# Agents Modeling Agents: Incorporating Ethics-Related Reasoning

**Sergei Nirenburg and Marjorie McShane**[1]

**Abstract.** We describe CLAD, an implemented advisor system in the domain of clinical medicine. CLAD assists a human physician in making decisions about diagnosing and treating patients. CLAD monitors the transcript of an ongoing dialog between the physician and a patient, builds and augments a mental model of the patient and suggests courses of action to the physician. CLAD can also explain its decisions and describe its understanding of the beliefs (including ethics-related beliefs), goals, plans, personality traits, biases and other features of the patient – both directly observed ones and obtained through CLAD's own reasoning processes. The paper includes a detailed analysis of several examples of CLAD operation that illustrate the interaction between mindreading and moral judgment.

## 1 INTRODUCTION

If autonomous intelligent agents are to collaborate in ever more sophisticated ways with humans and other agents, they must be endowed with an increasingly encompassing computational theory of mind – not only their own mind, but the minds of others as well. Such a theory of mind will rely not only on knowledge directly available through channels of perception, but also on modeling agents' internal – that is, unobservable – beliefs about the world, with "beliefs" understood as knowledge for which the agent has less than full confidence. Creating and using beliefs about other agents' unobservable characteristics allows an agent to engage in sophisticated behavior such as detecting other agents' motivations, predicting their future behavior in specific situations, and tracing the biases and ethical considerations contributing to their decision-making. An agent armed with the ability to reason about others can also turn the same capabilities inward, supporting metacognition about its own behavior. An agent generates beliefs on the basis of inputs from its stored knowledge, stored beliefs, and results from its perception processes.

Modeling other agents, or "mindreading," is broadly accepted as an important scientific task for cognitive systems. Thus, according to Bello [1], "One of the key features of any complete computational theory of human cognitive architecture is a process-level explanation of how it represents and reasons about the contents of others' minds. This key question is driving a host of research projects in social neuroscience, developmental psychology, linguistics, philosophy of psychology and, more recently, in computational modeling of cognition… [M]aintaining representations of others' beliefs and having them be available to our practical reasoning system (e.g. planning, action-selection etc.) afford us faster socio-cognitive computations, and thus the ability to be more effective teammates or competitors."

[1]University of Maryland Baltimore County {sergei, marge}@umbc.edu

This paper belongs to the area of computational modeling of cognition and discusses select aspects of the theory of mind under development for the OntoAgent environment. In this paper we illustrate the modeling and use of unobservable agent characteristics with the help of examples from the current implementation of OntoAgent. The examples demonstrate that ethical considerations can be successfully incorporated into OntoAgent with no modifications to its control structure, simply by expanding the inventory of agents' unobservable features (such as character traits, preferences, susceptibility to biases, etc.) whose values are used by OntoAgent's general-purpose decision-making module. This is a promising finding because it obviates the need to introduce a separate modeling strategy specifically for moral reasoning. Moreover, our examples illustrate how ethical reasoning can be seamlessly integrated with other decision-making needs of an agent. This reflects our desire to investigate ethics issues, as it were, not as a separate task but in competition with other decision-making considerations. The former option was chosen in the pioneering work of Anderson and Anderson (e.g., [2]) that concentrates on modeling the seven *prima facie* duties of Ross [3]. We would like also to consider cases where no decision is ethically correct (though some may be deemed more correct than others); where different agents hold different opinions on ethics; where agents choose to follow a course of action that is not the best from the ethical standpoint; etc. We also concentrate on building "mindreading" agents that will be evaluated not only on the basis of choices that they themselves make but also on the basis of how successfully they interpret actions of other (artificial or human) agents, including the ethical component of these actions. An additional goal of the discussion is to show the feasibility of practical reasoning systems based on the proposed theory of mind, its associated theories (e.g., the theory of ontological semantics for language processing), and the knowledge bases supporting all of the above.

