuntarily, but also those which are voluntary and planned – men obtain by difficulties. There is no pleasure in caring and drinking unless the pains of hunger and thirst have preceded. Drunkards even eat certain salt meats in order to create a painful thirst – and when the drink allays this, it causes pleasure. It is also the custom that the affianced bride should not be immediately given in marriage so that the husband may not esteem her any less, whom as his betrothed he longed for.

Sigmund Freud (1920, p. 4) wrote:

> We have decided to relate pleasure and unpleasure to the quantity of excitation that is present in the mind but is not in any way 'bound'; and to relate them in such a manner that unpleasure corresponds to *increase* in the quantity of excitation and pleasure to *diminution*. . . [and] the factor that determines the feeling is probably the amount of increase and diminution in the quantity of excitation *in a given period of time*. (emphases his)

As it can be noted, Aristotle proposes that pleasure is a dynamical notion, i.e. a "movement" toward a normal/natural state of being we call today a homeostatic equilibrium. Needless to say, in order to be able to move toward a state, one has first to deviate from the state. St. Augustine provides real-life examples about how people deliberately let their organisms deviate from the equilibrium, which results in such unpleasant experiences as painful thirst or unbearable longing, yet the unpleasantness magnifies the subsequent pleasure. If we admit that the deviation from the homeostatic equilibrium may result from the "unbound quantity of excitation", Freud's statement can be interpreted as a clear suggestion that, as for pleasure volume, the speed of return to the equilibrium matters a lot.

Mainstream AI/robotics still seems to be blind to the above tips; however, one can meet solutions that almost beg for being supplemented with mechanisms that would facilitate seeking pleasure as understood in the way Aristotle and Freud suggest. As an example, let us consider the drives on which the motivation system of the MIT *Kismet* (Breazeal 2002) is based. Each of the drives is represented as a device to which a small steady stream of "activation energy" is being provided. The volume of accumulated energy is being measured. If the volume exceeds the upper limit of a defined homeostatic range, this means that the robot is under-stimulated. When a satiatory stimulus is provided to the device, the activation energy starts escaping. However, if the energy volume falls below the lower limit of the homeostatic range, we may say that the stimulus is overwhelming. Depending on the state of its drives, *Kismet* seeks an appropriate stimulus or tries to ease itself of the most harmful one. This solution mimics the important homeostatic mechanisms that in the case of animals serve to maintain certain key physiological parameters within healthy limits (pp. 108-109). As for machine psychodynamics, what the discipline recognizes as a source of pleasure is not the state in which the activation energy is within a homeostatic range, but just the *process of approaching* the state which can take place only if a robot previously deviated from the state. The more the deviation diminished in a given period of time, the higher the pleasure. A psychodynamic robot is proposed to actively seek pleasure through a purposeful "playing" with the deviations.

Regrettably, *Kismet* makes no use of the dynamic of the activation energy. Although *joy* is on the list of the robot's "emotions" and even one of its conditions is defined as "achieving goal" (p. 111), the goal is defined as a particular relationship between the robot and its environment (p. 136). However, it cannot be denied that since the achieved relationship may mean the receiving of an appropriate satiatory stimulus, the solution serves well for homeostasis. And machine psychodynamics does not propose to resign from the mechanisms for homeostasis maintenance. What machine psychodynamics does propose is to supplement the mechanisms with a machinery that measures how the deviation from the homeostatic equilibrium changes in time and, based on the measurement, generates a pleasure signal whose properties correspond to the tips that come from the Aristotelian-Freudian view of pleasure. The pleasure signal can then be used by the robot as a universal reinforcer and motivator of the self-development of various motor skills and cognitive abilities. It can also be noted that purposeful (i.e., pleasure-aimed) "playing" with the deviations from the homeostatic equilibrium brings us closer to the notion of free will.

In order to formulate a technical definition of pleasure, let us first establish order in other related notions. An excessive amount of the activation energy that fills *Kismet*'s drives can be treated as a source of the Freudian unbound excitation. In the case of a deficit of the energy, the satiatory stimulus can be treated as the unbound excitation. In both cases, an "amount" of the excitation positively correlates with the value of deviation from the homeostatic equilibrium. Note that Freud (1940, p. 15), when discussing pleasure-related phenomena, replaced the "amount of unbound excitation" with the word 'tension'. This replacement is justified if we admit that such an unbound excitation is subject to integration, where the integral is tension volume. Hence, tension has been adopted by machine psychodynamics as its key concept – a variable that corresponds to the deviation of a part of the body from its resting state or the deviation of a mental state from a homeostatic equilibrium. The supply of the activation energy and satiatory stimuli are instances of the more general notion of *stimulus*.

Based on all of the above considerations, let us try to formulate a technical definition of pleasure (to be used in robotics rather than in a philosophical debate).

> ***Definition of pleasure*** (draft)**:** *Pleasure is a measurable quantity that reinforces certain reactions and behaviors of a creature and constitutes an attractive purpose of actions the creature may plan and undertake.*

We also need a statement that describes in possibly technical terms the properties of pleasure. The statement must reflect the relationship between pleasure dynamics and tension dynamics. As it was concluded before, tension rises when a related tension plummets. But what happens then? Everyone who can recall an experienced pleasure will surely agree that the feeling more or less abruptly rises and remains for some time after the moment the pleasure-causing stimulus stops being effective; however, the volume of the feeling decays with moderate speed. Indeed, this is the very nature of pleasure. Therefore, we can formulate a draft of a related law:

> ***First law of psychodynamics*** (draft)**:** *Pleasure volume rapidly rises when a related tension plummets, whereas it slowly decays when the tension either rises, remains constant, or diminishes with a relatively low speed.*

The law implies that, although there are pleasure-causing stimuli, the cause is not direct. What directly causes the rise of pleasure is an abrupt discharge of a tension. A stimulus can cause pleasure only via making a tension plummet. Once having risen, the pleasure does not plummet. Rather, it decays relatively slowly. The slow decay gives a creature an opportunity to efficiently increase pleasure volume via repeatable discharging of selected tensions.

Let us consider a mind whose innate feature is a perpetual overwhelming strive to enhance its pleasure record. Let the mind contain a collection of tension-dynamics-driven pleasure generators (working according to the first law of psychodynamics) and a collection of functions mapping certain stimuli onto dynamics of certain tensions. Let us imagine that the mind, in order to enhance its pleasure record, develops for itself not only pleasure-giving reactions to certain stimuli, but discovers and memorizes how pleasurable it is to repeat a certain move several times, or discovers and memorizes that a given kind of pleasure will be the strongest if a certain drive is allowed to deviate from its equilibrium to an extreme value (but not more), or discovers and memorizes a sequence of actions leading to the acquisition of an efficient tension-discharging stimulus, or discovers and memorizes a successful way of planning such sequences... The above vision outlines the *pleasure principle* in its version dedicated to machine psychodynamics.

