
Friedrich Gelbard, Dietmar Bruckner
Institute of Computer Technology, Vienna University of Technology
gelbard@ict.tuwien.ac.at, bruckner@ict.tuwien.ac.at

Abstract—Conflicting goals, rules, and/or input data can cause problems in the decision making unit of software agents. Hence, in this paper we introduce psychoanalytic defense mechanisms to be implemented in multi-agent systems to resolve these conflicts. We give a general insight into defense mechanisms in psychoanalysis and how we transfer this theory to multi-agent systems. We describe the kinds of conflicts in the decision making process of software agents where the implementation of psychoanalytic defense mechanisms help to improve the stability of the system and to enhance the performance of the decision making unit. We explain different aspects and important notions which have to be obeyed while implementing defense mechanisms in cognitive systems. Therefore, we provide some example detection rules to find conflicting data, conflicting drives, conflicting action plans, and conflicting intentions in cognitive systems. These detection rules help to exclude unreal input data from the environment by performing a reality check. Defense mechanisms which we implemented are repression, denial, and reaction formation. Finally, we give examples for applications of defense mechanisms in cognitive systems.

Keywords: software agents; cognitive science; defense mechanisms; psychoanalysis; cognitive automation; modelling the human mind;

I. INTRODUCTION

Conflicting inference rules, conflicting input data, and conflicting goals of software agents, often lead to evaluation instabilities and not necessarily converging algorithms in multi-agent systems. In this paper we introduce psychoanalytic defense mechanisms [1] to be implemented in computer science to resolve these conflicts. The implemented defense mechanisms are part of a larger project, ARS [2], which aims to provide a model of the human mind according to psychoanalytic notions.

In the current article we explain the general principle of defense mechanisms in cognitive systems, we describe how defense mechanisms can lead to more stability in the decision making unit of software agents, and we show an application example in our simulation environment for defense mechanisms in cognitive science. Furthermore, we explain how the emotion fear can be created during the process of defense and illuminate the term quota of affect that determines which defense mechanisms are chosen to resolve a certain conflict.

In Section II we give a brief overview of cognitive architectures of software agents as discussed in [3], [4], and [5]. We discuss symbolic cognitive architectures, connectionist cognitive architectures, and hybrid cognitive architectures.

In Section III we describe, first, the principle of psychoanalytic defense mechanisms and how we transferred the principle of defense mechanisms to be implemented in multi-agent systems.

Then we elucidate the situations and conflicts in the decision making process of software agent systems where defense mechanisms can lead to more stability and better performance of the system. We give examples of working rules of defense mechanisms and Super-Ego rules. (Section III.C) We explain quota of affect which is used to set the focus of attention. (Section III.D) We describe tasks of defense mechanisms in software agents (Section III.E) and emersion of fear (Section III.F).

II. STATE OF THE ART

As of today, there are some research institutions which try to model the human brain in order to find new notions for artificial general intelligence systems. One common issue of these artificial intelligence architectures is the striving to implement ideas of the human cognition in artificial general intelligence. These systems are subsumed under the title of cognitive architectures.

In the following section we introduce important cognitive architectures of autonomous software agent systems and summarize their most important aspects.

A. Cognitive Architectures

Cognitive architectures in artificial intelligence try to model the human cognition process and human problem solving capabilities. Cognitive architectures are divided into three kinds: Symbolic cognitive architectures, connectionist cognitive architectures - also called emergent cognitive architectures, and hybrid cognitive architectures which combine characteristics of both, symbolic and connectionist [3], [4], [5].

B. Symbolic Cognitive Architectures

Symbolic cognitive architectures have a pre-defined set of functionalities and generic rules which represent the human’s cognitive processes.

In 1997, Kieras and Mayers developed the EPIC cognitive architecture at the University of Michigan. EPIC stand for Executive Process - Interactive Control [6]. The architecture’s aim is to model human’s multimodal cognition and human’s capability for multi-task performance. EPIC’s processing unit
comprises a Production Rule Interpreter and the Working Memory. 

In 2004, ADAPT [7] was published. It implements active perception, e.g. goal oriented perception like “find the blue block”, and parallel processing of schemas for planning and perception. Schemas are the action plans of ADAPT and are represented declaratively in its working memory.

Another cognitive architecture project was issued in 2006. [8] In this article, Langley and Choi present ICARUS, a project based on cognitive psychology. They describe memory structures, presentation structures, performance structures, and learning structures. Cognitive processes in ICARUS are conceptual inference, goal selection and skill execution. Incremental learning is implemented in ICARUS by the ability to store new facts in short-term and long-term memory.

The most widely known symbolic cognitive architectures are ACT-R [9], [10] and Soar [11], [12]. The memory structure of ACT-R and Soar, both, comprise a long-term memory and a working memory. One of the differences between ACT-R and Soar is that in Soar two or more production rules can fire simultaneously [13].

