Transformation of Perceptions as Filter Mechanisms in Artificial Intelligence

An Implementation of Psychoanalytic Defense Mechanisms in Artificial Intelligence

Friedrich Gelbard, Dietmar Bruckner, Klaus Doblhammer, Zsofia Kovacs
Institute of Computer Technology
Vienna University of Technology
Vienna, Austria
{gelbard, bruckner, doblhammer, kovacs}@ict.tuwien.ac.at

***Abstract*—When using the human mind as a template for artificial intelligence systems, principles of human thinking have to be modeled. One such very important principle is defense. Hence, in this article we describe a formalism to implement psychoanalytic defense mechanisms in artificial intelligence systems. They can be seen as transformations. A tuple[1] consisting of perceptions or drives (of a software agent in our case) combined with a value indicating their strength is transformed into a different tuple, if a potential interpretation is considered inappropriate by the defense. Some types of defense mechanisms can easily be transformed. Other more difficult ones and abstract types of defense mechanisms are transformed by the use of transformation tables. The transformations can alter, suppress or pass each of the components of a tuple or the whole tuple.**
In this article we show the transformation tables, we give a categorization of defense mechanisms, we show the general form of the transformations and we show some examples of transformations and ways to implement them in artificial intelligence.

Keywords: aritificial intelligence; defense mechanisms; psychoanalysis; cognitive automation; human mind;

I. INTRODUCTION

During the ongoing project ARS (Artificial Recognition System) [1] a psychoanalytic model of the human mind has been developed. A team of computer programmers, psychoanalysts and automation engineers are working on the implementation of the psychoanalytic model of the human mind and simulate the results on a computer platform. One important part of this model are the defense mechanisms of the human mind. Anna Freud gives in [2] a comprehensive introduction into defense mechanisms of the human mind and describes the conflicts of Ego, Super-Ego and Id[2] which lead to the activation of defense mechanisms. One can compare the defense mechanisms with filter mechanisms in artificial intelligence. In our system, inputs are perceptions from the environment and drives like hunger and thirst. All these perceptions are filtered by the defense mechanisms. The defense mechanisms can suppress, alter or merely let pass perceptions or drives. Further explanations of the defense mechanisms in our model are given in [3].

[1] A tuple or n-tuple is a list of elements. For example: (a, 26, house, 1.45)
[2] According to the Freudian Second Topographical model of the human mind

The aim of this publication is to explain the principles how defense mechanisms are implemented in an artificial intelligence system, which data structures we used for perceptions, which data structures we used for drives and how the mapping of incoming data and activation of defense mechanisms is devised.

In Section II we introduce the data structures for perceptions and the data structures for drives and we show how the data structures are combined with a value indicating their strength, namely: quota of affect. Quota of affect is the intensity of the focus of interest on a specific perception or drive, that means how important the perception or drive is right now.

In Section III the data structures are shown in context of the transformations which represent the defense mechanisms and in Section IV the defense mechanisms for drives are introduced. Examples how they work and how they are calculated are given. The defense mechanisms introduced in this paper are only a short extract of defense mechanisms of the human mind. We have chosen those which fortunately fulfill two requirements: they are the most important ones according to psychoanalytic studies and they are easy to implement.

Sections V and VI demonstrate the functionality of the defense mechanisms of perception and the way of implementing them in an artificial intelligence system. In Sections VII and VIII the intensity and activation processes of defense mechanisms are explained and example tables are given to show which defense mechanisms are activated under which conditions.

Finally, we added two sections where we discuss similar projects and give reasons and challenges of implementing psychoanalytic defense mechanisms in artificial intelligence systems.

II. PRELIMINARY DEFINITIONS

In order to define the transformation for drives and perceptions, we need to define the set of tuples to be transformed, first, followed by the data format of the tuples. Furthermore, in this article we distinguish between transformation of drives and the transformation of perceptions as it is done in psychoanalytic theory.

We start by defining the tuples for transformation of drives. Such tuples consist of drive source, drive aim, drive object and quota of affect. For example, in the drive "I want to eat an apple." "I" is the drive source, "eat" is the drive aim and "apple" is the drive object.

Drive tuple = (drive source, drive aim, drive object, quota of affect)

Fig. 1 shows the data structure for drives. The two bars at the beginning and at the end symbolize a linked list because the input of defense mechanisms is always a list of drives or a list of perceptions. In our paper we set focus on only one single drive or perception and, for simplifying the concept, we show how a single drive or perception is altered by the defense mechanisms.