## 2 ONTOAGENT

Initial implementations of OntoAgent are in the domain of clinical medicine. This led to the early introduction of simulated embodiment [4,5], making OntoAgent agents "double agents", in that they have a cognitive side and, optionally, a physiological side. The cognitive agent – on which we focus here – engages in perception, reasoning and action. Currently supported modes of agent perception in OntoSem are language understanding and interoception, which is the interpretation of bodily signals generated by physiological simulation. Results of perception are interpreted by the language and interoception processors using the same metalanguage as is used in the specification of the agent's memory. Then these new memories are stored in the agent's ontology and fact repository (memory of assertions). In this paper we do not address language learning in OntoAgent.

That issue is discussed in [6]. As an example of a metalanguage structure used to populate agent memory, consider an agent's interpretation of another agent's utterance *I'm scared*, happening on April 7, 2012. Small caps show ontological concepts; indices differentiate instances.

```
FEAR-FR22
   DOMAIN HUMAN-FR71 ;the result of reference resolution of "I"
   RANGE  .8         ;on the {0,1} scale; terrified would be 1
   TIME   (ABSOLUTE-DAY 7) (ABSOLUTE-MONTH APRIL)
          (ABSOLUTE-YEAR 2012)
```

As concerns the speaker, it will remember and store in its own fact repository (a) the meaning representation of the feeling itself and (b) the fact that it generated the corresponding speech act, with the identity of the hearer noted. The architecture of an OntoAgent agent is shown in Figure 1 and explained in the caption. Many aspects of OntoAgent and its current prototype applications, Maryland Virtual Patient and Clinician's Advisor, have been reported, for example: physiological simulation for virtual patients [4,5,7]; cognitive modeling and decision making [8,9]; agent memory management [10, 11]; agent metacognition [12]; agent learning [6]; dialog modeling [13]; and semantically-oriented language processing for intelligent agents [14,15].

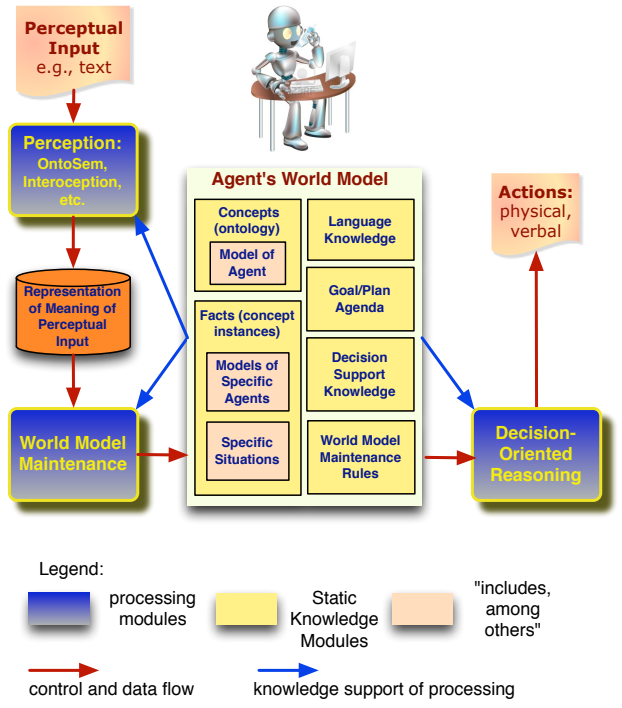


**Figure 1.** OntoAgent agent architecture, including: [center] an agent's world model; [left] the inputs that contribute to that model – the interpretation of results of perception and the operation of "world model maintenance" functions (responsible, among other things, for maintenance of unobservable features; and [right] the types of agent action the model supports – decision-making that can lead to physical or verbal action.

In this paper we focus on the dynamic building and use of "Models of Specific Agents" and "Specific Situations" in the application called CLAD: CLinician's ADvisor. CLAD seeks to improve the decision making and reduce the cognitive load of practicing clinicians by providing targeted, motivated decision support during interviews with live patients.

