Comparison of Technical Filter Mechanisms and Defense Mechanisms of the Human Mind

Why do We Need Psychoanalysis in Artificial Intelligence?

Friedrich Gelbard, Christian Brandstätter, Klaus Doblhammer, Isabella Hinterleitner, Stefan Kohlhauser, Zsofia Kovacs, Heimo Zeilinger
Institute of Computer Technology
Vienna University of Technology
{gelbard, brandstaetter, doblhammer, hinterleitner, kohlhauser, kovacs, zeilinger}@ict.tuwien.ac.at

Abstract—Looking for new paradigms in artificial intelligence, we are investigating functionalities of the human thinking process to manipulate information and filter perceptions. In this paper we introduce defense mechanisms of the human mind to be applied in artificial intelligence. We compare functionalities of defense mechanisms of the human mind with nowadays used filter mechanisms in artificial intelligence and explain reasons why defense mechanisms of the human mind open a broad new spectrum of possibilities and opportunities for artificial intelligence. In particular are these the defense mechanisms repression, displacement, sublimation, projection, disavowal, isolation, separation, depreciation and idealization. These defense mechanisms were chosen and devised with a team of psychoanalysts. We compare state-of-the-art artificial intelligence with psychoanalytical notions in our ongoing ARS project and explain why psychoanalysis is important for future developments in artificial intelligence. Finally, we give examples of similar projects.

Key words: artificial intelligence; bionic inspired intelligence; defense mechanisms; human mind; psychoanalysis

I. INTRODUCTION

New paradigms and methodologies for designing artificial intelligence systems are needed. The use and deployment of artificial intelligence systems in new areas of life and in new technical applications are desired. A promising and broad field of research is what is called bionic inspired artificial intelligence and the striving to model functionalities of the human mind for implementation in artificial intelligence [1]. One important part of the human mind are the so called defense mechanisms. This paper shows which aims can be pursued in artificial intelligence by using functionalities of defense mechanisms of the human mind as filter mechanisms in artificial intelligence systems. The aim of the paper is to bring light into the various kinds of defense mechanisms, to give examples and to compare them to nowadays applied filter mechanisms in artificial intelligence.

In Section II we explain why psychoanalytical notions can lead to new paradigms in artificial intelligence. In Section III we compare statistical implementations in artificial intelligence, predicate logic and neuronal networks with psychoanalytical notions [2] and their possible use cases in artificial intelligence.

In Section IV we give reasons why the functionality of defense mechanisms of the human mind should be applied in artificial intelligence. In Section V we compare psychoanalytic defense mechanisms with filter mechanisms in technical systems. We give examples and propose interesting fields where defense mechanisms of the human mind can be used to fulfill certain tasks. Examples of such tasks are: neglecting data which are not part of the pre-defined goals of the system, converting data to be aim directed or certain functionalities of a technical system where a certain object is regarded as having always good properties or is blocked because it is regarded as having always bad properties.

In Section VI we describe limitations and possible problems of the implementation of psychoanalytic defense mechanisms in artificial intelligence systems. Certain issues of cooperation with psychoanalysts [3] are mentioned and we tried to find out how sufficient psychoanalytic knowledge is on the functionalities of the human mind to be able to implement these notions and to point out which functionalities are described insufficiently and can only be implemented rudimentarily.

Finally, in Section VII we give a short overview of projects which tried or are still trying to find a functional model of the psychoanalytic view of the human mind and which try to implement these findings.

II. WHY DO WE NEED PSYCHOANALYSIS IN ARTIFICIAL INTELLIGENCE?

Psychoanalysis and the model of the human mind described by psychoanalysis are new promising paradigms in artificial intelligence especially if one wants to develop an artificial general intelligent system. The question often posed is: “Are you sure that the psychoanalytical notion will be successful in artificial intelligence?” The answer is: “It is a promising and powerful tool which can be implemented in artificial intelligence systems.” We will give examples and further explanations in Section IV and V. Furthermore, in our ongoing project ARS (Artificial Recognition System) [4] we show ways to develop program code for psychoanalytic notions in artificial
intelligence and to simulate the results in an agent simulation environment.

The point is, psychoanalytical notions combined with artificial intelligence are a path which is entered by many research institutions. The responsibility of these institutions is to describe the path and to explore whether it is successful, where the bottlenecks are, where the big advantages are and how the path can be entered by successor-generations. Why is psychoanalysis in artificial intelligence important? Because psychoanalytical notions can help implementing new algorithms for choosing actions and action plans from a predefined pool, they can help creating data structures and memory structures for ontologies of artificial intelligence systems, they can help filtering incoming data in certain ways (see Section V) and they can help equipping systems with primitive automatisms derived from human drives.

Fig. 1 shows a schematic view of the defense mechanisms. The input data are filtered by the following criteria: filter forbidden action plans (aims), filter forbidden incoming data and prove reliability to other agents. The applied criteria are stored in the internalized rules memory. Examples of internalized rules are:

- Forbidden aims like “It is not allowed to eat other agents.” or “It is not allowed to go to a certain place.”
- Forbidden data are perceptions from the environment which are too dangerous to become conscious and would result in fear.
- Prove reliability: Be honest to other agents and do not “lie” about data or action plans so your offered services are assessed in a positive manner by other agents.

![Defense mechanisms of the human mind.](image)

As of today, ARS is able to fulfill easy standard tasks with its agents in an agent simulation environment like finding nutrition, finding the way through an environment consisting of stones, walls and enemies and avoiding confrontations with enemy agents. Helpful tools are the defense mechanisms of psychoanalysis which will be built in into our ARS system.

Defense mechanisms will help in fulfilling more complex tasks in a complex environment and are one step towards implementing the human cognitive processes in artificial intelligence.

According to our mind, ARS and the psychoanalytical approach open new opportunities in artificial intelligence and in particular in the area of artificial general intelligence. Humans, always, used nature as role model for technical inventions. An important step will be to transfer ARS from a simulation status to a real-world application combined with a robot which can interact directly with humans. This project is already in a planning state.

### III. STATE OF THE ART ARTIFICIAL INTELLIGENCE VERSUS THE ARS PROJECT

Today, artificial problem solving algorithms and programs are intended for a specific field or for a specific group of problems. A general software package to yield artificial general intelligence is not in practical use nowadays and no application is deployed so far. First rudiments are tested in different labs around the world.

Neuronal networks, statistical methods of artificial intelligence, the so called data mining, and logic oriented concepts are capable of fulfilling important artificial intelligence tasks. The problem with these systems is that they need to be developed or at least configured for each task separately. So the desire to develop a universal system which can fulfill different tasks of different complexities arises.

The aim of our ARS project is to build such a general intelligence system. What distinguishes ARS from other general intelligence systems is the strict inherence to the psychoanalytic model of the human mind. One main task of the ARS project is to implement this model in artificial intelligence. This way, we hope to be able to solve complex problems in complex environments.

In the following we will refer to our model for artificial general intelligence, ARS [4], and compare the properties with nowadays used paradigms and methods in artificial intelligence. Advantages and downsides are depicted.

#### A. Neuronal Networks

The strengths of neuronal networks are the adaptability to various tasks and the simplicity of the principle. Compared to our psychoanalytic model of the human mind they, too, try to build a resemblance to the human mind and in addition to the human’s neuronal network. To make our model more human-like a neuronal hardware architecture could support our software model. In this way, neuronal networks can be regarded as an ideal sub-layer for our psychoanalytic model.

We could imagine, that a neuronal hardware architecture would bring us closer to a human-like artificial general intelligence.

So according to our opinion, neuronal networks are a good (hardware) architecture but they lack a meta-layer for artificial general intelligence.

Intel has already developed a hardware architecture for neuronal purposes. This hardware architecture is currently in a testing state.

Compared to our model one can say that the weighted parameters of neuronal networks are similar to our quota of

---

1 Human drives are, for example, hunger or thirst
affect. Quota of affect determines which perception from the environment or which drive is the most important one. In other words, quota of affect sets the focus on specific perceptions or drives. Also, in our model we assign weights and strengths of associations to perceptions and drives which can be seen similar to the weighted functions in neuronal networks.

B. Predicate Logic

Logic based artificial intelligence has the advantage that it is exact, deterministic, the results are traceable, easy to follow and an explanation unit can document and explain exactly how an action plan was developed and which steps were taken to reach a certain goal. This advantage is opposed by rising calculation time and memory usage if the problem becomes more and more complex. Exceeding a certain level of complexity, logic based algorithms are no longer useable because of the rising calculation time they need for the reification algorithms.