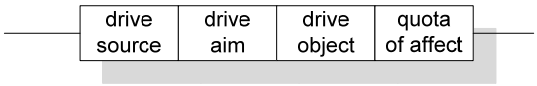

Figure 1. Data structure for drives.

On the other hand, in order to describe the tuples to be transformed concerning perception we need the definitions of thing-presentations and thing-presentation meshes: A thing-presentation is a representation of a rudimentary perception, for example, the shape of a thing, a color of a thing, the rudimentary smell of a thing, haptic perceptions of a thing. All these perceptions of a thing together are called thing-presentation mesh. A thing presentation mesh can be regarded as shown in Fig. 2.

TP …… thing-presentation

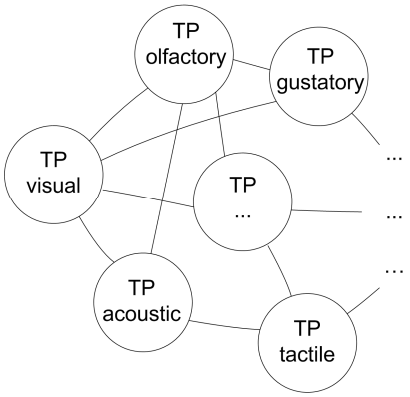

Figure 2. Thing-presenation mesh.

A tuple for transformation of perceptions consists of a thing-presentation mesh and quota of affect. The quota of affect is added to the thing-presentation mesh during the process of perception. Fig. 3 and the following formula show the data structure for perceptions, the perception tuple.

Perception tuple = (thing-presentation mesh, quota of affect)

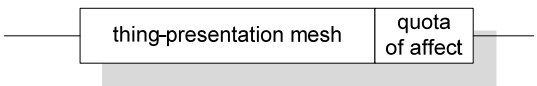

Figure 3. Data structure for perceptions.

III. KINDS OF TRANSFORMATIONS

Having all the definitions we can start looking at the kinds of transformations. To keep the concept of this article as simple as possible we limited the number of basic transformations to the following eight transformations for drives: repression, turning against the self, intellectualization, reaction-formation, displacement, reversal into the opposite, sublimation, and projection. Additionally, the following five defense mechanisms are designated for perceptions: rationalization, projection, separation – depreciation, separation – idealization, and disavowal[3].

In the following, we use the abbreviations:

dt ..... drive tuple
dt' ..... drive tuple' = altered drive tuple after defense mechanisms
pt ..... perception tuple
pt' ..... perception tuple' = altered perception tuple after defense mechanisms
dmd ..... defense mechanisms for drives
dmp ..... defense mechanisms for perception
ds ..... drive source
da ..... drive aim
do ..... drive object
qa ..... quota of affect
tp ..... thing presentation
tpm ..... thing-presentation mesh[4]

Hence, defense mechanisms in formal notation are:

$$dt' = dmd\ (dt) \tag{1}$$

$$(ds', da', do', qa') = dmd\ (ds, da, do, qa) \tag{2}$$

$$pt' = dmp\ (pt) \tag{3}$$

$$(tpm', qa') = dmp\ (tpm, qa) \tag{4}$$

In these equations dmd() and dmp() are functions[5]; ds, da, do, qa, and tpm are the input values for the defense mechanisms and ds', da', do', qa', and tpm' are the output values. The functions dmd() and dmp() map the input values to output values. (1) and (2) show the function for defense mechanisms for drives whereas (3) and (4) show the function for defense mechanisms for perceptions.

Fig. 4 shows the functions and the principle how the defense mechanisms work. In Fig. 4 *defense drives* stands for the function dmd() (defense mechanisms for drives) and *defense percept.* stands for the function dmp() (defense mechanisms for perception). The tables are explained in the following two sections. *Drives* stands for the tuple (drive source, drive aim, drive object, quota of affect), *perceptions* stands for the tuple (thing-presentation mesh, quota of affect),

[3] In literature, disavowal is also called denial.
[4] Tpms is used as plural of tpm (thing-presentation meshes)
[5] The return value of the function is one data structure consisting of several elements.

altered perceptions stands for the tuple (thing-presentation mesh', quota of affect') and *altered drives* stands for the tuple (drive source', drive aim', drive object', quota of affect')

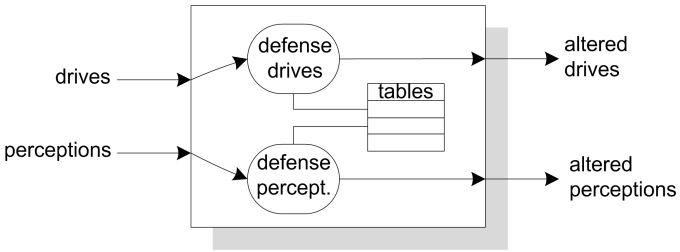

Figure 4. Defense mechanisms.