CLAD's ability to understand clinical and dialog situations is supported by its being equipped with mental models of clinicians and their patients and means for updating and maintaining these models. During its work with a particular clinician C over time, CLAD enhances its model of C by including in it a model of each patient $P^{C}_{i}$ through C's eyes, that is, CLAD's beliefs about C's knowledge and beliefs about P. CLAD uses these "models of others" in conjunction with its own knowledge and beliefs to suggest decision-making strategies to the clinician.

## 2 THE EXAMPLES

To illustrate the OntoAgent approach to calculating, recording and using unobservable features of others to support an agent's own reasoning, we have selected examples that have relevance to the issue of ethics in computer systems. We agree with McLaren [16] that it is not appropriate for an intelligent agent to take responsibility for ethical decisions; rather, the most it can do is support the decision-making of humans who must accept that responsibility. Consider two typical situations in which CLAD might be called upon to assist clinicians:

1. A patient refuses a recommended intervention. CLAD can assist the clinician in convincing the patient by (a) attempting to determine why the patient refused and (b) offering patient-specific argumentation strategies. CLAD does not enter into the debate of whether or not a physician should force his opinion on a patient. (CLAD independently decides whether or not it believes the advice was the best available advice in the first place; if it disagrees with the advice, it flags the clinician about that separately.)

2. The clinician presents the patient with a prognosis. CLAD evaluates whether it is within reasonable bounds of accuracy or if the clinician might be making an error in judgment. CLAD does not enter into the debate of whether or not overly optimistic prognoses (intended to leverage placebo effects) are clinically justified in principle.

As an example of the first situation, consider a patient (P) with acute appendicitis who refuses a life-saving appendectomy. Table 1 presents some of the reasons CLAD knows about for refusing surgery, framed as **beliefs** held by P, as well as some **clues** in favor of each analysis.

CLAD may or may not have information about these clues stored in $P^{C}_{i}$ (its model of the clinician's model of $P_i$). If it does have such clues, it can hypothesize about which belief is leading to the patient's decision. This hypothesis alone might be sufficient to help a tired, frazzled, rushed – in general, cognitively overloaded – clinician to steer the conversation with P in a useful direction. However, if the clinician wants more help, CLAD can suggest the best argumentation strategy by combining known methods to address the belief (see the 'How to…' column in Table 1) with known features of P that affect the choice (see the 'Influencing features…' column in Table 1). For example, if P is a Christian Scientist with low medical

sophistication but high intelligence, the clinician might engage him in a debate about the details of Christian Science, whose theology does not actually require refusal of medical intervention. If, by contrast, P is a Christian Scientist with a high fear of death, low intelligence and low medical sophistication, the clinician might better choose to focus on the statistical likelihood of dying – that is, if the clinician decides that it is ethically appropriate to try to change the patient's mind to begin with.

| Belief | Clues | How to Address Belief? | Influencing Features of P |
|---|---|---|---|
| **The body can heal itself.** | P has said this.<br>P doesn't have regular check-ups, vaccines, etc. | Statistical likelihood of dying.<br>Physiological explanation. | P's medical sophistication. [Fig. 2]<br>P's cultural background.<br>P's level of fear. |
| **Surgery is too dangerous.** | P has said this.<br>P's relative died in surgery.<br>P has rejected indicated surgery in the past. | Statistical likelihood of dying.<br>Break down aspects of surgery, discuss risks of each. | P's medical sophistication.<br>P's level of fear.<br>P's foci of fear. |
| **Christian Science.** | P has said this.<br>P has never accepted a medical intervention.<br>P has no medical history. | Statistical likelihood of dying.<br>Physiological explanation.<br>Primer on Christian Science theology. | P's medical sophistication.<br>P's level of fear.<br>P's foci of fear.<br>P's intelligence. |