Logic based artificial intelligence has few in common with our psychoanalytical model and there is no congruence to our project. Our project, ARS, is designed for complex environments. The developed action plan can depend on little differences of two different situations – similar to the butterfly effect. Therefore, action plans of ARS can vary a lot while the initial situations seem to be very similar. ARS lacks of an explanation component where the steps to reach an action plan are explained and the reasons for taken actions are given.

C. Statistical Artificial Intelligence

Statistical methods such as Hidden Markoff Models or Bayes’ classifier have the ability to cope with input data with several thousand features like fingerprint recognition, face recognition or music style classification. Also the ARS approach should cope with a huge number of input data and a large number of features.

One purpose of ARS is to handle complex situations and reduce and filter input data from the environment.

D. BDI Architecture

The characteristic of a BDI architecture is the selection of a plan from a pool of available action plans to alter the environment. ARS, too, choses from a pool of available action plans. In ARS these action plans are associated to objects, perceptions and drives. In future ARS implementations we intend to provide the possibility to develop new action plans.

The principle of a BDI architecture to alter the environment in accordance to the desires of the agent comes close to ARS notions where drives bring about tensions which lead to actions.

Tab. 1 summarizes the main properties of neuronal networks, predicate logic, statistical artificial intelligence, BDI architecture and our ARS project. One can say that the ARS architecture is the most general architecture of the five compared architectures. Statistical artificial intelligence is nowadays the most promising architecture with regard to practical applications like classification and clustering. Logic and reasoning algorithms are used in semantic web applications.

| TABLE 1: COMPARISON OF STATE OF THE ART ARTIFICIAL INTELLIGENCE WITH OUR ARS PROJECT |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Neuronal networks               | Predicate logic                 | Statistical artificial intelligence (HMM, Bayes, ...) | BDI architecture               | ARS project (based on psychoanalysis) |
| Main applications               | Classification, encoding data, clustering | Problem solving, action plan generation, diagnostics | Feature recognition, classification, clustering | Software agents, selection of action plans | In development status, action plan development, model of the human mind |
| Decision unit                   | Weights of paths                | Resolution calculus              | Arithmetic distance             | Logic based algorithms           | Based on drives and tensions     |
| Explanation unit attached       | Not included                    | Included                         | Not included                    | Included                         | Not included                    |
| Calculation time vs. amount of input data | In several neuronal network applications linear | Amount of input data limited due to rising calculation time | Depending on the chosen classifier linear to the amount of input data | Calculation time increases drastically with the amount of predefined action plans | In the current state only tested with few input data |
| Weighted associations           | Yes                             | No                              | In some way yes                 | No                              | Yes                             |
| Used filter mechanisms for incoming data | Data reduction by inner layers with less nodes | No                              | Sampling rate or resolution of input data can be altered | No                              | Defense mechanisms               |
| Comparable to functionalities of the human mind | distributed architecture, no further similarities | No parallels                     | Statistical methods \(\Rightarrow\) no similarity to functionalities of the human mind | Selection of predefined action plans, BDI theory of human practical reasoning | Tries to implement the functionalities of the human mind |
IV. WHY DO WE NEED DEFENSE MECHANISMS IN ARTIFICIAL INTELLIGENCE?

Defense mechanisms of the human mind can be seen as very powerful and universally usable and applicable filter mechanisms in technical artificial intelligent systems. They are capable of reducing incoming data to a minimum, they are able to direct incoming data towards predefined and internalized aims, and they can neglect certain components of incoming data to be more focused and more concentrated in reaching internalized aims. They can cope with incoming data which are counterproductive and may even have negative impact on the system by altering them or turning them into the opposite.

Furthermore, the advantage of using psychoanalytic defense mechanisms in artificial intelligence is that they open a broad spectrum of new ways, opportunities and insights for artificial intelligence which we do not even forebode now.

From a psychoanalytical point of view, the human psyche consists of three main components – or, as psychoanalysis says – main conflicting parties: The Id, the Ego and the Super-Ego. When these three counterparts are conflicting, the defense mechanisms start to act and try to alter or suppress drives and/or perceptions in order to reduce the conflict and allow a conflict-free perception of the environment and conflict-free inner drives.

So in other words, defense mechanisms of the human mind resolve conflicts which are caused by incoming data in comparison with, by the system, allowed and admitted data. Certain incoming data or actions of the system are not allowed due to, sometimes, safety-reasons, boundaries within a system has to work, functionalities which are possible but not allowed or not allowed in cooperation with other systems because other systems could be damaged or hurt and due to pre-programmed (internalized) rules. Furthermore, a system needs to prove its reliability to other systems or individuals. This is a kind of social component which is used to judge a service or system for its trustworthiness. Therefore, unreliable actions and action plans must be barred in order to gain trust from other systems. The statements above apply, as well, to conflicting outgoing data, actions and action plans.