IV. DRIVE TRANSFORMATIONS

In this section we show how the function dmd() (defense mechanisms for drives) is applied to parts of drive tuples. We only show the altered parts. Unchanged data are omitted here for a simpler enumeration. We enumerate the defense mechanism followed by the component of drive tuple which is altered and how it is altered. The constant NULL means that the component is erased. SELF represents the agent itself. The keywords "intellectualization", "opposite", "different from drive object on same social and cultural level", "complementary active-passive", "social and cultural higher" are resolved with help of the tables as explained later.

- Repression: drive source' = NULL and/or drive aim' = NULL and/or drive object' = NULL
- Turning against the self: drive object' = SELF
- Intellectualization: drive aim' = intellectualization of drive aim
- Reaction formation: drive aim' = opposite of drive aim
- Displacement: drive object' = different from drive object on same social and cultural level
- Reversal into the opposite: drive aim' = complementary active-passive drive aim
- Sublimation: drive aim' = social and cultural higher drive aim
- Projection: drive source = SELF, drive source' = drive object, drive object' = SELF

In order to apply the transformations "intellectualization", "reaction formation" and "reversal into the opposite" we make use of Tab. I. To apply the transformation "displacement" we use Tab. II and for "sublimation" we use Tab. III. The remaining three transformations: "repression", "turning against the self" and "projection" can be applied without the use of a table. The tables were created with help of our team of psychoanalysts.

As an example, we transform the incoming drive tuple (SELF, exhibitionism, people, quota of affect = 0.7) by using intellectualization. Hence, we look up in Tab. I and find in the first column "exhibitionism". We find in the second column the expression "physical body culture". So our transformation results in (SELF, physical body culture, people, quota of affect = 0.7)

The exhibitionist will train and strengthen his body and show it – maybe – during a contest to people. This way the exhibitionistic, originally forbidden, drive is converted to a hobby which is regarded as sports.

In the same way displacement works: Let the incoming drive tuple now be (SELF, want to punch, parents, quota of affect = 0.6). Here the defense mechanism displacement is activated. We look up in Tab. II and find the entry "(punch) cushion" in the second column beside "(punch) parents" in the first column. The resulting output drive tuple is therefore (SELF, want to punch, cushion, quota of affect = 0.6).

If a child is very angry with his/her parents and wants to punch the parents, displacement leads the anger towards a cushion. The child punches the cushion.

Finally, an example for the defense mechanism sublimation given in Tab. III, where the drive aim is shifted toward a social and cultural higher and by internalized rules allowed drive aim'. Incoming drive tuple: (SELF, play, excrement, quota of affect = 0.3), outgoing drive tuple: (SELF, sculpting, clay, quota of affect = 0.3)

TABLE I. INTELLECTUALIZATION, REACTION FORMATION, REVERSAL INTO THE OPPOSITE[6].

Action	Intellectualization of action	Opposite of action (reaction-formation)	Complementary active/passive action (reversal into the opposite)
drink alcohol	being very thirsty, being cool	aggression against all alcohol consumers	become barkeeper
exhibitionism	physical body culture	dresses always correctly	voyeurism
smoker	just adjusting to others	aggression against smokers	sells cigarettes

TABLE II. DISPLACEMENT[6].

Drive object	Drive object' for displacement
(punch) parents	(punch) cushion
(angry with) a friend	(angry at) traffic

TABLE III. SUBLIMATION[6].

Drive aim	Drive aim' for sublimation
play with excrement	sculpting
masturbation	artistic work
egoistic behavior	push own interests as manager

V. THING-PRESENTATION MESH VERSUS LIST OF THING-PRESENTATION MESHES

Transformations of perceptions are more difficult to implement in artificial intelligence systems than the transformations of drives because perception data are not as structured as drives. Whereas the data structure for drive comprises drive source, drive aim, drive object and quota of affect, the data structure for perceptions consists of a thing-presentation mesh (see Fig. 2) and quota of affect.