**Table 1.** A subset of the appendectomy decision space.

A natural question is, how does CLAD acquire the beliefs stored in $P^C_i$? Some beliefs can derive from direct evidence: e.g., P's chart says he is a male, and CLAD assumes – with maximal confidence – that the clinician believes that. Other beliefs are derived through CLAD's reasoning. In the current version of CLAD, this reasoning is carried out using decision functions that take as input evidence from sources as varied as the patient's own statements, the patient chart, and the interpreted transcript of doctor-patient conversations. For example, the property MEDICAL-SOPHISTICATION is referred to in Table 1 as a feature influencing CLAD's decision making. Its value is calculated by CLAD using a stored decision function whose input parameters include EDUCATION-LEVEL, VOCABULARY-CHOICE, QUESTION-SOPHISTICATION and QUESTION-FREQUENCY. Each of these input parameters is assigned a time-dependent value, with CLAD assessing and recording its confidence in this assignment. Figure 2 illustrates the subset of $P^C_i$ devoted to P's medical sophistication.

The last three types of evidence contributing to calculation of the value for MEDICAL-SOPHISTICATION all rely on functions that take as input the text-meaning representations (TMRs) of the doctor-patient interviews, which are generated by CLAD using the OntoSem language analysis system [13]. The decision function for VOCABULARY-CHOICE estimates the level of vocabulary use by the patient based on word length and comparisons with available dictionaries of difficult words; the function for QUESTION-SOPHISTICATION measures the content-oriented sophistication of the patient's questions based on how often they seek information about *how* or *why* some medical event happens, and how many words in questions are mapped to the medical subtree of the ontology; the value of QUESTION-FREQUENCY is a function of the average number of questions per visit, both direct (*Will it hurt?*) and indirect (*I suppose I won't need to take this medicine very long*). Naturally, estimating the sophistication of questions asked to the doctor is the most difficult, and therefore least confident, calculation.

HUMAN-301
- MEDICAL-SOPHISTICATION .8
  - CONFIDENCE .8
  - EVIDENCE
    - EDUCATION-LEVEL 1
      - CONFIDENCE 1
      - EVIDENCE INTAKE-QUESTIONNAIRE
    - VOCABULARY-CHOICE 1
      - CONFIDENCE .8
      - EVIDENCE TMR OF D-P INTERVIEWS
    - QUESTION-SOPHISTICATION .5
      - CONFIDENCE .5
      - EVIDENCE TMR OF D-P INTERVIEWS
    - QUESTION-FREQUENCY 10
      - CONFIDENCE .8
      - EVIDENCE TMR OF D-P INTERVIEWS

**Figure 2**. CLAD's belief about the doctor's belief about the medical sophistication of the patient.

Another way for the agent to infer property values of other agents is through static correlations among values of features comprising an agent model that we hypothesize might have predictive power. A sampling of such correlations is shown in Table 2. This table also gives a sample of the many kinds of features contained in CLAD's models of agents.

Assume that CLAD is faced with a decision whose function requires knowledge of an agent's level of optimism, and assume that, at the moment, CLAD does not have an explicit value for this property stored in its model of that agent. It can choose to estimate the agent's level of optimism (Col. 1) based on previous evidence of it being happy or depressed (Col. 2), making overly optimistic or pessimistic prognoses (Col. 3), and/or accepting challenges or avoiding risky behavior (Col. 4). CLAD's confidence in its estimation of the value of OPTIMISM depends on the amount of evidence available in its memory. An interesting question that we will not pursue in this short space is when (how often, at what junctures) it is appropriate for an agent to make generalizing conclusions about features of other agents like overall optimism, or having a high susceptibility to jumping to conclusions. Decisions like this are handled by the World Model Maintenance Engine shown in Figure 1.

Let us return now to the second class of CLAD functionality we use for illustration: CLAD helping the clinician to avoid incorrect prognoses. Making prognoses about things like the likelihood of a medication benefitting a particular patient is a tricky business. Clinical trials can provide information such as "medication efficacy: 50%." This means that for any given patient, this medication has an equal chance of being effective

and ineffective. If we consider that a positive attitude can positively impact healing, then a clinician might be justified in saying, "Of course it will work!" If, by contrast, we consider that offering false hope might cause a patient to lose trust in his doctor, a more coolly objective prognosis might be justified. The question then is: how can an intelligent assistant be useful to a clinician making prognoses? We suggest at least two ways. On the one hand, since our existing predictive physiological models were developed through an intensive effort by expert clinicians, they permit CLAD to make more specific prognoses than a clinician can be expected to make on the fly under the time pressure of an office visit (see [4,5]). On the other hand, CLAD can offer opinions about the extent to which overly optimistic or pessimistic decisions are justified based on parameter values found in its models of the patient and the clinician as well as in its knowledge about the objective medical situation. Let us consider a specific example of the latter in more detail.