3. Pleasure applied

The psychodynamic view of pleasure implies that the process of pleasure generation consists of two sub-processes. The first sub-process is a detection of the plummeting of a tension. The second sub-process is following the actual result of the detection with some inertia. If we treat the actual value of a tension as a continuous function of time, a detector of the plummeting can be described using the notion of a derivative. So, the detector can be defined as $\lambda(-\dot{q})$, where q is a plot of a tension volume changing in time, the dot above the character is the symbol of differentiation, and λ is a modifying function that passes arguments of high value through and amplifies them, whereas it suppresses arguments of low value (especially negative ones). Accordingly, the detector will produce a signal as long as the first derivative of a related tension is strongly negative. As for following with inertia, let us note that such a process takes place when the speed of a follower toward the escaper is proportional to the distance between them. Hence, when an output signal is to follow a signal of reference, we may achieve this by assuring that an infinitesimal increment of the output signal is always proportional to the difference between the signal of reference and the output signal. So, a pleasure generator will work when a derivative of the pleasure signal p is always proportional to $\lambda(-\dot{q})-p$. Hence the formula that grasps the first law of psychodynamics quantitatively:

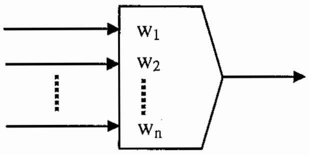
$$\dot{p} = \lambda(-\dot{q}) - \frac{p}{T} \tag{1}$$

where p stands for pleasure volume, q stands for the volume of a related tension, λ is a nonlinear function that favors high arguments, and T is a proportion coefficient to be interpreted as a time constant. Indeed, the higher the T, the stronger the inertia and, accordingly, the slower the decay of pleasure volume.

Generating a pleasure signal is thus solving equation (1) given a plot of a tension volume. Hence, as it can be noted, two operations on signals are necessary: integration and differentiation. In human-level-machine-intelligence-oriented research a recommended method of calculus seems to be one that is possibly close to an economy physical implementation, where the notion of economy applies both to costs of related electronics and to the space in a robot's body to be occupied by the electronics. The effect of inertia can be achieved using an integration element with a negative feedback. Integration –

ubiquitous in Nature and control engineering – is realized using numerous devices, from simple pieces of elastic tissue, through voltage-accumulating membranes, to op-amps. Since an economy differentiation element is still an issue, we can roughly substitute it by a combination of a single integration element with a single delay element. As for delaying devices, Nature employs neural collaterals of various lengths, along which pulses propagate with stable velocities. The sequential circuit technique offers a shift register that can substitute a collateral.

Let us introduce the notion of *integ*, which stands for a device that receives simultaneously a plurality of signals, multiplies each of them by a specified weight, and integrates the weighted sum, where the integral is subject to saturation such that the output value can never go beyond the range [0, 1]. Let the graphical symbol of integ be



where w_1, w_2, …, w_n are weights of related input signals. Figure 1 shows the scheme of a pleasure generator built of two integs and one delay element. Figure 2 shows the answer of the generator to a single act of tension discharge. Figure 3 shows the effect of pleasure accumulation caused by a series of rhythmic discharges of a tension.

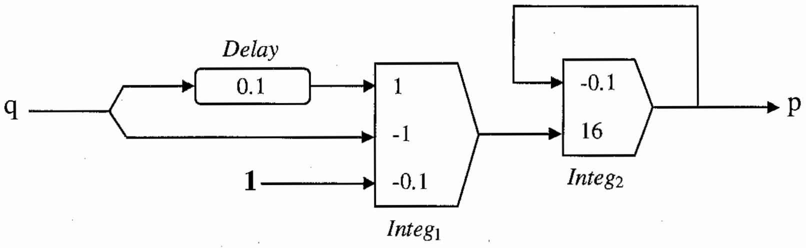


Fig. 1. Pleasure generator. *Delay* is a device that returns an incoming signal unchanged after a specified time; *Integ* is a device that integrates incoming signals, each multiplied by a specified weight while trimming the output value so it never leaves the range [0, 1]; **1** is a constant signal.

As an example of a pleasure-generator application, let us consider a psychodynamic creature I call *psychod*. Let us assume that the creature is equipped with a pair of tentacles that themselves serve as tension accumulators. The greater the deformation of a given tentacle, the higher the tension volume. Assume the creature has no innate mechanism for obstacle avoidance. So, initially, upon each encounter with an obstacle, the psychod performs random movements. When one and only one tentacle touches something and then, by accident, stops touching (which means that a tentacle-related tension plummeted), the pleasure generator connected to the tentacle produces a signal that strengthens the role of the psychod's brain circuits that most substantially contributed to the recent movement. In this way, psychod learns, not supervised at all, to more and more smoothly avoid obstacles, driven only by pleasure defined in psychodynamic terms.

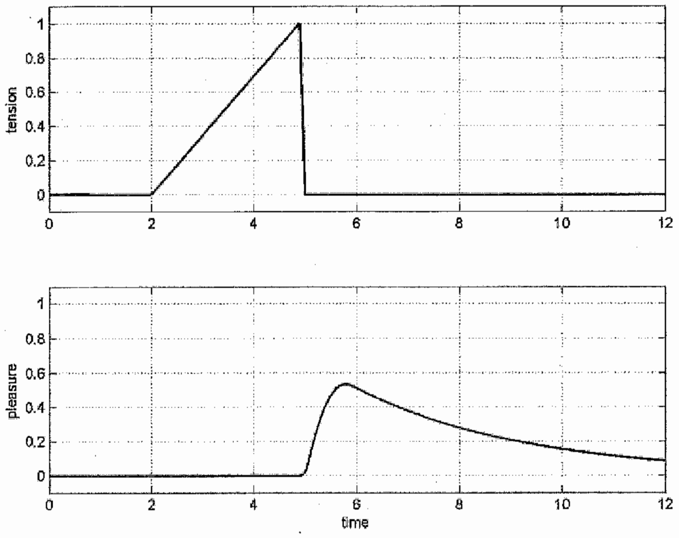


Fig. 2. Pleasure as a measurable quantity that rises when a tension plummets.

What was important in the above example was the principle that after each act of strengthening a behavior, the role of a generator of signals causing random moves diminished a little. The recommended mechanism for this kind of learning should be able to suppress random-signal production with increasing strength to eliminate its impact when the objective of the learning is accomplished. However, the mechanism should revert to randomness when the learning does not succeed within a substantial amount of time or when conditions change such that the already learned behavior stops working. Hence, another law emerges:

Second law of psychodynamics (draft): *The degree of randomness of behavior in a given situation is inversely proportional to the progress in learning what to do in such situation.*

The law is not specific to the psychodynamic perspective. Nonetheless, machine psychodynamics has no choice but to add it to its theory.