Further examples of symbolic cognitive architectures are GLAIR [14], CoSy [15], and EPIC-Soar [16].

C. Connectionist Cognitive Architectures

In contrast to symbolic cognitive architectures, connectionist cognitive architectures do not have pre-defined action plans per se. The cognitive functions of connectionist architectures are more or less emergent and arise out of a combination of hierarchical layers. An example is Rodney Brooks’ Subsumption architecture [17].

Another connectionist concept is the Global Workspace Theory by Baars [18]. There, our consciousness is compared to a spotlight in a theater. The spotlight is the focus of attention and the accompanying perceptions are stored in a short-term working memory. Everything remaining in the dark is the sub-conscious. Shanahan [19] developed in 2006 a cognitive architecture which is based on Baars’ Global Workspace Theory.

Further connectionist cognitive architectures are for example I-C SDAL [20] and DARWIN [21].

D. Hybrid Cognitive Architectures

Hybrid architectures combine the processing of high-level symbols with the emergent architecture of low-level self-organizing properties.

In [3] hybrid cognitive architectures are categorized into two groups: localist-distribute architectures and symbolic-connectionist architectures. The first one combines local principles of data representation with distributed memory concepts. The second one combines rule-based memory with emergent low-level perception processing.

The classification of cognitive architectures varies. E.g., [3] and [22] classify the widely known architecture ACT-R as a hybrid cognitive architecture because it uses symbolic constructs like chunks and on the other hand ACT-R’s production system uses a sub-symbolic architecture. In contrast, [5] classifies ACT-R as a symbolic cognitive architecture.

Further examples of hybrid cognitive architectures are iCub humanoid robot [23], LIDA [24], CLARION [25], and PACO-PLUS [26].

III. APPLICATION OF PSYCHOANALYTIC DEFENSE MECHANISMS IN SOFTWARE AGENTS

In psychoanalysis, the second topographical model by S. Freud [27] explains the interaction of Super-Ego, Ego, and Id (see Fig. 1). Based upon this model, Moser [28] devised in 2009 an implementable model of Super-Ego, Ego, and Id. In contrast to Moser, our ARS model [5] comprises the decision making process of software agents as well and applies Suppes and Warren’s transformation rules for defense mechanisms (see [30]). Thus, ARS is the first implementable comprehensive psychoanalytic model of the human mind which incorporates Freud’s second topographical model.

A. Defense Mechanisms in Psychoanalysis

Fig. 1 shows the functionality of the defense mechanisms and displays S. Freud’s second topographical model.

---

1 Sub-symbolic architectures do not use symbols but different structures like, for example, neuronal networks.
Fig. 1 shows the Id which represents external perceptions and inner drive demands and the Super-Ego which represents social rules. Super-Ego’s social rules provoke conflicts with the drive demands of the Id. The Ego, on the other hand, tries with the help of the defense mechanisms to avoid the conflicts. The defense mechanisms alter the conflicting drive demands until they allow a conflict-free processing in the consciousness. I.e. the conflicts caused by Super-Ego rules can be resolved by the use of defense mechanisms.

B. Transferring Psychoanalytic Defense Mechanisms to Software Agents

After having explained the psychoanalytic ideas of defense mechanisms we show how to transfer these ideas to software agents. The main notion is to use defense mechanisms from psychoanalysis to help resolving conflicts during the decision making process in software agents.

The following mapping rules hold for the transfer of defense mechanisms from psychoanalysis to software agents:

- First, input data of a software agent system correlate to human’s perceptions of the environment. In our ARS model [5] we assign the preprocessing of input data to that part which represents the subconscious of our model. During the per-processing, raw input data are bound to associations and expectations from the psychoanalytic Ego.
- Second, inference rules of a software agent system and predefined goals are correlated to human’s action plans and internalized rules of the psychoanalytic Super-Ego. At which, the inference rules are ascribed in our model to the psychoanalytic Ego.
- The third rule concerns humans’ inner drive demands like the demand to eat or to drink. Transferred to software agents or robotics this could be the desire to maintain temperature levels of a robotic system or the desire to recharge batteries of a robot. The inner drive demands are assigned to the psychoanalytic instance Id.

After having transformed psychoanalytic defense mechanisms to software agents, we look for mapping psychoanalytic conflicts to conflicts in the data base and rule base of software agents.

The three conflicting parties in psychoanalysis are Id, Ego, and Super-Ego. In software agents conflicts in the decision unit and/or conflicting action plans can arise from the following components:

- Input data from the environment conflicting with desires of the agent and predefined goals.
- Contradictory inference rules in the planning section of the agent
- Contradictory facts in long-term and short-term memory
- Social rules in multi-agent systems conflict with desires of the agent

C. Defense Mechanisms to Filter Conflicting Goals

To keep a cognitive system up-and-running and to keep a system capable of acting in an inconsistent environment, defense mechanisms of the human mind - adapted and transferred to multi-agent systems - can be of help. Defense mechanisms work, mostly in the subconscious.

In a first step, the defense mechanisms have to detect conflicts. And, in a second step, the detected conflicts have to be resolved by using different defense mechanisms (see Fig 2). One criterion which defense mechanism will be activated is the strength of a drive demand - like hunger or thirst (see section III.D).

In our ARS project [5] we use simple detection rules to detect conflicts of the system. The left side of the rule defines a conflicting condition and the right side of the rule activates a certain defense mechanism.

For example

\[ \text{hungry, sees(food source), sees(enemy) } \rightarrow \text{repress(hunger)} \]  

(1)

Rule (1) means: If an agent is looking for nutrition and the agent sees a food source but at the same time the agent sees an enemy agent who also wants the food and where a fight can be caused, then the agent represses the hunger.

This way, the conflict between the drive demand to eat the food source, and the internalized rule “Do not provoke a fight.” is resolved.
In our ARS software architecture the defense mechanisms are located before the action planning module, shown in Fig. 2. The defense mechanisms (Fig. 2) filter conflicts before they reach the action planning module.

Hence, conflicting action plans do not pass the defense mechanisms and, therefore, the action planning module can easily choose an action plan without having troubles with conflicting rules or plans.

As shown in Fig. 2, the action planning module has a backward loop to perception. This backward loop represents altered expectations in different contexts and different situations. The expectations are used to perform a reality check for perceptions from the environment. (see Fig. 3)

But also, a forward loop from defense mechanisms to action planning is shown in Fig. 2. The forward loop is labeled “Fear”. Using this forward loop, defense mechanisms can send under certain circumstances the signal Fear to action planning. (see Section III.F)

D. Quota of Affect

One important component of the detection rules for conflicts is quota of affect. The quota of affect describes how strong a certain drive demand of an agent is or how strong the agent wants to do a certain action. Over time the quota of affect of a drive demand increases. If the drive demand e.g. hunger is satisfied by eating the quota of affect is decreased.

Taking the quota of affect into account the example detection rule (1) is extended to

\[
\text{hungry}, \text{sees(food source), sees(enemy), quotaOfAffect(hungry)} < 0.4 \rightarrow \text{repress(hunger)} \tag{2}
\]

In accordance to the quota of affect different defense mechanisms can be activated. For example, the detection rule (2) can be altered for a different quota of affect to result in a different defense mechanism - see (3). In (3) the drive demand to eat is stronger than a threshold quota of affect and therefore the defense mechanism reaction formation is chosen (right side of rule (3)).

Reaction formation means to alter the drive aim. Drive aim is for example the drive demand to eat or to drink. Hence, the agent will not repress the hunger like in (1) and (2) but the agent will choose a different drive aim to (virtually) satisfy the hunger. The agent will shift the desire to eat towards a more aggressive attitude and could try to bite the other agent.

\[
\text{hungry}, \text{sees(food source), sees(enemy), quotaOfAffect(hungry)} \geq 0.4 \rightarrow \text{reaction\_formation(hunger)} \tag{3}
\]

E. Tasks of Defense Mechanisms in Software Agents

The following tasks and fields of application of psychoanalytic defense mechanisms in multi-agent systems can be implemented to discharge the decision making unit and to improve the convergence of the decision making process (see Fig. 3).

F. Fear

In our ongoing project ARS [5] we set up a model of the human mind to be implemented in artificial intelligence. The project team of ARS consists of team members from different scientific fields. Among the team members are psychoanalysts, too. Together with them we elaborated an important functionality of psychoanalytic defense mechanisms: to produce fear.

In psychoanalytic theory, fear is produced by the defense if too many perceptions or drive demands are denied or repressed, respectively.

Thus, in our ARS project defense mechanisms have a direct communication channel to the decision making unit where they can send a “fear signal” (see Fig. 2). The fear signal can have three levels: prickle, worriedness, and anxiety. Depending on the sum of quotas of affect of the repressed drive demands, one of the three fear levels is sent to the decision making unit.
If the decision making unit receives, for example, a fear signal of level three, the action plan “flee” becomes the most prioritized for the software agent.

G. Implemented Example

As of now, in our ARS project we have implemented three defense mechanisms: repression, denial, and reaction formation.

The functionality of repression is to transfer drive demands to a Repressed Content Storage where they remain until they are, eventually, re-attached to a similar incoming drive demand. The process of repression sends fear to the decision unit – that is in accordance to psychoanalytic theory, where too much repressed content causes spontaneous fear.

The functionality of denial is to erase unreal perceptions. In contrast to repression the denied content is indeed erased and not immediately stored and retrieved later on.

Reaction formation is a little more sophisticated than the two defense mechanisms repression and denial, mentioned above. Reaction formation alters the drive aim. A drive aim is, for example, the desire to eat. The current drive aim is altered into the opposite drive aim. The psychoanalytic opposite of eat is to bite someone. So the defense mechanism reaction formation alters the drive demand to eat into an aggressive drive demand to bite another agent. The agent will no longer search for food but will be aggressive against other agents.

Considering the above said, we can describe our example implementation in the ARS project where we used the simulation environment MASON. We implemented two software agents, agent A and agent B. Agent A starts to search for nutrition and eventually will find a food source. In our example, agent A approaches the food source but at the same time agent A perceives another agent, agent B.

The decision unit of agent A has now two conflicting action plans: The first one is to eat the food and the second one is: “Avoid enemy agents!”

Here, the defense mechanisms start to act. The defense mechanisms are located before the decision unit and have built-in rules\(^*\), too. In our example, rule (2) from Section III.D holds because the drive demand hunger is not too big, i.e. the quota of affect (see Section III.D) of drive demand hunger is below 0.4. The right side of rule (2) is: \( \rightarrow \) repress(hunger) - see Figure 4.

So, in this situation agent A represses the drive demand hunger. Moreover, the defense mechanism repression sends the signal “fear” to the decision making unit (see Section III.F). Hence, the decision making unit gets the perception “enemy agent” and the signal “fear”. The resulting action plan can be determined without conflicting goals. The inferred action plan is “flee”.

Agent A leaves and in the meanwhile the drive demand hunger increases. The signal “fear” is no longer present and the new action plan is inferred again as “search for nutrition”. Thus, agent A starts again to search for nutrition.

\(^*\) Agent A ends up perceiving the food source again and perceives the enemy agent, as well. This time the quota of affect of the drive demand to eat is above 0.4 (see Fig. 4, right branch). In other words, this time agent A is very hungry. Now, detection rule (3) from Section III.D holds true and the defense mechanism reaction formation is activated to resolve the conflicting action plans for the decision making unit.

The functionality of reaction formation is to shift quota of affect\(^*\) from “life” to “aggression” or vice versa (see above in this Section). In our case, quota of affect is shifted partly from the drive demand “eat” to the drive demand “bite”. Accordingly, we defined in ARS the two drive demands “eat” and “bite” as a pair of opposite drive demands. The drive demand “eat” is not fully omitted but only diminished. It remains active in parallel to the drive demand “bite” which is now stronger.

That means that the defense mechanisms of agent A, this time, do not repress the drive demand hunger but shift quota of affect partly towards aggressive drives. The decision unit of agent A continues to infer the action plan “go forward” towards the food source while being aggressive against agent B. This time, no fear is sent to the decision unit while the goals “eat” and “be aggressive” are not conflicting.

The outcome is: Agent A eats the food and is not afraid of agent B. The quota of affect of the drive demand to eat of agent A is decreased after he has found nutrition.

This example shows how we implemented defense mechanisms in software agents as a preprocessing unit to action planning. Conflicting drive demands and Super-Ego rules are filtered and/or altered by the defense mechanisms to resolve conflicts in the decision making unit of software agents.

\[^*\] Quota of affect see Section III.D

![Diagram](image-url)
IV. CONCLUSION AND OUTLOOK

We explained the functionalities of psychoanalytic defense mechanisms and we showed how psychoanalytic defense mechanisms can be transformed to be used in multi-agent systems. Hence, we gave mapping and transformation rules to transfer defense mechanisms from psychoanalysis to software agent systems. We mapped psychoanalytic conflicts of Super-Ego, Ego, and Id to conflicts in perceptions from the environment, conflicts in the knowledge base, fact base, and rule base. Furthermore, we showed fields of application of defense mechanisms in software agent systems.

We gave example detection rules for conflicts in software agent systems to discover inconsistencies and discover unreal perceptions. Furthermore, we explained the generation and application of “fear” in multi-agent systems by defense mechanisms (see Fig. 2 and Section III.F).

Defense mechanisms in multi-agent systems make the decision and action planning unit more stable by excluding conflicting perceptions, drive demands, rules, and action plans from the decision making process.

Defense mechanisms in cognitive systems can be applied to resolve conflicts caused by inconsistent data bases and inconsistent rule bases in situations, where reasoning or inference engines fail. The convergence of inference algorithms is improved and unreal input data and perceptions are mostly suppressed.

In future developments in our project ARS we plan to implement more kinds of defense mechanisms, other than repression, denial, and reaction formation (see Section III.G). Other kinds of defense mechanisms can lead to a more fine grained alteration of perceptions, rules, and facts and could lead to intelligent error correction in perception and inference.

REFERENCES