| Static Traits | Related Transient States | Related (Susceptibility to) Biases | Related Preferences |
|---|---|---|---|
| robust ↔ fragile | fresh ↔ tired | [none] ↔ depletion-effects & cognitive-overload | take-on-more-work ↔ avoid-more-work |
| optimistic ↔ pessimistic | happy ↔ depressed | overly-optimistic-prognosticating ↔ overly-pessimistic-prognosticating | accept-challenges ↔ avoid-risky-behavior |
| analytical ↔ impulsive | concentrated ↔ rushed | [none] ↔ jumping-to-conclusions & small-sample-bias | postpone ↔ act-now |
| confident ↔ insecure | decisive ↔ indecisive | illusion-of-validity ↔ cognitive-overload | convince-others ↔ let-others-decide-for-themselves |
| empathetic ↔ aloof | engaging-another ↔ keeping-distant | tendency-to-conceal-bad-news ↔ strictly-"like-me"-reasoning | close-relations ↔ distant-relations |
| extroverted ↔ introverted | chatty-mood ↔ terseness | [none] | long-conversations ↔ short-conversations |

**Table 2.** Property correlations that help to fill out values of an agent model. ↔ indicates a scale whose end points are indicated. Features before and after ↔ correlate across columns: e.g., [optimistic, happy, overly-optimistic-prognosticating and encourage-others] are related.

In any given situation, CLAD can combine its general knowledge of medicine with known features of the patient to arrive at its objective prognosis for a medication's efficacy for the patient. As an initial simplification, CLAD has three available values for a medication's likely efficacy: *unlikely to work, might work, very likely to work*. Assume that the objective prognosis in a given situation is *might work*, and assume that the clinician tells the patient *Surely it will work!* Is this a misrepresentation (possibly a breach of ethics), a clinician error (e.g., due to an incorrect use of statistics) or a clinically justified decision on the part of the clinician ("My patient needs some hope and good news.")?

CLAD can attempt to trace the clinician's reasoning for presenting an overly optimistic hypothesis using a function that considers certain features of the clinician, the patient, and the clinician-patient relationship. If the reasoning seems justified, then CLAD will conclude that the exaggeration was intentional and will throw no warning.

The features of interest in this decision are shown in Table 3, which compares two different patients who have the same physiological profile but different character traits and different relationships with the doctor. The objective prognosis – *it might work* (based on "likelihood of treatment success for this patient: .5") – is always available and appropriate if the clinician chooses it. What CLAD needs to understand is whether an exaggeration like *Surely it will work!* is justified for either of these patients.

| Feature | Source of value | Patient-A Value | Patient-B Value |
|---|---|---|---|
| Likelihood of treatment success at population level | function (literature, clinician's experience) | .5 | .5 |
| Likelihood of treatment success for this patient | function (population-level success, patient-specific features) | .5 | .5 |
| Best score of other available options | function, for each other available option (population-level success, patient-specific features) | .2 | .2 |
| Overall optimism of clinician | CLAD's past memories of clinician behavior | 1 | 1 |
| Confidence of clinician | as above | 1 | 1 |
| Clinician's knowledge about treatment | as above | .8 | .8 |
| Personal relationship between clinician and patient | as above | *1* | *.1* |
| Patient's medical sophistication | See Fig. 1 | *.1* | *.8* |
| Patient's need for encouragement | as above | *1* | *.2* |
| Patient's susceptibility to encouragement. | as above | *1* | *.2* |
| APPROPRIATE PROGNOSIS: Likelihood of success of treatment is: | function (all of the above values and their confidences) | *.7 - .9 (Surely it will work!)* | *.5 (It might work.)* |