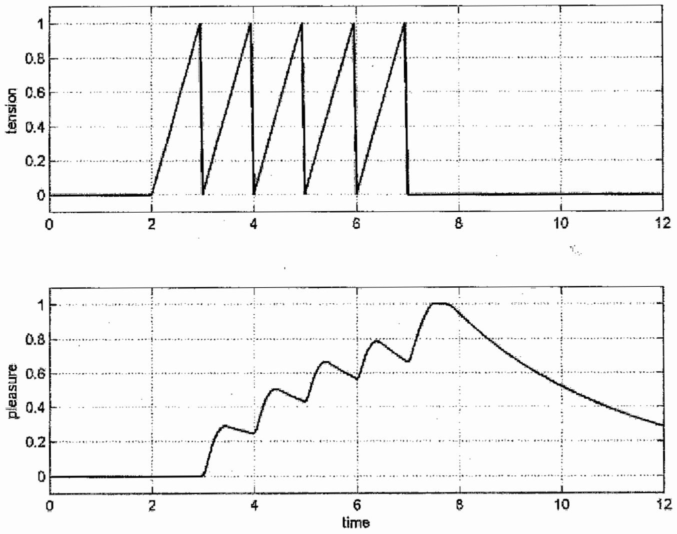


Fig. 3. Accumulation of pleasure via repetitive discharging of a tension.

Let us return to the psychod. When its sensorium gets equipped with a camera and a module for the recognition of objects of interest, the creature becomes capable of learning, pleasure-driven, to approach the objects. However, in this case the brain circuitry must be capable of establishing or strengthening new tension accumulators or links between existing ones to be later used for the reinforcement of behaviors that move the robot from an actual state to a state in which a higher pleasure can be acquired. Hence,

Third law of psychodynamics (draft): *When state Y is a result of behavior B executed in situation X, and Y coincides with the acquisition of pleasure P, then, for situation X, behavior B gets reinforced by the pleasure P, while for other situations behaviors resulting in finding oneself in situation X start*

> *being reinforced by a pleasure whose volume is somewhat smaller than the volume of P.*

The challenge is to handcraft (or evolve) an initial structure in which the third law of psychodynamics could efficiently work toward a general intelligence.

As for approaching objects, note that approaching something by oneself is not the way of object-of-desire-acquisition that human infants master first. Babies first learn a "social way" of handling the environment (Minsky 1985). When something interesting is actually beyond reach, a child produces sounds that sound meaningless, but in fact they express a desire to be provided with a particular item. A good caregiver learns how to interpret the sounds – starting from a pure trial-and-error strategy, but, later, reinforced by the child's smiles resulting from providing the truly desired item. Gradually, according to the second law of psychodynamics, the trial-and-error strategy fades in favor of better and better guessing. An analogical process happens in the child's mind; however, it results in a purposeful modification of the produced sounds to make them more and more similar to the proper names of the items of interest[c]. This effect was observed when *Miao-V* (a simulated mobile robot equipped with a camera, microphone, and speaker) interacted in the way described above with a human investigator. Desire related to particular objects rose randomly and the creature reacted with sounds – initially random, later (according to the laws of psychodynamics) more or more purposeful. This means Miao-V, together with its caregiver, developed for themselves a mutually understandable proto-language. In other words, an emergence of communication behavior took place. When the caregiver was temporarily unavailable, Miao-V (also driven by the laws of psychodynamics) started trying to use its mobility potential to approach the object of desire (Liu et al. 2006).

4. Machine adventurousness

As Marvin Minsky proposes, one 'secret of creativity' may be to develop the knack of accepting the unpleasantness that comes from awkward or painful performances (Minsky, to appear). Indeed, only an adventurous individual may deliberately select a challenging

[c] Of course, such kind of language acquisition is not the only means of developing babies' linguistic competence. They also learn associations between objects/situations and words properly pronounced. Indeed, a caregiver who "lovingly" imitates children's improper speech, may slow down their cognitive development.

gain over an easy one. Psychodynamic robots are adventurous owing to the pleasure principle and the adventurousness pays. First, adventurousness may facilitate survival. Second, it may accelerate cognitive growth.

As for survival, let us consider a situation where a robot's habitat is separated from the rest of the environment by an unsafe zone. Let us assume that the supply of vital resources started decaying in this habitat. Let us also assume that the robot's knowledge includes neither the dimensions of the unsafe zone nor what lies beyond the zone. Needless to say, in this situation, if the robot were not psychodynamic, its fate would be sealed. Fortunately, unlike conventional robots that have no mechanisms facilitating an emergence of the idea to engage in a "purposeless" risk, a psychodynamic robot from time to time ventures into the unsafe zone and deliberately exposes itself to dangers - just to increase the fear-related tension and to get pleasure from discharging it. If the robot does not carry things too far (or is simply lucky), it has a good chance of discovering an area rich in vital resources beyond the unsafe zone (Buller 2006a). The above vision has been implemented as a simulated environment populated by two colonies – a colony of normal food-seeking creatures and a colony of creatures that evolved a psychodynamic mechanism for pleasure seeking via accumulating and discharging a fear related to entering the unsafe zone. The fates of the populations were precisely such as predicted above – the rational food-seekers became extinct, while the adventurous pleasure-seekers found their new niche.[d]

Analogically, the pleasure principle may make a psychodynamic robot "purposelessly" penetrate various areas of its own memories. Unlike a non-psychodynamic robot that confines the usage of its long-term memories to finding only data that are helpful to solve a problem that is already being faced, its psychodynamic cousin may daydream in the literal meaning. Daydreaming allows it to experience, to a certain extent, an increase of bodily and psychic tensions, as well as pleasures resulting from the discharging of the tensions. In order to magnify such substitute pleasures, the robot may embellish facts, design new adventures, or even imagine completely fantastic worlds. It can later try to implement the ideas it has dreamed out. Verily, circuitry that facilitates creativity may emerge just as a result of the robot's strive for more pleasure (Buller, 2006a). In order to achieve machine day-dreaming, it seems unavoidable to equip

[d] M. Joachimczak, personal communication.

a robot with a collection of internal sensors and effectors (or ensure that they would self-develop).

Machine adventurousness may also consist in manipulations on the probability of acquiring a painful strike. Let us assume that pain, like pleasure, is a measurable quantity; however, unlike pleasure, it rises when a tension exceeds a certain level. Let us also assume that pain may also suppress current pleasure. In a simple mobile robot, pain-related tension accumulators can be fuelled by accelerometers. Accordingly, when the robot bounces into an obstacle, the event will result in the appearance of a pain-signal. Let us now consider a simple mechanism for pain avoidance. Let us assume that, owing to an associator, the robot may quickly learn that a new portion of pain is usually preceded by a rapid distortion of one of its tentacles. Let us also assume that an innate mechanism for pain-avoidance-aimed learning reacts to an expected pain with increased randomness of behavior (the second law of psychodynamics), which may appear as panic, yet is quite purposeful. If the expected pain accidentally does not come, the pain-expectation-related tension will plummet. Consequently, a related pleasure will be generated and the circuitry that contributed to the accidentally good maneuver will be strengthened. This mechanism may work in parallel with the obstacle-avoidance-learning mechanism that the psychod which was described in the previous section was equipped with. But, what is most important in the idea of expected-pain-related tension accumulators is that a higher-order control system can use them as a means of an extraordinary (as for robotics) way of pleasure acquisition. Having the accumulators the robot may deliberately undertake a risky action – which may result in severe pain with high probability. But, when the action succeeds and the pain finally does not come, the pain-expectation-related tension will plummet and the robot will acquire a portion of great pleasure.

As a final example, let us consider a little child that deliberately irritates her father. She may do so not because she is a bad child. She may simply engage herself in a kind of risky adventure. The possible penalty may be a reprimand or a spanking. The result may even be only her father's sad face (which is a very painful experience, indeed). Anyway, a penalty-related fear develops. The fear is an uncomfortable tension. But, if the child proves to be smart enough to not overstrain the cord and, furthermore, smart enough to succeed in restoring the lovely atmosphere (e.g., via a comically apologizing smile), the abrupt disappearance of the fear-grounded tension causes a rise in pleasure volume. The side effect of the emotional play may be a new portion of the child's social-cognitive experience. Note that during such play the child more or less consciously induces joy in her parent – just via developing his anger-grounded tension and its abrupt discharge. Verily, if these psychodynamic mechanisms were added to the brain of Kismet, the repertoire of the sociable humanoid's interactions with humans would dramatically increase.

5. Constructive ambivalence

Is the distant object a snake, or only a snake-shaped branch? To accept the already faced challenge, or to give up? To select a longer but safer route, or a shorter but riskier one? To imitate the other individual's behavior, or rather refrain from the imitation? A conventional robot in the face of such dilemmas tries to quickly work out an explicit decision, whereas a psychodynamic robot may endure ambivalence. In a psychodynamic mind contradictory ideas may coexist and each of them may try to suppress all others, for a domination over rival ideas gives the winning one an access to motor drives and, in general, an influence on the course of things. However, fortune is fickle. The winning idea may, after a while, lose to a rival one, and after an unpredictable time win again. An intrinsic dynamics of the process may cause irregular switches of judgments, hesitation, or some inconsistencies in the robot's behavior. In mainstream AI/robotics such hesitation and inconsistencies would be hardly welcomed. Yet, as for cognitive self-development, they may play a pivotal role. Ambivalence may force a robot to develop new methods of judgment and to test their efficiency versus those developed earlier. Also, ambivalence gives a chance to sometimes implement a stupid idea (and, consequently, face an "unnecessary" trouble to cope with), which, as long as the resulting behavior is not too devastating, may give the robot useful knowledge about its own physical or mental capacity.

As an example of a mechanism that facilitates irregular switching let us consider a pair of integs interconnected in such a way that Integ$_1$ receives an external signal r$_1$, (with weight u$_1$), a constant signal **1** (with weight –0.2), and the output from Integ$_2$ (with weight –1), whereas Integ$_2$ receives an external signal r$_2$, (with weight u$_2$), a constant signal **1** (with weight –0.2), and the output from the other Integ$_1$ (with weight –1). The constant signal **1** together with the related negative weight represent a constant leak. Let us assume that weights u$_1$ and u$_2$ are variables whose values can be provided as a pair of another external signals (Fig. 4). If both r_1 and r_2 are constant, the outcome is easily predictable. For $u_1r_1 = u_2r_2 \leq 0.2$ the integs accumulate nothing; accordingly, both outputs are constant and equal 0. For $u_1r_1 = u_2r_2 > 0.2$, the output will increase until achieving the value of $min\{1, u_1r_1 - 0.2\}$. For $u_1r_1 > u_2r_2 > 0.2$, both outputs will initially increase; however, at certain moment q_2 will start diminishing and doing so until it reaches 0, while q_1 will continue increasing until reaching 1.

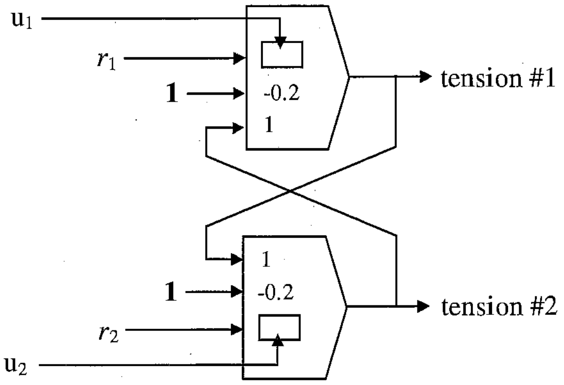


Fig. 4. A combination of two integs for conflict resolution. u_1 and u_2 are contradictory urges, r_1 and r_2 are instances of a random binary series (every two seconds a coin is tossed and in the event of tails 0 is returned and kept until the next toss, whereas in the event of heads, 1 is returned and kept until the next toss); **1** is a constant value; the rectangles are slots to be filled with values of u_1 and u_2 that serve as changing weights to the integs; and tension #1 and tension #2 are associated with the urges u_1 and u_2, respectively.

Let us now consider r_1 and r_2 as instances of such a function of time, that every two seconds a coin is tossed, and in the case of heads, the function will return 1 for the next two seconds, whereas in case of tails, 0 will be returned and kept for the next two seconds. Let us assume that the values of r_1 and r_2 come from separate tossing. What will the outputs from the integs be if, say, in the period of interest u_1 is constant and equal to 0.5, while u_2 is also constant; however, equal to 0.31? An intuitive guess may be that the outputs of the integs will behave in the way similar to the last case considered for constant r's, i.e., both outputs would initially increase (perhaps staggering) and then, at a certain moment, q_2 would get suppressed until it reaches 0, while q_1 would reach the stable value of 1 and keep it. That guess is generally correct. It is obvious that for $Integ_1$ the chance to suppress the rival is much bigger. Hence, $Integ_2$ seems to be doomed to lose the ability to produce a signal soon. In fact, despite such a smaller supply of data to be integrated, $Integ_2$

behaves as if it did not give up and fights bravely (Fig. 5). As a matter of fact, it can count on a certain beneficial property of random series. Namely, it sometimes happens that r_1 suffers a long series of tails, while r_2 enjoys a long series of heads. In such a rare moment tension #2 may achieve 1 and successfully suppress tension #1 at least for a while. Does this phenomenon not match the event when an individual considers two choices for some time – one that by all means is wise and the second one that is visibly stupid – and, surprisingly, decides for the second choice? Events of such kind are not seldom in real life – maybe all of us remember at least a couple of cases of cursing not bad luck but our own stupidity immediately after committing to a decision.

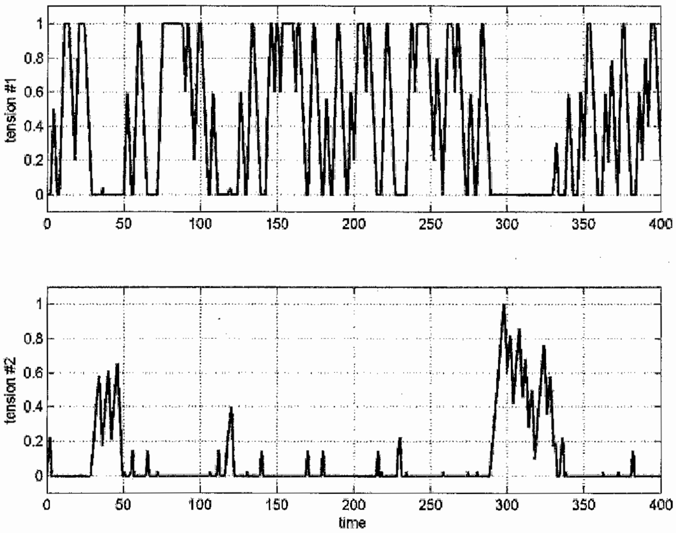


Fig. 5. A counterintuitive behavior of the schematic from Fig. 4 for $u_1 = 0.5$ and $u_2 = 0.31$. It may happen that a tension that comes from a small urge can suppress, at least for a while, a rival tension – even one that comes from a much stronger urge. Perhaps the underlying property of randomness contributes to the phenomenon of intentionality.

The phenomenon discussed above suggests another law whose draft is stated as below:

Fourth law of psychodynamics (draft)**:** *A non-zero tension always has a chance to suppress rival tensions – even those that are much stronger.*

But one can hardly admit that a human being may deliberate in the course of using just a pair of integs. Yes, in fact, the above example was only a simple illustration of the essence of a fight of an idea vs. another idea. However, as it was experimentally confirmed, the fourth law of psychodynamics may work well even on the level of basic urges. *Miao-1* the robot is a simulated mobile creature, in which two contradictory drives fight for the access to motor control. The first drive is hunger related to the state of a battery, and the second one is excitation caused by a toy. A fragment of a related report states:

> *Miao punches the ball / it stopped punching and looks toward the battery charger / Miao turns back to the ball and punches it (though not too vigorously) / suddenly it resigns, turns, and slowly approaches the charger / It gets very close to the charger / Miao sadly looks back at the ball, then turns and starts recharging...*

Of course, the word "sadly", if treated literally, would be by all means farfetched. But the point is that the robot's slow turn and gaze *looked* sad and looked so not because somebody intentionally programmed a masquerade sadness, but because the robot's brain circuitry allowed for a psychodynamic process resulting in such expression (Buller et al. 2005).

As for human-level deliberation, not only the basic urges, but also sophisticated ideas seem to fight. Hence, to build machinery for such a fight, we need suitable representations of such ideas and, based on them, build (or let a robot's brain develop by itself) an appropriate circuitry. A step in this direction is an implementation of *MemeStorm* – a grid of processing nodes inhabited by populations of identical pieces of information called *memes* (Buller & Shimohara 2001). The populations fight against one another for domination in the grid. A single meme means nothing. A single meme can only contribute to a belief. Only when a population of identical memes wins and expels rival populations does it mean an emergent judgment or belief. Memes are in perpetual motion by jumping from one node to another. When two memes meet in a node, they may interact. If they are not related, they will simply change the directions of their motions like balls that have elastically collided. If they represent contradictory facts, both will be annihilated. If one meme contributes to the belief that **"if x, then y"**, while the second one contributes to the belief that fact **x** takes place, a new copy of a meme in favor of the belief in fact **y** will be born. If one meme contributes to the belief that **"if x, then y"**, while the second one contributes to the belief that fact **"not x"** is true, a new copy of a meme in favor of the belief that **"not y"** will be born. It was experimentally confirmed, that when streams of contradictory memes flow to the grid, the dynamics of concluding beliefs resembles the dynamics of q_1 vs. q_2 demonstrated in the experiment with two integs mentioned above

(Buller 2006). But MemeStorm processes propositions, i.e., something more sophisticated than a pair of signals representing basic drives. Moreover, the MemeStorm nodes can be developed to process multimodal memes (Buller 1995). Hence, MemeStorm is proposed as a new concept for working memory and considered as a challenge to the Baddeley-Hitch (1974) view of working memory.

6. Proto-intentionality

Valentino Braitenberg (1984) proposed that free will can be attributed to *Vehicle 12* – a robot equipped with a module that, based on the number of brain elements activated at a given moment calculates the number of brain elements to be activated in the next moment. The related function was a U-curve, inverted and somehow distorted, so the generated numbers were virtually unpredictable for a human observer. To the possible remark that Vehicle 12 only looks as if it had free will Braitenberg answers:

> ...whoever made animals and men may have been satisfied, like myself, a creator of vehicles, with something that for all intents and purposes looks like free will to anyone who deals with his creatures. This at least rules out the possibility of petty exploitation of individuals by means of observation and prediction of their behavior. Furthermore, the individuals will themselves be unable to predict quite what will happen in their brains in the next moment. No doubt this will add to their pride, and they will derive from this the feeling that their actions are without casual determination (p. 69).

Rodney Brooks (2002) attributes an intentionality to *Genghis* – a legged insect-like robot whose brain is a collection of interconnected AFSMs (augmented finite-state machines), where no AFSM has an intelligence higher than a soda machine. When Genghis's array of sensors caught sight of nothing, it waited. When perceiving a moving infrared source, the robot treated it as prey and chased it scrambling over anything in its path. Brooks argues that it was the robot's own will, since there was no place inside the control systems of Genghis to represent any intent to follow something (pp. 48-50).

Regardless of whether one agrees with the thesis that Vehicle 12 and Genghis are intentional creatures, we can at least admit that Braitenberg and Brooks proposed two criteria based on which we may estimate a potential for intentionality. The *Braitenberg criterion* is the lack of causality in a creature's behavior, while the *Brooks criterion* is the lack of a particular goal-related place in a creature's control system. In order not to provoke justified objections from philosophers, let us introduce the notion of *proto-*

intentionality as the name of a feature one may attribute to a robot based on the Braitenberg criterion, Brooks criterion, and other criteria of this kind. Let as, therefore, try to determine the proto-intentionality of *Neko* the robot.

Neko (Buller et al. 2005) has two wheels (each propelled by a dedicated motor), a speaker, a camera, and two touch sensors. The robot's brain contains tension-accumulation units that represent boredom, excitation, fear, and anxiety. Boredom accumulates when Neko does not perceive any object of interest and discharges when it sees one. When the camera detects a green object, the level of excitation increases and remains high as long as the object remains in the visual field. The level of fear becomes high as a reaction to the appearance of a red object, remains high as long as the object remains in the visual field, and solely drops after the object's disappearance. As for anxiety, each of its three accumulators increase spontaneously and independently from the other. Any time Neko turns left, right or back, a resulting discharge of the related accumulators takes place. An arbitrary hardwiring determines which tensions can be suppressed by a given tension and when. A winning tension can activate a behavior module. Fear can cause the robot to turn back and escape. Excitation forces the robot to chase a related object. Anxiety can cause the robot to look around. Neko can learn by itself how to cope with boredom. It can choose between producing a meow, looking around, and going forward. Going forward increases the chance of meeting an object of interest. Meowing can make somebody bring something. The learning is reinforced in a psychodynamic way, i.e., by a signal caused by tension discharge. Owing to the tension representing "irrational" anxiety, in the event of the lack of an object of interest Neko behaved as an animal in a cage, i.e., it wandered back and forth, and "nervously" looked around. As could be noted, Neko's behaviors were virtually unpredictable; however, those who had the possibility to monitor the states of the tension accumulators could know what the robot would do in the next couple of seconds. What was truly unpredictable was the speed of increase of a given tension and, accordingly, whether it would manage to suppress other tensions soon. Therefore, we can say that Neko moderately fulfils the Braitenberg criterion. As for the Brooks criterion, note that, although Neko's designers hardwired a relationship between dominating tensions and particular actions, there was no module containing a definition of an overall goal. Hence, we may admit that Neko strongly fulfils the Brooks criterion of intentionality. On the other hand, note that whatever Neko does, it acts in search of pleasure (related to tension-discharge). Pleasure is a universal reinforcer of pleasure-acquisition-aimed changes in its brain functionality and the general motivator of doing anything. Unlike Genghis, Neko did not stay motionless when it saw nothing. Neko could get bored and start actively seeking excitement. Hence the idea of another criterion of intentionality to be explained in the next paragraph.

Although Neko's mind, as well as the mind of Miao-V mentioned in the section devoted to pleasure generation, includes a tension-discharge-grounded signal to reinforce beneficial changes in its structure, it seems that this is not sufficient to attribute a pleasure-related proto-intentionality to the robots. Machine psychodynamics intends for the pleasure signal not only to reinforce the learning of particular reactive behaviors but, aiming at the self-development of a human-level intelligence, intends to use pleasure as a motivator of the sophisticated planning and executing of plans. A fully psychodynamic robot must be able to deliberately expose itself to inconveniences and dangers – just to increase related tensions and then let them discharge, which might result in pleasure-signal generation. Hence, the *psychodynamic criterion* of intentionality is the ability to achieve a state defined as pleasurable by deliberately plunging oneself into a state defined as unpleasant. This would mean that the creatures who evolved a habit of entering a dangerous zone in order to discharge fear, which I mentioned in the previous section, fulfill the psychodynamic criterion. Nevertheless, the creatures are still so simple that it is not easy to recognize them as free-will-driven creatures, despite the fact that they somehow also meet the Braitenberg criterion and Brooks criterion. What do they lack?

Christof Koch (2004) suggests that one of the signs that a creature may be endowed with consciousness is a behavior revealing hesitation about what to do. Although the suggestion applies to living creatures, we could apply it also to artifacts, provided that hesitation demonstrated by an artifact is by no means a masquerade. This criterion, let us call it the *Koch criterion*, fits Miao-1, the robot mentioned in the previous section. Miao's hesitation results from a fight between contradictory urges. It may change its mind in an unpredictable moment. This property seems to elevate Miao's mind to a level that is not available to insects (regardless of the fact that Miao-1 lacks innate mechanisms facilitating flying, navigating, and the "four F's"[e] that insects are endowed with). Is a frog's mind at a comparable level? To be able to answer that question we would have to have an idea about whether the frog hesitates whether to jump or not when the overall innate condition for jump triggering is satisfied in a relatively small degree. It is quite possible that a simple

[e] feeding, fleeing, fighting, and reproduction.

threshold-based trigger is employed in the frog's brain. But, as for dogs, there is perhaps no doubt that they sometimes hesitate whether to attack or to escape, or whether to obey a calling or to ignore the caller.

There seems to be no way to complete the list of intentionality criteria in a predictable time. Yet dealing with the criteria may appear stimulating for the designers of machine intelligence, especially when the objective is just human-level intelligence. Let us, therefore, devote yet some paragraphs to this topic.

Richard Dawkins hypothesizes that consciousness may arise when the brain's simulation of the world becomes so complete that it must include the model of oneself (1999, p. 59). As can be noted, the *Dawkins criterion* already has been defined. Having one day completed a robot equipped with machinery for handling a world model including the model of oneself, we may make a breakthrough in the issue that today is maybe the hottest one in the field of robotic intelligence – learning from observation and instruction. Related projects have resulted in giving robots the ability to imitate selected human behaviors (e.g., Bentivegna et al. 2004) or learn from verbal communication (e.g., Weng 2004). Machine Psychodynamics intends to supplement these achievements with mechanisms that will make a robot actually want to learn. Let us imagine a robot whose memory hosts two world models – a *model of perceived reality* and a *model of desired reality*. The desired reality may develop driven by, among other things, several sorts of challenges. Let us consider the following story: The robot notices a person (or other robot) juggling balls. A question emerges: "Would I be able to do the difficult thing that the other individual can do?" The question induces a challenge that results in the mental image of oneself juggling too. The difference between the desired reality and the perceived reality may be a source of a strong tension. How to discharge the tension and have the resulting pleasure? Just by learning. Needless to say, the learning can be recognized as a *voluntary* learning. But what to do with such tension when the learning cannot succeed? The theory of machine psychodynamics considers defense mechanisms, i.e., the possibility of redirecting the desire toward another challenge and acting toward a substitute satisfaction. Another defense may consist in a distortion of perceived reality, e.g., a repression of inconvenient memories to "unconscious" zones of the mind, which may reduce the difference between perceived and desired reality (Buller 2004). To have the collection of world models complete, let us also mention a *model of ideal reality* to be acquired during upbringing. Having such a model the mind may develop tensions related to moral dilemmas that may result from the difference between the ideal reality and desired reality (Buller 2002).

Let us now recall the *I, robot* movie – especially a scene in the house of Dr. Lanning – a background character (yet a very important one) whose death is the subject of investigation. A detective switches a video on and watches a recorded lecture for a while. Dr. Lanning spokes from behind the grave:

> There have always been ghosts in the machine. Random segments of code, that have grouped together to form unexpected protocols. Unanticipated, these free radicals engender questions of free will. Creativity. And even the nature of what we might call the soul. Why is it that when some robots are left in darkness, they will seek out the light? Why is it that when robots are stored in an empty space, they will group together, rather than stand alone? How do we explain this behavior? Random segments of code? Or is it *something more*? When does a perceptual schematic become consciousness? When does a difference engine become the search for truth? When does a personality simulation become the bitter mote... of a soul? (my emphasis)

The above Isaac Asimov vision is still pure fiction. But the MemeStorm – a psychodynamic working memory (described in the section devoted to constructive ambivalence) – when developed to process multimodal memes (also mentioned there) has a good chance to make the vision reality. The related criterion of intentionality, let us call it the *Asimov criterion*, is therefore the possibility of free interplay of segments of a code. But, the 'something more' that Dr. Lanning addressed perhaps should not be left disregarded. In reference to this let us consider the message by Yingrui Yang and Selmer Bringsjord (2003):

> Cognitive modelers need to step outside the notion that mere computation will suffice. They must face up to the fact, first, that the human mind encompasses not just the ordinary, humble computation that Newell and all his followers can't see beyond, but also *hyper*computation: information processing at a level *above* Turing machines, a level that can be formalized with help from chaotic neural nets, trial-and-error machines, Zeus machines, and the like. (emphases their)

Even if Yang and Bringsjord are right, yet there are so many mental phenomena that remain to be covered below the Turing limit that there is no need to even consider an abandoning of machine psychodynamics in favor of hypercomputation. Moreover, even if hypercomputation someday riches the stage of practical implementation, it will probably be possible to add a hypercomputing layer to a psychodynamic below-Turing-limit architecture. Anyway, the *Bringsjord criterion* deserves to be added to the current list of the criteria of machine intentionality. The list may now be called the Humble Seven.

Unlike the Brooks's Big Five, an integration of the items from the list of the intentionality criteria is neither practical nor profitable. The criteria are proposed only as a toolset for analyzing or comparing robotic solutions as for their potential of intentionality. On the other hand, any subset of the criteria may serve well as an inspiring target of a particular sub-project in the field of human-level-intelligence-oriented robotics. The Humble Seven by no means can serve as arguments in philosophical debates on a nature of consciousness or free will.

7. Frequently raised objections

Objection 1: *Machine psychodynamics is inspired by Freudian psychoanalysis, which is unscientific.*

Answer: The statement that machine psychodynamics was inspired by *selected elements* of Freud theory would be true. The myth that psychoanalysis is unscientific, which has been adhered to for decades by a fraction of the scientific community, now seems to be facing a fast track to oblivion. After reviewing the recent findings in neurobiology vs. Freud's ideas, Eric Kandel, 2000 Nobel laureate, has concluded that psychoanalysis is "still the most coherent and intellectually satisfying view of the mind". The quote comes from the May 2004 issue of *Scientific American*, not from *Unscientific American*.

Objection 2: *There is no proof that psychodynamic mechanisms implemented in the presented robots can scale.*

Answer: Not everything has to scale. For example, note that human working memory can handle in a given moment only 7±2 items and there is no grounds for even speculation that more would work better. As for other mechanisms, there is no such a law that would state that a researcher must keep the scientific community uninformed about his results until *everything* related is proven. By the way, those who deny one's right to speak about his approach because of a lack of proof of scalability would have probably blocked a publication of the news about landing on the Moon. Indeed, NASA had not proved that their method of landing scales to other celestial bodies.

Objection 3: *Robot hesitation can be achieved very simply – via an oscillator; hence, no exotic psychodynamic mechanism has to be introduced.*

Answer: An oscillator could only help in mimicking a superficial expression of hesitation, which is scarcely the point. In the related experiment the hesitation was a result, not an objective, of the underlying psychodynamic process. Note also that in real life hesitation is by no means regular. Of course, one can artificially distort the time constant of the oscillation, but it would only be a kind of "cargo cult", not a way to a robot's cognitive self-development.

Objection 4: *In which area has a psychodynamic solution outperformed the best of the traditional methods?*

Answer: The objective of developing machine psychodynamics is not to outperform anything. Machine psychodynamics is not an incremental research. The target product is a robot demonstrating a human-level intelligence. Before the objective is accomplished, any comparison of related results seems to be groundless. Indeed, in publications devoted to Kismet – a flag representative of the class of sociable robots – I have not found any suggestion that the solution has outperformed a rival solution. No surprise. Related solutions are unique and incomparable. The race for machine human-level intelligence is going on. It is not too serious idea to judge contestants when they are still so far from the winning-post. One had better bet.

Objection 5: *If machine psychodynamics had made any sense, related papers would have been widely cited, but they are not.*

Answer: The popularity of a scientific idea comes not only from its scientific merits. It is also a function of affiliation, acquaintances, personal charisma, writing style, and, last but not least, the luck of the idea's proponent. Indeed, machine-psychodynamics-related papers are hardly welcomed by renowned conferences or journals. And what is provided as the "reason" of opposition? The four above objections, which are virtually groundless. Unfortunately, there is no opportunity to argue with anonymous referees. By the way, let us recall the *four stages of acceptance* attributed to J. B. S. Haldane, geneticist and evolutionary biologist: i) *this is worthless nonsense;* ii) *this is an interesting, but perverse,* *point of view;* iii) *this is true, but quite unimportant;* iv) *I always said so.* Verily, machine psychodynamics is still somewhere between the second and third stage. And, I guess, one day the approach will achieve the fourth stage, but nobody will remember the conferences from 2000-2005 in which it was presented.

Objection 6: *If machine psychodynamics had made any sense, it would have been invented at a leading university.*

Answer: This statement begs for justification; however, one may hardly be able to provide any. Lo, where was the first successful airplane built? At Harvard? At Berkeley? At MIT? I am sorry, no. The Wright brothers had nothing in common with academia. Consequently, the poor young men did not know that a flying machine is impossible (which scientific authorities preached). And maybe that was the reason that they succeeded.

Objection 7: *Why waste other people's money for such useless research? Would it not be better to buy food and feed poor children?*

Answer: The suggestion that building machine human-level intelligence is useless has poor grounds. The related research gives us a chance to uncover the most intriguing mysteries – the conditions of emergence of thought, the limits of machine intelligence, the very essence of human nature. Even if not all of the mysteries are uncovered satisfactorily, ultra-intelligent robots may become well integrated with people's everyday lives. Let us assume that for all of the funds actually spent for machine-psychodynamics-related research we buy bananas and give one each to one poor child. The children would eat their bananas and in one hour feel hungry again. If machine psychodynamics succeeds and makes a breakthrough on the way to human-level intelligence, the resulting robots will be able to create a giant industry that will give jobs to the fathers of the children. And one day, not the fathers of the children, but humanoid robots will be killed in the fight against terrorism. Would the children prefer to get just one banana? I doubt it.

8. Concluding remarks

I outlined machine psychodynamics – a paradigm of building brains for robots intended to achieve a human-level intelligence. Machine psychodynamics admits that a target robot is to develop its cognitive structure by itself. What is novel is that the robot's behaviors and structural changes that lead to cognitive self-development are to be reinforced by pleasure understood as a measurable quantity. I proposed drafts of four laws of psychodynamics. The first of the laws states that pleasure volume rapidly rises when a related [bodily or psychic] tension plummets, whereas it slowly decays in other events. This law may be employed by a mechanism that forces a robot to develop, all by itself, smarter and smarter methods of changing its relation to the environment or changing the environment itself, just in order to acquire various tension-discharging patterns. The second and third law of psychodynamics provide tips for the construction of such mechanisms.

Machine psychodynamics also proposes that some ambivalence may accelerate a robot's cognitive growth. In the psychodynamic decision-making process contradictory judgments and beliefs fight against one another to dominate a working memory. The winning idea starts being processed toward appropriate action. Nevertheless, owing to a certain counterintuitive property of random series, the actually winning idea may in an unpredictable moment lose to a rival idea. Consequently, the robot may hesitate or abruptly change its mind. These phenomena are reflected in the fourth law of psychodynamics, which states that a non-zero tension-related signal always has a chance to suppress rival-tension-related signals – even those that are much stronger. Mechanisms for pleasure generation and ambivalence jointly make a psychodynamic robot an adventurous creature. The four laws have been formulated as a conclusion from several experiments with simulated and physical robots. However, only for the first law of psychodynamics a general formula has been contrived; the remaining three laws are still narrative drafts, which constitute a challenge motivating further development of the theory.

Machine psychodynamics does not intend to replace the homeostatic mechanisms employed by mainstream AI/robotics. The psychodynamic approach only proposes to supplement the mainstream solutions with mechanisms for active pleasure-seeking, some deliberate irrationality, and constructive ambivalence. These mechanisms are expected to bring a breakthrough in the quest for a machine's self-development toward a human-level intelligence. This belief is justified by several theoretical arguments and a set of experimental results including an emergence of communication behavior. A set of methods for seeking pleasure that a psychodynamic robot should develop may lean toward a state that is compatible with the caregiver's system of values. Hence, despite the pleasure-principle-based "selfishness", a psychodynamic robot may become useful to its human master.

Machine psychodynamics seems to have the potential to substantially contribute to research that aims to give robots the ability to imitate human behaviors and to learn from verbal instruction. The contribution would be a mechanism that makes a robot actually want to learn. I proposed seven criteria for estimating a robot's potential for intentionality and introduced the notion of proto-intentionality to facilitate related discussion.

Sadly, a robot designed to deliberately expose itself to inconveniences and dangers may be hardly welcomed by today's corporate investors. The same undoubtedly applies to a robot that displays visible signs of indecisiveness. Nonetheless, I argue that such troublesome properties may be an unavoidable price for the robot's cognitive self-development up to a level beyond that which can be achieved via handcrafting or simulated evolution.

Acknowledgement: This research was supported by the National Institute of Information and Communications Technology of Japan (NICT), Grant 13-05.

References

1. Aristotle, *Rhetoric, Book I*, Ch. 11, 350 BC (translation by W. Rhys Roberts).
2. R. Arkin, *Behavior-Based Robotics* (The MIT, Press Cambridge, Mass., 1998).
3. St. Augustine of Hippo, *The Confessions, Book VIII*, Ch. 3, AD 397 (translation by A. C. Outler).
4. A. D. Baddeley and G. J. Hitch, Working Memory, In: G. Bower (Ed.) Advances in learning and motivation, 8 (Academic Press, New York, 1974), pp. 47-90.
5. Ch. Becker, S. Knopf, and I. Waschmut, Simulating the Emotion Dynamics of a Multimodal Conversational Agent, in E. André et al. (Eds.), *Affective Dialogue Sysyems (ADS 2004)*, LNAI 3068 (Springer-Verlag, Berlin, 2004), pp. 154-165.
6. D. Bentivegna, C. G. Atkeson, A. Ude, and G. Cheng, Learning to act from observation and practice, *International Journal of Humanoid Robots, 1* (4), 2004, pp. 585-611.
7. V. Braitenberg, Vehicles: Experiments in Synthetic Psychology (The MIT, Press Cambridge, Mass., 1984/1986).
8. C. Breazeal, *Designing Sociable Robots* (A Bradford Book/The MIT Press, Cambridge, Mass., London, England, 2002)
9. R. A. Brooks, Prospects for Human Level Intelligence for Humanoid Robots, in *Proc. 1st Int. Symposium on Humanoid Robots (HURO-96)*, Tokyo, Japan, October 30-31, 1996.
10. R. A. Brooks, *Flesh and Machines: How Robots will Change Us* (Pantheon Books, New York, 2002).
11. A. Buller, Operations on Multimodal Records: Towards a Computational Cognitive Linguistics, *Technical Report TR-95-027*, International Computer Science Institute, Berkeley, 1995.

12. A. Buller, Psychodynamic robot, *Proceedings, 2002 FIRA Robot World Congress*, Seoul, May 26-29, 2002, pp. 26-30.

13. A. Buller, From q-cell to artificial brain, *Artificial Life and Robotics, 8* (1), 2004, pp. 89-94.

14. A. Buller, Building Brains for Robots: A Psychodynamic Approach, in: S. K. Pal, S. Bandyopadhyay, and S. Biswas (Eds.) *Pattern Recognition and Machine Intelligence: First International Conference, PReMI 2005, Kolkata, India, December 20-22, 2005, Proceedings,* LNCS 3776-0070, (Springer-Verlag, Berlin, 2005), pp. 70-79.

15. A. Buller, Machine Psychodynamics: A Key to Artificial Brains, Research Memo, January 15, 2006, ATR Network Informatics Labs., Kyoto.

16. A. Buller, Mechanisms underlying ambivalence: A psychodynamic model, *Estudios de Psicologia, 27* (1), 2006 [in printing].

17. A. Buller, M. Joachimczak, J. Liu, J., & K. Shimohara, K. ATR Artificial Brain Project. 2004 Progress Report, *Artificial Life and Robotics, 9* (4), 1995, pp. 197-201.

18. A. Buller and K. Shimohara, On the dynamics of judgment: doses the butterfly effect take place in human working memory? *Artificial Life and Robotics, 5* (2), 2001, pp. 88-92.

19. P. Coiffet, An introduction to bio-inspired robot design, *International Journal of Humanoid Robots, 2* (3), 2005, pp. 229-276.

20. R. Dawkins, *The Selfish Gene* (Oxford University Press, Oxford, 1976/1999).

21. S. Freud, *Beyond the Pleasure Principle* (W. W. Norton & Co., New York, 1990/1920.

22. S. Freud, *An Outline of Psycho-Analysis* (W. W. Norton & Co., New York, 1989/1940.

23. R. Halavati, S. H. Zadeh, and S. B. Shouraki, Evolution of Pleasure System on Zamin Artificial World, *Proc. 15th IASTED Int. Conference on Modeling and Simulation (MS'04), Marina de Rey, USA.*

24. A. E. Henninger, R. M. Jones, and E. Chown, Behaviors that Emerge from Emotion and Cognition: Implementation and Evaluation of a Symbolic/Connectionist Architecture, *The Second International Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS'03), July 14-18, 2003, Melbourne, Australia*, pp. 321-328.

25. Ch. Koch, *The Quest of Consciousness: A Neurobiological Approach* (Roberts & Co. Publishers, Englewood, Colorado, 2004).

26. J. Liu, A. Buller, and M. Joachimczak M Self-motivated learning agent: skill-development in a growing network mediated by pleasure and tensions, *Transations of the Institute of Systems, Control and Information Engineers, 19* (5) (2006) [in printing].

27. M. Minsky, *The Society of Mind* (Simon & Schuster, New York, 1986).

28. M. Minsky, *The Emotion Machine* (Simon & Schuster, New York, to appear).

29. A. Nowak and R. A. Vallacher, *Dynamical Social Psychology* (Guilford Press, New York, 1988), pp. 89-102.

30. J. D. Velásquez, Modeling Emotions and Other Motivations in Synthetic Agents, *Proc. AAAI-97*, pp. 10-15.

31. J. Weng, Developmental robotics: Theory and experiments, *International Journal of Humanoid Robots, 1* (2), 2004, pp. 199-236.

32. D. Westen, Psychology. Mind, Brain, & Culture. Second Edition (John Wiley & Sons, New York, 1999).

33. Y. Yang and S. Bringsjord, Newell's program, like Hilbert's, is dead; let's move on, *Behavioral and Brain Sciences, 26* (5), 2003, p. 627.