[6] Tables are shortened for demonstration purposes

In contrast, in our simulation model called ARS [1] which is not part of this article, perceptions are characterized by a list of thing-presentation meshes. We used a list of thing-presentation meshes because humans perceive many things at a certain time all of which have to pass the defense mechanisms. And we split the list of thing-presentation meshes in our model to pass only one single thing-presentation mesh at a time through the defense mechanisms.

The disadvantage of this method is that we lose associations of one single thing-presentation mesh to all other multimodal thing-presentation meshes perceived at the same time. This could distort the performance of the defense mechanisms. In a future article it is planned to take these associations of thing-presentation meshes into consideration while passing the perceptions through the defense mechanisms.

In the meanwhile, we limit one perception to one thing-presentation mesh for demonstration purpose and for easier explanation of the principle. See Eq. (4).

It should be noted that after the perception-data passed the defense mechanisms the thing-presentation mesh or parts of it can be void in case of disavowal and also quota of affect can be void.

Furthermore, in our simulation environment [4] we define a set of possible perceptions. These perceptions are thing-presentation meshes in the subconscious. Examples are:

- shape or smell of a person, sound of steps, shape of male sexual characteristic, sound of a male voice
- shape of female sexual characteristic, sound of a female voice
- shape of nutrition, smell of nutrition, gustatory perception of nutrition, tactile perception of nutrition
- shape of enemy

VI. Perception Transformations

This section is to show how the function dmp() (defense mechanisms for perception), introduced in Eq. (3) and Eq. (4), is applied to thing-presentation meshes and their components, the thing-presentations. Thing-presentations can be altered (rationalization), thing-presentations or parts of it can be erased (separation – deprecation, separation – idealization, disavowal) and quota of affect can be erased or altered, respectively. We enumerate the defense mechanisms followed by the rule that applies and alters the thing-presentation mesh or quota of affect. Again, the constant NULL means that the component is erased. For perceptions we define the following defense mechanisms.

- Rationalization: thing-presentation' = rationalized thing-presentation
- Projection: SELF in thing-presentation mesh is replaced by a different object/person
- Separation – depreciation: All positive thing-presentation meshes' = NULL (see list of positive connotations)
- Separation – idealization: All negative thing-presentation meshes' = NULL (see list of negative connotations)
- Projective identification: thing-presentation mesh' = NULL, thing-presentation mesh of a different agent = thing-presentation mesh, quota of affect' = NULL, quota of affect of a different agent (individual) = quota of affect
- Disavowal or denial: thing-presentation mesh' (or parts of it) = NULL

We can apply the defense mechanisms "projection", "projective identification" and "disavowal or denial" without any table, for the defense mechanisms "separation–deprecation" and "separation–idealization" we use Tab. V. In the first column of Tab. V some examples for positive connotations are given and in the second column examples for negative connotations are listed. The functionality of separation is as follows: Separation-deprecation erases[7] all positive connotations of the thing-presentation mesh and separation-idealization erases all negative connotations of the thing-presentation mesh.

Furthermore, the defense mechanism "rationalization" is, again, implemented by use of a table (see Tab. IV). Rationalization is mostly directed against inner perceptions like feelings, thoughts but also actions. For example, if an individual feels fear but rationalizes the fear the individual can say: "I am just careful." Another example for rationalization is: If someone turns insults against the self or feels always guilty he/she can say: "I am just so selfless."

TABLE IV. RATIONALIZATION[8].

Thing-presentation mesh	Rationalized thing-presentation mesh
being too fearful	carefulness
always feeling guilty	selflessness
alcohol addiction	just a little thirsty, acting cool

About "projective identification" one has to say that we need two agents (individuals). In our simulation environment we have the agent SELF and a second agent. Projective identification means that the SELF-agent loses his perception and quota of affect, for example inner perception of fear, and the perception and quota of affect is projected (transferred) to the second agent. In our simulation environment that is easy to implement but for future projects some (subconscious) signals have to be provided to transfer perceptions and quota of affect from one agent to another.

TABLE V. POSITIVE AND NEGATIVE CONNOTATIONS[8].

Positive connotations	List of negative connotations
shape of male sex symbols	shape of enemy
shape of female sex symbols	aggressive behavior
shape of nutrition	screaming person
voice of caring mother	voice of monitory mother

[7] Erase means to set the value of the thing-presentation to NULL.
[8] Tables are shortened for demonstration purposes

Finally, one example for disavowal or denial is given: If someone denies a certain perception, the perception data structure looks like: (NULL, quota of affect = 0.8). The quota of affect remains and can later on reappear and be attached to a new perception. In this case, a new perception gets a different quota of affect and the individual does not know where the quota of affect comes from.