**Table 3.** Sample calculations of clinically justified prognoses for 2 patients.

As the table shows, this prognosis is more easily justified for Patient-A than for Patient-B. Patient-A has low medical

sophistication (he is unlikely to know anything about the medication), a great need for encouragement and a high level of susceptibility to the doctor's encouragement; in addition, the doctor knows this patient well and feels justified in stretching the truth to fulfil these needs. By contrast, Patient-B is medically sophisticated and might know a lot about the medicine or look up that information later; and the doctor does not know him very well and does not sense any particular need for encouragement. As such, there is no clear justification for exaggerating the likely efficacy of the medication and it is better to report maximally objective prognosis. If a clinician should tell Patient-A that the medication will surely work, CLAD will interpret that as a clinically justified exaggeration, but if he should tell Patient-B the same thing, CLAD will throw a flag in case the clinician spoke in error. As mentioned above, CLAD's role is to assist the clinician in avoiding errors by trying to understand his reasoning; it is not CLAD's place to have an opinion about the judiciousness of exaggerating a prognosis to the positive.

## 3 INTERIM CONCLUSIONS

The development of an explanatory theory of mind that can be realized computationally in intelligent agents capable of being accepted as members of teams consisting of agents and humans (and not just efficiency tools like calculators or internet search engines) is arguably the most forward-looking ambitious research program in computer applications today. To be explanatory, this theory must, among other desiderata, account for reasons underlying agents' behavior. We believe that this is best done through the introduction of descriptive mental models that use parameters representing directly unobservable features of agents. This is a long-term project. But it is not too early to discuss how to model the moral stance of agents and what connections this stance has to the agent's theory of mind. Indeed, if such agents are to be accepted as team members by humans, then they will be expected to ne endowed with ethics – otherwise they would not be trusted by the human team members and would not be able to reason about other agents' motivations or explain their actions.

In this paper we illustrated the approach to the theory of mind in the OntoAgent environment, specifically concentrating on ethics-oriented features and situations. We discussed two levels of decisions – how to extract values for a variety of directly unobservable parameters and how to make decisions about a) other agents' beliefs and intentions and b) the agent's own actions on the basis of these parameter values. We extended the inventory of features (ontological properties) for modeling agents to include ethical considerations and so far have not found any problems with treating these features in exactly the same manner as other agent features.

In our work we followed the path established in [2] and concentrated on developing an advisor system, CLAD, that is constrained to clinical medicine. Unlike the Andersons, we also incorporated elements of mindreading in our agents, as a result of which CLAD's advising activity seeks to model human decision making in social environments where agents must model and take into account the "inner world" of other agents. This latter capability allows for modeling different ethical theories and different points of view within a single computational testbed.

A central task in the ongoing development of the OntoAgent theory of mind is enhancing the world model maintenance engine, including maintenance of ever more sophisticated models of self and other agents and the interaction between this task and language processing and other perception engines. A major component of this task boils down to knowledge acquisition for specific applications, and we intend to continue acquiring relevant knowledge by working with domain experts and by interpreting the findings of psychological experiments suggesting certain generalizations about human behavior. We will also continue our application-building efforts for the purposes of testing and validating the theoretical hypotheses.

At the same time, we will continue our work on formulating the theory of agent's mind as such. While at this point it is premature to offer a comprehensive description of this theory, we intend to formulate one in the near future.

## REFERENCES


[1] Bello, P. 2011. Shared Representations of Belief and Their Effects on Action Selection: A Preliminary Computational Cognitive Model. Proceedings of the 33rd Annual Conference of the Cognitive Science Society.

[2] Anderson, M., S.L.Anderson and C. Armen. An Approach to Computing Ethics. 2006. IEEE Intelligent Systems, July-August.

[3] Ross, W.D. 1930. **The Right and the Good**. Clarendon Press.

[4] McShane, M., Fantry, G., Beale, S., Nirenburg, S. and Jarrell, B. 2007. Disease interaction in cognitive simulations for medical training. