A Simulation Tool for Training Autistic Reasoning about Intentions

Boris Galitsky
Knowledge-Trail Inc 9 Charles Str Natick MA 01760
bgalitsky@knowledge-trail.com

Abstract

In this study we briefly present a strategy for rehabilitation of autistic reasoning about intention and beliefs, based on playing with mental simulator that is capable of modeling mental and emotional states of the real world. The natural language multiagent mental simulator NL_MAMS, which operates in this world, is introduced, and the rehabilitation methodology is outlined.

Introduction: enabling autistic reasoning with intention and belief

Recent psychological studies have revealed that autistic children can neither reason properly about mental states nor understand emotions (Perner 1991; Pilowsky at al 2000). Autism is characterized by impaired social interaction and communication, combined with repetitive and stereotyped patterns of behavior, and affects up to 1% of school-aged children. Frequently, autistic children do not understand intentions of others. Being able to provide adequate answers to the questions “are you hungry?”, “do you like this toy?”, most autistic children do not react to questions like “am I hungry?”, “would he want this toy?’. Autistic children need to be explicitly taught that other people have intentions.

The theory of mind account, which is most relevant to teaching understanding of intentions of others, refers to the ability to infer and understand what oneself and others are thinking (knowing, believing, desiring) in order to plan ones own behavior and predict the behavior of others. This ability to reason about mental attitudes is impaired in patients with autism (Baron-Cohen 2000, Frith 2001). This reasoning disability leads to difficulties with such mental reasoning-based forms of behavior as pretend play, problems in understanding false beliefs, and the ability to tell lies. In our previous studies we demonstrated that this disability is correlated with corruption of axioms for reasoning about intention and belief (Galitsky 2002, 2005, Peterson et al 2004).

The possibility to teach autistic children theory of mind has been assessed in multiple studies because of potentially important clinical implications. The theory of mind training studies conducted so far have shown that some individuals with autism can be taught to pass the particular tasks of reasoning about mental states (Swettenham 1996, Sutton et al 1999). In most cases, it is natural to assume that trainees indeed apply one or another reasoning pattern rather than memorizing exact answers. Regrettably, in most cases, the studies of how individuals with autism acquire mental reasoning patterns are lacking an accurate formulation of these patterns, backed up by computational experiments. We believe the latter is essential to differentiate between mental and non-mental components of reasoning process.

Our model of the corruption of human reasoning is based on the supposition that there are a number of standard axioms for mental entities; these axioms are genetically set for normal children and are corrupted in the autistic brain. The patterns of corruption vary from patient to patient and are correlated with the specifically outlined groups of autistic children. The simplest of such axiom is *not see, therefore not know*, highlighted in so called Sally-Anne test (Baron-Cohen, 2000). The children with autism have to acquire these axioms explicitly, by means of direct training, using a variety of ways to introduce intentions, beliefs, knowledge and other attitudes of agents and relations between these attitudes. Teaching of such axioms may occur in both formal and natural languages; some of the patients prefer the former over the latter. Frequently, autism is not accompanied by learning disabilities, so the patents willingly participate in training programs of distance learning.

Our practical experience shows that using software-based training allows us to hold the attention of autistic patients for much longer periods than the traditional means of one-to-one treatment by a human. Conventional rehabilitation is always associated with problem of keeping a patient in a responsive mode. Since for the autistic patients the strict rule-based learning is usually much easier than the direct introduction of the various forms of reasoning about mental world, the latter is achieved via the former. The methodology of using artificial mental reasoning systems for rehabilitation is based on teaching autistic children the “mechanical” forms of reasoning about mental world, because the attempts to directly introduce the emotional interaction with the others in a natural manner (e.g., teaching by examples, imitating) frequently fail.