VII. Intensity of the Transformation

It is very important to take always the quota of affect into consideration. Quota of affect means how strong the desire is to satisfy the drive. That means, there is a value how high the quota of affect of the transformed desire of perception is in dependence of the quota of affect of the original desire.

As one defense mechanism can just as well act smoothly as altering the drive or perception almost 100%, a numerical value for the intensity of the defense mechanism is added during the triggering and execution of the defense mechanism. The range of the intensity of a defense mechanism can be between 1% and 100%. I.e., the function call of the defense mechanism repression for a certain drive can look like:

defense_mechanism_drive (repression, 60%, drive source, drive aim, drive object, quota of affect);

Moreover, the interesting thing of defense mechanisms is that after the defense it is possible that two drives or perceptions exist: The original not fully defended drive or perception and the altered or partly altered one. The important thing is that the quota of affect must be divided according to the rate quota of affect of original, reduced drive to new drive which was created or altered by the defense.

Again, we look at our example defense mechanism repression of a drive. We assume that the intensity of the defense mechanism is 90% and take into account the previously defined transformation rules for drives. Thus, the outcome of the transformation will be: First, the original drive still exists but has now a quota of affect of 10% of the original quota of affect of the drive. And second, the denied drive appears holding a quota of affect which corresponds to 90% of the quota of affect of the original drive:

- Original drive: (drive source, drive aim, drive object, 10% quota of affect)
- Denied drive: (drive source, denial of drive aim, denial of drive object, 90% quota of affect)

The quota of affect of the original drive is now very low, namely 10% and a big share of psychic energy is invested into the countercathexsis to keep the denial of the new drive.

VIII. Activation of Transformations

One final question is left open: When are the transformations activated and how are they activated? Therefore, we installed another table which tells us which kind of defense mechanism has to be activated, with which intensity it is activated and in reaction to which drive or perception it is activated. Tab. VI is intended for activating the defense mechanisms for drives and a separate table, Tab. VII, shows the activation of defense mechanisms for perceptions. In these tables we give examples for the activation of defense mechanisms for a random individual who activates the following defense mechanisms under the given circumstances:

- He starts to chaw on a pencil if he feels the desire to smoke.
- Intellectualization, if he feels sexually attracted to his mother. He explains the sexual attraction with: I like my mother. I like to be with her. It is not a crime to like someone.
- He projects his fear of the aggressive father to his uncle and says: “My uncle is anxious about my father.”
- He rationalizes his sadness about his friend who is aggressive against someone else. He says: “My friend just tells the truth and directs the other person to the right way.”

A. *Activation for Drives*

In the first four columns of Tab. VI drive source, drive aim, drive object and the quota of affect are quoted. If a drive is detected which fulfills the shown criteria of the first four columns in Tab. VI the defense mechanism shown in column *defense mechanism to activate* is activated with the given intensity in the column *intensity of defense mechanism*.

B. *Activation for Perceptions*

Similarly, in Tab. VII the first three columns show the activation criteria for defense mechanisms for perception and in the last three columns the defense mechanisms to activate, the intensity of activation and the altered thing-presentation mesh are given.

TABLE VI. ACTIVATION OF DEFENSE MECHANISMS FOR DRIVES (EXAMPLES).

Drive source	Drive aim	Drive object	Affect	Quota of affect	Defense mechanism to activate	Intensity of defense mechanism
nicotine homeostasis	light a cigarette	cigarette	greed	50%	displacement (displaced object: pencil)	80%
hormones, sexual organs	sex	mother	sexually attracted	30%	intellectualization	100%

TABLE VII. ACTIVATION OF DEFENSE MECHANISMS FOR PERCEPTION (EXAMPLES)

Incoming thing-presentation mesh	Affect	Quota of affect	Defense mechanism to activate	Intensity of defense mechanism	Altered thing-presentation mesh
thing-presentation of aggressive, angry father	anxious	80%	projection	100%	thing-presentation mesh of uncle who is anxious about father
thing-presentation mesh of a friend who is aggressive against someone else	anger	60%	rationalization	70%	thing-presentation mesh of a friend who says the truth to someone else who deserves it

IX. SIMILAR PROJECTS

In [5] a similar notion is applied to defense mechanisms of the mind. Suppes and Warren speak of propositions rather than drives and perceptions. They use triples of actor-action-object in accordance to Freud's subject-verb-object. We think that a triple does not describe a drive or perception satisfyingly. Thus, we used more detailed data structures for drives (drive source, drive aim, drive object, quota of affect) and for perceptions (thing-presentation mesh, quota of affect).

The authors of [5] leave quota of affect out of consideration and limit the outcome of a transformation to one proposition. We apply a defense mechanism only to a certain degree and keep the original drive or perception to a certain limited degree. [5] lacks of suggestions how transformations of propositions can be implemented. In our article we use tables to map altered parts of the drive data structure to original ones.

In [6] a model for the conflicts of Ego, Id and Super-Ego is developed. [6] regards the model from a general point of view without mentioning the data structures or presenting examples how his defense mechanisms work. He gives an insight into the data flow of his model, though, and explains the interaction process of the components.

Andrzej Buller presents a "psychodynamic model" in [7] and describes the embedded functions of the model. In his model defense mechanisms are included that "can cause changes in the models of reality". He quotes the defense mechanisms repression, projection, denial, rationalization and sublimation and he speaks of tension reduction by applying the defense mechanisms. Due to the defense mechanisms in his model three types of reality arise: ideal reality, perceived reality and desired reality. Data structures and implementations are mentioned for the "working memory" of his model but not for the defense mechanisms.

X. DISCUSSION AND CONCLUSION

We introduced new paradigms in artificial intelligence in this paper and showed a way how to implement and test them in a simulation environment. Special attention was given to a certain part of our psychoanalytic model, the so called defense mechanisms of the human mind. The defense mechanisms represent the filter mechanism for incoming data from the environment and filter mechanisms for inner drives. In particular we explained data structures for incoming perceptions and showed data structures for inner drives which represent the desires of the system. We developed a way to implement the defense mechanisms in an artificial intelligence system. One big advantage of our model is that it is universally applicable for various fields and problems in artificial intelligence and that complexity of incoming data is quickly reduced in our system due to defense mechanisms.

Our model [8] and the incorporated defense mechanisms show new ways and methods in artificial intelligence. There are some similar projects and notions which explain the implementation of defense mechanisms of the human mind in artificial intelligence systems, though. Examples for projects are given in Section IX and in [5], [6], [7], [9] and [10], but none of them is as comprehensive and as detailed as our ARS project.

Our approach is seminal but there are some drawbacks: First, the capabilities of the agent, the defense mechanisms and its environment in our simulation program are very limited. Second, there are no "real world" implementations so far – but are planned in the future. The potential and the scope of our project are not fully exploited, yet. Nevertheless, the way of implementing psychoanalytical notions in artificial intelligence is promising and pivotal and can bring great opportunities in the near future.

REFERENCES

[1] D. Dietrich, D. Bruckner, G. Zucker, B. Müller, and A. Tmej, Psychoanalytical Model for Automation and Robotics, invited, Proceedings of the 9th IEEE AFRICON'09.

[2] A. Freud, The Ego and the Mechanisms of Defense, London, Hogarth Press and Institute of Psycho-Analysis, GB, 1937. Revised edition: 1966 (US), 1968 (UK)

[3] C. Riediger, Psychoanalytical Defense Mechanisms Applied to Autonomous Agents, Vienna University of Technology, Faculty of Informatics, Institute of Computer Technology, Austria, 2009.

[4] T. Deutsch, H. Zeilinger, R. Lang, Simulation Results for the ARS-PA Model. In: Proceedings of 2007 IEEE International Conference on Industrial Informatics INDIN07, 2007, S. 1021 - 1026.

[5] P. Suppes, H. Warren, On the Generation of Classification of Defense Mechanisms, The International Journal of Psychoanalysis, 56: 405-141, USA, 1975.

[6] U. Moser, Theorie der Abwehrprozesse, Brandes & Apsel Verlag GmbH, Frankfurt am Main, Germany, 2009.

[7] A. Buller, Volitron: On a Psychodynamic Robot and Its Four Realities, ATR Human Information Science Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan, 2003.

[8] D. Dietrich, G. Fodor, G. Zucker, D. Bruckner (Hrsg.): Simulating the Mind - A Technical Neuropsychoanalytical Approach, Springer-Verlag, 2008, S. 436.

[9] C. Becker-Asano and I. Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. Autonomous Agents and Multi-Agent Systems, 20(1):32-49, 2009.

[10] U. Ramamurthy, B. J. Baars, S. K. D'Mello, and S. Franklin. Lida: A working model of cognition. Proceedings of the 7th International Conference on Cognitive Modeling, pages 244-249, 2006.