All these functionalities can be accomplished by defense mechanisms. We think that some reasons why defense mechanisms are not implemented, yet, in practical applications are the complexity of the whole apparatus of defense mechanisms, the various possible application fields, the difficulty to implement a universally usable (program-)structure and the lack of developers’ experience in that field. Defense mechanisms are tested, though, in some research projects.

In the following section we give some insight into the functionalities of defense mechanisms, give some hints of how to use defense mechanisms in technical applications and how to approach implementing defense mechanisms in artificial intelligence systems.

V. WHAT HAVE TECHNICAL FILTER MECHANISMS AND DEFENSE MECHANISMS OF THE HUMAN MIND IN COMMON?

The defense mechanisms of the human mind have been defined by Anna Freud [5] and in rudiments in the work of Sigmund Freud about repression [6]. Fig. 2 shows the defense mechanisms explained in this paper. One important thing is that more than one defense mechanism can be active at a time. That is symbolized in Fig. 2 by having a parallel architecture of the defense mechanisms repression, displacement, projection, disavowal, isolation and separation.

![Figure 2. Defense mechanisms as filter mechanisms in artificial intelligence.](image-url)
reached. These aims are programmed\(^2\) mainly by parents, teachers and the society. Therefore, no “wrong” or forbidden aims are pursued.

An example of displacement is: I want to punch somebody I am angry with but I punch a cushion. An example of sublimation is: I am an egocentric person. This is not regarded a positive trait. So I sublimate my egocentric behavior and become a manager who pursues the interests of his/her company in projects.

C. Projection

The functionality of perception is as follows: If a human cannot cope with a drive or perception the drive or perception can be ascribed to a different individual. If I have fear and cannot cope with my fear, I can project the fear to someone else. So I got rid of my fear and I think now, that the other one is very anxious.

Examples and possible technical applications are hard to find. One possible application in artificial intelligence could be if an agent tells another agent: “No, it wasn’t me who had fear! It was someone else.” So projection needs a higher degree of social and cultural development. We have not reached, yet, such a development with our agents in artificial intelligence.

D. Disavowal

The functionality of the defense mechanism disavowal is to deny reality, i.e. to deny incoming data. Sometimes the denied incoming data are replaced by “better fitting” data.

In artificial intelligence or other fault tolerant systems inconsistent input data are corrected by error detection and error correction automatically. Some protocols simply estimate or approximate data, some have a built in error correction. Like, for example, data read from a DVD are interpolated if data are missing. Data streams from the internet are corrected if errors are detected and a certain amount of errors is not exceeded. Above a certain number of errors the error correction can correct wrongly and alter correct parts of an incoming data package.

In a similar way, disavowal replaces incoming perception data by fantasies which fit better and strengthen the human’s psyche. The human’s perception is so complex that sometimes it is hard to say what is real and what is fantasy.

In childhood disavowal can compensate weaknesses and help support development of the individual. But sometimes, especially in grown up age, disavowal causes alterations of perception and causes serious personality problems. In this case, so to say, wrong error detection is taking place.

E. Isolation

The defense mechanism isolation splits the perception content and the accompanying quota of affect. Only the perception content can pass the defense mechanisms and the quota of affect is blocked. Hence, objects are perceived in a neutral way without any classification into good or bad nor positive or negative.

From a technical point of view one can say that incoming data are not classified. Incoming data are stored as homogenous data without any qualifier. Data are perceived in a neutral way without classifying them. An example could be a technical video surveillance system with face recognition. In normal operation mode the system classifies people into “right to access”, “access denied” and “un-decidable”. The defense mechanism isolation would correspond to an operation mode of the security system where faces of people are perceived but access is granted to everyone and the functionality of the system to judge people is disabled.

F. Separation – Idealization

A similar defense mechanism is separation-isolation. From a psychoanalytic point of view, this defense mechanism neglects all negative qualities of an object or person. The good and bad attributes of a person or object are separated. To the extreme, babies cannot assign bad and good attributes to one and the same object. For example, for a baby the good mother is a different person than the bad mother.

To stay with or example of video surveillance system, separation – idealization means that the security system lets all people pass because they are “good” from the system’s point of view. Maybe a better fitting example would be a firewall which is in installation mode. In this mode the installation program has full access to the resources of the computer because an installation program is considered only to have “good” qualities. One of our next ideas is to implement a computer firewall by using psychoanalytic defense mechanisms.

So one can state that a technical system which runs in a mode where it classifies all incoming data as good data works with the same notion as humans defense mechanism separation – idealization.

G. Separation – Depreciation

The defense mechanism separation – depreciation does exactly the opposite of the defense mechanism separation – idealization. If a human sees all attributes of an object as only bad, this is called depreciation.

Again, technical systems which can be compared to the psychoanalytic defense mechanism separation – depreciation can be seen as system which regards all incoming data from an object as bad. An example would be a firewall which blocks all data of a computer from a foreign domain because this computer is regarded as “bad”.

VI. LIMITATIONS AND OUTLOOK

One disadvantage is still that there are only few research groups and only implementations in testing state of human defense mechanisms in artificial intelligence. So, there are few experiences and rarely available publications on the usability and capabilities of defense mechanisms of the human mind in artificial intelligence. In further projects we will develop a functional model to describe defense mechanisms from a technical point of view, together with help of a team of psychoanalysts.

\(^2\) In psychoanalytical terms programmed rules are called internalized rules.
A second weakness is: Psychoanalytic knowledge of defense mechanisms is in many points diverse, different opinions exist and sufficient knowledge is lacking on some topics to describe psychoanalytic phenomena sufficiently and to describe them from a technical point of view. One important aim of our project is to clarify and systemize psychoanalytic knowledge and to discover grey areas of psychoanalysis.

Concerning ARS, in future projects we will extend our approach to more complex simulated situations of the agents and to more complex environments. As of now, the capabilities of ARS are very limited and the entities, like stone, wall or enemy, used in the simulation environment, are very simple and few in number.

VII. SIMILAR PROJECTS AND DISCUSSION

First of, P. Singh and M. Minsky [7] defined a model of cognitive diversity using agents that can switch between multiple representations and more important multiple layers of reflection. The layers included among basic learned reactions, deliberative thinking, reflective thinking, self-reflective, and self-conscious thinking also innate reactions. Thus, the model contains early “filter mechanisms” for the detection of problems and appropriate reactions on the problems based on instinctive reflexes and learned responses. For example, if one hears a loud noise, one will move to a quieter place to avoid the noise. Finally, the model of Singh and Minsky is consistent with early views of Sigmund Freud, who saw the mind as a system for resolving (or for ignoring) conflicts between our instinctive and acquired ideas.

Furthermore, Buller [8] on a more theoretical level built a similar model on the concept of psychodynamics. Buller postulated a model based on the crucial role of unconscious processes and the existence of mental forces and defense mechanisms. To use the concept of psychodynamics Buller includes the technical interpretation of tension, thoughts, feelings, pleasure, wishes and conflicts. In his architecture he implemented the defense mechanisms of Repression, Projection, Denial, Rationalization and Sublimation.

Forrest [9] explains interesting comparisons of psychoanalytical mechanisms and in particular defense mechanisms with efforts of artificial intelligence to implement psychoanalytic notions. In the section “New Ways of Neural Nets” he compares human learning capabilities to the capabilities of neural nets in artificial intelligence and parallel computing. Forrest describes his ideas from a psychoanalytical point of view followed by examples in artificial intelligence.

To enable the usage of the defense mechanisms in autonomous agents, various mechanisms have been analyzed and a general technical model was developed by Riediger 2009 [10]. Lang 2010 [11] investigates ways of implementing psychoanalytical ideas and also psychoanalytical defense mechanisms in artificial intelligence.

Other important projects in the field artificial general intelligence are [12], a logic based approach, and [13] where emotional states in artificial intelligence are simulated.

Concluding, we can state that psychoanalysis and the comprising defense mechanisms are new aspects and notions in artificial intelligence. We showed in this paper that psychoanalytic concepts can be useful in artificial intelligence. We mentioned examples and application areas of psychoanalysis in artificial intelligence. We explained several defense mechanisms of the human mind and their possible applications in artificial intelligence, and we compared state of the art artificial intelligence concepts to our psychoanalytical ideas and implementations in the ongoing project ARS. Psychoanalysis is a new path in artificial intelligence which needs to be treaded, so successors on this path have landmarks and know what to expect. A path is made by walking. (Lao Tse)

REFERENCES