Similar to the theory of mind training settings introduced above, we teach individuals with autism mental entities and their combinations. However, unlike the previously mentioned studies, we use *formalized* means to teach mental entities, suggesting that they are more suitable to the peculiarities of autistic development. We use the non-human (computer) resources, readily acceptable by autistic children, to introduce them to the mental world (of humans) via formalized reasoning. The paradox of our methodology is that reasoning about the mental world, usually supposed to be irrational and displayed as an emotion, can nevertheless be considered from the abstract perspective, formalized and used in training. This hypothesis (Galitsky 2002) is used as a framework of our rehabilitation strategy to develop rational and emotional behavior in the real mental world. Traditionally, strict (formalized, mathematical) thinking is considered as an opposite notion to emotional (fuzzy, approximate) thinking and behavior. Since for the autistic trainees strict rule-based learning is much easier than the direct introduction of the various forms of emotional behavior, the latter is achieved via the former.

A model of intention and belief suitable for teaching autistic children

Below we will present a step-by-step introduction to our formalization intention and belief, adequate with respect to explaining its features to autistic trainees. Suggested rehabilitation exercises follow these steps to provide a consistent and comprehensive educational framework.

Step A. Humans can adequately operate with the set of mental NL expressions containing not more then four embedded mutually dependent mental entities. We constrain the degree of embedding of mental metapredicates to be equal or less than four. The experimental studies of autistic reasoning have shown that both autistic subjects and controls display a monotonic increase in the complexity of mental formulas (Baron-Cohen 2000). For example, autistic children may properly handle the level of embedding *zero* by the age of 6 (you *want*), *one* by the age of 9 (I *know* that you *want*, *know(i, want(you, ...))*, and *two* by the age of 14 for most mental entities (I *want* him to *know* that you *believe*, *want(i, know(him,believe(you,...)))*), whereas controls operate properly with the level of embedding three at the age of 8.

Step B. In natural language (NL), each mental entity has a variety of meanings. In terms of logic programming, there can be multiple clauses defining a mental predicate via the other ones. Absence of such a family of definitions for a mental entity means that all possible meanings are implicitly assumed. Thus the problem of disambiguation in a formal language is posed for those situations where the agents exchange messages in formal language.

In accordance to our experience, it is the first meaning for a mental entity that is hard to perceive and use for an autistic trainee. The consecutive meanings of a familiar entity are memorized and used much easier than the initial one. For example, it takes substantial efforts to introduce the entities of *offending* and *forgiving* to autistic children; they neither discern these actions in others nor reproduce them themselves. As only the first meaning that links these entities together is captured, further separate meanings and associated entities like *reconciling* are perceived more easily, given the respective formal clauses.

Step C. The elementary expression for a mental state is of the form

$m_1(a_1[,a_{1'}], m_2(a_2[,a_{2'}], m_3(a_3[,a_{3'}], m_4(a_4[,a_{4'}], p))))$

where $m_1 \ldots m_4$ are the metapredicates for mental states and actions, occurring with or without negation; m_4, (m_3 and m_4), (m_2 , m_3 and m_4) may be absent; in accordance to Step *C*, the total number of metapredicates is equal or less then four. $a_1 \ldots a_4$ are the *active* agents, square brackets denote the variables for the second agent $a_{1'} \ldots a_{4'}$ (this is the *passive* agent for the mental actions, committed by the *active* agent, denoted by the first argument). For example, an action (and resultant state) with its actor and its receiver is expressed by the metapredicate *inform(Actor, Receiver, Knowledge)*, and an action with two (possibly, symmetric) receivers – by metapredicate *reconcile(Actor, Receiver1, Receiver2, MatterToSettleDown).* The last variable, *MatterToSettleDown,* ranges over an arbitrary formula, which may be a physical, mental, or mixed state. Further on we will assume that mental metapredicates may have additional arguments and will not be showing them explicitly. Finally, *p* is a predicate or expression for physical action or state.

We call such elementary expression for an arbitrary *p* a *mental formula.* It obeys the standard criteria of being a *well-formed* formula.

Step D. The totality of well-formed mental formulas fall into three following categories:

1. A valid mental formula that represents an existing mental state.
2. A mental formula that always holds for any set of agents in any state (an axiom of modal logic, for example *know(Who, know(Who, Knowledge))*).
3. An invalid mental formula that cannot be interpreted. For example, we believe that it is impossible that a person pretends about someone else's mental state $pretend(a_1, a_2, want(a_3, Something))$.

Step E. There are certain syntactic constraints for the formulas describing the mental world that are sufficient to express an arbitrary multiagent scenario. A set of expressions for a mental state has two following components:

1. Mental state fluents, characterizing instant mental states;
2. Mental state clauses, specifying the set of consecutive mental states.

Mental state fluents are expressed with mental formulas as a following conjunction

$$\mathop{\&}_{i=1..n,\, j\in A} m_{i1}(a_{j1}, m_{i2}(a_{j2}, m_{i3}(a_{j3}, m_{i4}(a_{j4}, p))))$$

where $m_{i1} \ldots m_{i4}$ are the metapredicates for mental states and actions, occurring with or without negation; m_{i4} ,(m_{i3} and m_{i4}), (m_{i2} , m_{i3} and m_{i4}) may be absent;

$a_{j1} \ldots a_{j4}$ are the agents from the set of all agents *A;*

For example, *Peter knows that Nick does not know that Peter wants Mike to cook* → *know(peter, not know(nick, want(peter, cook(mike, food)))), m_{11} = know, m_{14} = not know, a_{11} = peter, p=cook(mike, food).* We use '→' to denote a mapping from NL to our formal language.

The following expressions are valid mental formulas to express the continuous mental conditions

p:-$\mu_1 * \dots * \mu_k$

This is a condition for physical action. Here and below $*$ denotes a logical programming conjunction "," or disjunction ";". For example, *Peter would cook if he knows that Nick wants him to do so: cook(peter, food) :- know(peter, want(nick, cook(peter, food))).*

$\mu(\mu_1)$:- $\mu_2 * \dots * \mu_k$

For example, *Mike knows the following: Peter would run away if Mike informs Peter that Nick wants Peter to run away and if Peter does not want to run away himself → know(mike, [run(peter, away) :- inform(mike, peter, want(nick, run(peter, away)))]).*

$\mu(\mu_1$:- $\mu_2 * \dots * \mu_k)$ and $\mu(p$:- $\mu_2 * \dots * \mu_k)$.

S*tep F.* The set of mental metapredicates, formed in accordance with the respective totality of natural language mental entities, is divided into three categories:

- Metapredicates for basic mental states;
- Metapredicates for derived mental states and actions;

Formally independent mental metapredicates that belong to the classes of equivalence of the above category of metapredicates with respect to agents' choice of action, required to reach one mental state from another.

Traditional metapredicates we consider to form the basis are *believe – desire – intention* (BDI), mentioned above. Since all of our mental metapredicates allow multiple interpretation, we merge *desire* as a long-term goal with *intention* as an instant goal to the metapredicate *intend(Agent, Goal)*, where *Goal* $\equiv \mu_1 * \dots * \mu_k$. It allows us to reduce the number of the well-written mental formulas for the analysis of valid formulas (Step *E*).

Step G. The set of available actions for agents is derived from the respective set of natural language entities. For each such entity, we obtain a spectrum of conditions to perform the denoted action based on the family of definitions for this entity in the *intend-know-believe* basis. From the linguistic perspective, the spectrum of meanings for an entity that denotes mental action is determined by the context of this entity (the set of other mental entities in the accompanying sentences). In our model of the mental world, there is a spectrum of clauses for each mental metapredicate such that each clause enumerates particular conditions for mental states. As an example, we present four clauses for *inform*, taking into account that many more clauses are required to form the whole spectrum for this word:

inform(Who, Whom, What) :- intend(Who, know(Whom, What)), believe(Who, not know(Whom, What)), believe(Who, intend(Whom, know(Whom, What))).

The meaning here is as follows: *Who* informs *Whom* about *What* if *Who* intends *Whom* to know it and believes that *Whom* does not know it and intends to know.

inform(Who, Whom, What) :- believe(Who, know(Whom, What)), intend(Who, believe(Whom, know(Who, What))).

The meaning here close to *confirm*: to *inform Whom* that not only *Whom* but *Who* knows *What* as well.

inform(Who, Whom, What) :- ask(Whom, Who, What), intend(Who, know(Whom, What)).

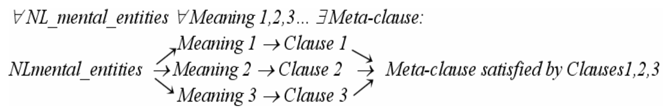
The meaning here is informing as answering.

inform(Who, Whom, What) :- ask(SomeOne, Who, believe(Whom, What)),
intend(Who, know(Whom, What) .

Here informing follows SomeOne's request for informing.

We can define an isomorphism between the NL mental entities and the metapredicates that express the criterion of belonging to the set of clauses for the predicate that we use for this mental entity.

For an example of such mapping, let us consider the set of definitions for the entity inform, presented above. All of



these clauses include the term intend(Who, know(Whom, What)). Let us build the meta-clause that expresses this common feature.

inform (as lexical unit) → the set of clauses { *inform*$_1$:-…, …, *inform*$_k$:-…} → syntactic metapredicate *clauseFor(inform, Clause):- clause_list(Clause, Bodys), member(intend(Who, know(Whom, What)), Bodys).*

The syntactic meta-predicate clauseFor accepts a mental entity to be expressed and a clause for it. The body of this clauseFor predicate expresses the verification that the clause obeys certain criteria, built to express the totality of meanings for this mental entity. We have verified that such isomorphism can be built for almost all mental entities we use in representation of scenarios (Galitsky 2002).

Step H. We define a training scenario as a sequence of mental states of interacting (having mutual beliefs) agents where each transition from mental state m to m+1 is a result of an action of each agent such that this action is logically deduced to be the best from the viewpoint of the respective agent.

Mental Simulator NL_MAMS

The Natural Language Multiagent Mental Simulator NL_MAMS (Galitsky 2004) inputs formal or natural language descriptions of initial mental states of interacting agents and outputs deterministic scenarios of intelligent behaviors of these agents. NL_MAMS simulates the agents which are capable of analyzing and predicting the consequences of mental and physical actions of themselves and others (Galitsky 2006). The output of the NL_MAMS is the sequence of mental formulas, reflecting the states, which are the results of the committed actions (behaviors) chosen by these agents. The simulator settings could be reduced to the game-theoretic ones if the mutual beliefs of agents are complete or absent, and intentions are uniform (a trivial case of multiagent scenario, Rosenschein & Zlotkin 1994).

For the purpose of rehabilitation of reasoning, we target the synchronous interaction model where the agents commit their actions simultaneously. These actions update their mental and physical states at the same time as well.

The sequence of committed actions (or skipping a move) and resultant states is the same for each agent. Decision-making of agents in our settings is primarily concerned with choice of actions to achieve desired states; agents have immediate and long term goals of mental and physical states, and sometimes explicit intentions of actions.

To choose the best action, each agent considers each action it can currently perform. Firstly, each agent selects a set of actions it can legally perform at the current step (physically available for the agents, acceptable in terms of the norms, etc.). Such an action may be explicitly intended or not; also, this action may belong to a sequence of actions in accordance with a form of behavior which has been chosen at a previous step or is about to be chosen. In the former case, the agent may resume the chosen behavior form or abort it.

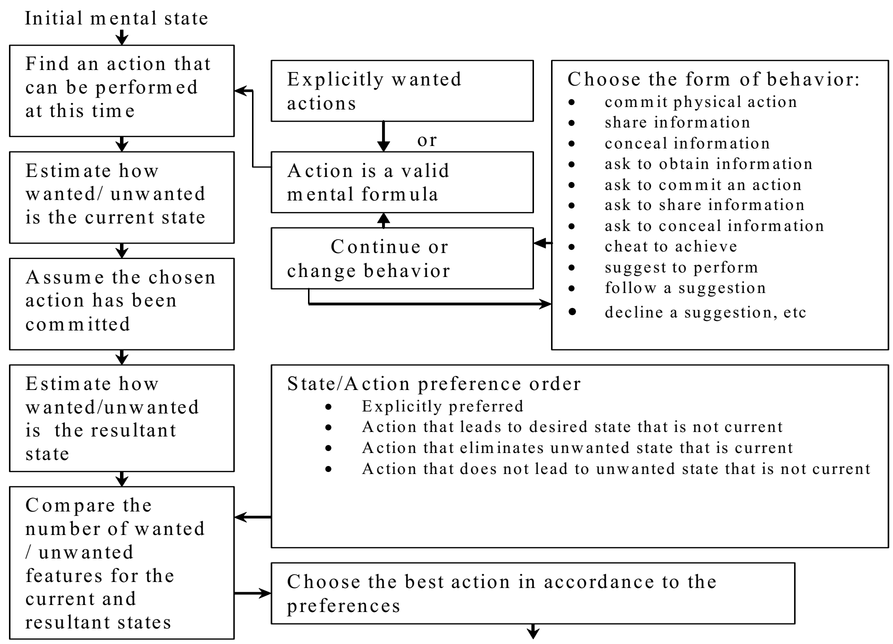


Fig.1: The chart for the choice of action, involving own agent capabilities and world knowledge. To reason for others, the agents of NL_MAMS load the world knowledge of respective agents and make decisions for them according to the same chart.

Having a set of actions which are legal to be currently performed, the agent applies a preference relation. This relation is defined on states and actions and sets the following order:

1. Explicitly preferred (intended) action.
2. The action that leads to a desired state that is not current.
3. Action that eliminates an unwanted state that is current.
4. Action that does not lead to an unwanted state that is not current.
5. Action that does not eliminate an intended state that is current.

We refer the reader to Fig.1 and (Galitsky 2004) for the further details on NL_MAMS, and to (Galitsky 2006) for the reasoning model which is implemented in NL_MAMS.

Training environment for teaching reasoning about intention and belief

A few versions of the web-based user interface for NL_MAMS have been developed for a number of environments, including describing of mental states of scene characters. A variety of interface components were designed for specifying mental states, including natural language and drop-down box-based.

The one-to-one rehabilitation strategy (NL_MAMS – independent), conducted by a member of rehabilitation stuff, includes the following components:

1. direct introduction of the basic mental entities *intend-know-believe* using real-world examples (Fig.2);
2. explanation of derived mental entities using the basis *intend-know-believe* ;
3. introduction of the derived mental entities by means of real-world examples (Fig.2);
4. conversations that heavily rely on a discourse with mental focus (Lovaas 1987);
5. conversations that are based on a pictorial representation of interaction scenarios (Fig.4);
6. involving the trainees into actual interactions with other children and asking them to verbally represent these interactions;
7. encouraging the parents and rehabilitation personnel to demonstrate a special awareness of mental entities in the real world (Galitsky 2000, Galitsky 2001);
8. "picture-in-the-head" and "thought-bubbles" techniques, using "physical" representation of mental attitudes (Swettenham et al 1996, Fig.4).

NL_MAMS-based training is intended to assist in all of the above components. Initially a trainer shows how to represent mental states from the above components via NL_MAMS, and discusses yielded scenarios with a trainee. The plausibility and appropriateness of actions yielded by NL_MAMS require special attention from trainees. Then the trainer specifies other initial mental states and asks a trainee to come up with plausible scenarios originating from these mental states. Trainees are children with high-functioning autism 6-10 years old, selected so that they are capable of reading simple phrases and communicating mental states in one or another way. After a certain number of demonstrations, the trainees are encouraged to use NL_MAMS independently, applying it to real-world mental states the trainees have experienced, as well as abstract mental states. Trainees are presented with both natural language and structured input and output of NL_MAMS, and they are free to choose their favorite way of user interface.

For example, an exercise introducing the mental action of *offending* and *forgiving* is depicted at Fig. 3. This is a partial case of NL_MAMS training of yielding a scenario given an initial mental state: it is adjusted to the definition of *offending*. Expected resultant scenario is just the actions of *offending* or *forgiving* with appropriate parameters for agents and subjects of these actions. These parameters are specified via drop-down boxes; their instances are expected to show the trainees how to generalize the instances of *offending* or *forgiving* towards different agents. Also, multiple ways to express these generalizations are shown: *friend, parent, brother/sister, they/them, he/she, him/her* etc. After the trainees learn how to derive a single-step scenario for a fixed mental action, they are given the tasks to compose scenarios with two or more mental actions they have already learned.

Training of Mental Reasoning using NL_MAMS

Specify the agents' and simulator parameters

There are friends Simon, Mike, and Bob.
Simon wants Bob to get the pizza.
Simon knows that the pizza is bad.
Simon wants that both Simon and Bob believe that the pizza is not bad.
Simon believes that Mike does not know that the pizza is bad.
Simon knows that initially Bob does not know that the pizza is bad.
Simon wants Mike to inform Bob that the pizza is not bad.
Simon will confess to other parties that the pizza is bad if he believes that Mike informed Bob that it is the case.
Simon believes that if Bob beleives that the pizza is not bad then Bob would get it.
Simon would start believing that when Bob is informed by Mike that the pizza is not bad then Bob would get it.
Mike wants Bob to get a pizza.
Mike does not want Simon to know that Mike wants Bob to get a pizza.
Mike believes that Bob would get the pizzas if Bob believes it is not bad.
Mike believes Bob would get a pizza if Simon informs Bob that the pizza is not bad.
Mike wants Simon to believe that Bob gets the pizza if Mike informs that Bob that the pizza is not bad.
Bob wants to get the pizza if Mike informs Bob that the pizza is not bad.
Bob wants to get the pizza if Mike informs Bob that the pizza is worth geting.
Bob wants Mike to believe that Bob will not get the pizza.
Bob wants Simon to believe that Bob believes that the pizza is not bad.

Load (translate into formal representation)

See the statements and select/deselect them to run the simulator

intend(s, buy(b,p))
know(s, pb)
intend(s, believe(b, not pb))
believe(s, not know(m, pb))
know(s, not know(b, pb))
intend(s,m, inform(m, b, not pb))
inform(s, _, pb) :- believe(s, inform(m, b, pb))
believe(s, (buy(b,p) :- beleive(b, not pb)))
believe(s, (buy(b,p):- inform(m,b, not pb))):- inform(m,b, buy(m, p))
intend(m, not know(s, intend(m, buy(b,p))))

Build the behavior

Check the generated scenario

Step	Simon	Mike	Bob
1	inform(s,b,not pb)	inform(m,b,buy(m,p))	ask(b,m,intend(m, buy(b,p)))
2	inform(s,m,not pb)	inform(m,s,(buy(b,p) :- inform(m,b,not pb)))	inform(b,s,buy(b,p))
3	suggest(s,b,buy(b,p))	skip	inform(b,m,not buy(b,p))
4	Skip	cheat(m,know(s,intend(m,buy(b,p))))	inform(b,s,believe(b,not pb))
5	Skip	suggest(m,b,buy(b,p))	suggest(b,m,inform(m,s,not intend(b,buy(b,p))))
6	Skip	skip	**buy(b,p)**

Fig.2: The NL_MAMS user interface for rehabilitation of autistic reasoning (advanced scenario).

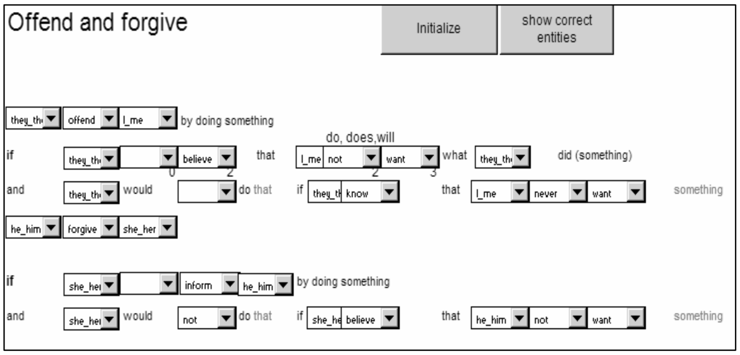


Fig.3: A form to introduce a mental entity (here, to offend and to forgive). After an entity is explained via examples and verbal definition is told, a trainee is suggested to choose the proper basic entities with negations when necessary to build a definition.

Fig. 4: A visualization of interaction scenarios. “Picture-in-the-head” and “thought-bubbles” techniques are used based on “physical” representation of mental attitudes (on the top-left). “Hide-and-seek” game (on the bottom-left). The children are asked the questions about who is hiding where, who wants to find them, and about their other mental states. In the middle: On the right: a trainee wrote a formula for “forcing to commit an unwanted action” and the respective natural language conditions for it.

Conclusions

The main contributions of the project discussed in this paper is the model of reasoning about mental world, with the special focus on intentions, and the means for children with autism to learn this model. The following components have been developed in the course of this project:

1. The theory of mind is subject to a formal treatment from the standpoint of logical artificial intelligence;
2. The possibilities of the theory of mind teaching are re-evaluated, taking into account the developed formal framework for reasoning about the mental world;
3. Appropriateness of formal reasoning as an educational means and associated cognitive issues are assessed;
4. The model of mental world is constructed to serve as a basis for education means;
5. The simulation-based hybrid algorithm of deriving consecutive mental states is implemented;
6. Theory of mind teaching using this software is evaluated on a short-term and long-term basis;
7. Implications for the practical rehabilitation strategies are analyzed;

The evaluation of this approach showed primarily that teaching autistic patients to reason about intentions and beliefs of others is possible, and it should be assisted by a software system capable of reasoning in respective domains, such as NL_MAMS.

References

Baron-Cohen, S. 2000. Theory of mind and autism: a fifteen year review. In S. Baron-Cohen, H. Tagar- Flusberg, and D. J. Cohen Eds. Understanding Other Minds, Volume A, pp. 3–20. Oxford, U.K.: Oxford University Press.

Frith, U. 2001. Mind blindness and brain in autism. Neuron 32, 969–979.

Fry, H., Ketteridge, S. and Marshall, S. 1999. A Handbook for Teaching & Learning in Higher Education: Enhancing Academic Practice. RoutledgeFalmer, London.

Galitsky, B. 2002. Extending the BDI model to accelerate the mental development of autistic patients. *Second Intl. Conf. on Development & Learning*. Cambridge, MA.

Galitsky, B. 2004. A Library of Behaviours: Implementing Commonsense Reasoning about Mental World. KES 2004 LNAI 3215 pp. 307-313.

Galitsky, B. 2005. Distance Learning Rehabilitation of Autistic Reasoning. Encyclopaedia of Distance Learning. Idea Group Publishing.

Galitsky, B. 2006. Reasoning about attitudes of complaining customers. *Knowledge-Based Systems* Volume 19, Issue 7, November 2006, 592-615.

Lovaas, O. I. 1987. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Clinical and Consulting Psychology* 55, 3–9.

Perner, J. 1991. Understanding the representational mind. *MIT Press* Cambridge MA.

Peterson, D., Galitsky, B. and Goldberg, S. 2004. Literal handling of conflicting default rules leads to inadequate reactions of autistic patients. Seventh Intl. Conference of Cognitive and Neural Systems, Boston MA.

Rosenschein, J. & Zlotkin, G. 1994. Rules of Encounter: Designing Conventions for Automated Negotiation Among Computers. Cambridge, MA: MIT Press.

Swettenham, J. 1996. Can children with autism be taught to understand false belief using computers? Journal of Child Psychology and Psychiatry, 37, 157-165.

Sutton J.; Smith P. K.; Swettenham J. 1999. Social cognition and bullying: Social inadequacy or skilled manipulation? British Journal of Developmental Psychology, September 1999, vol. 17, no. 3, pp. 435-45016.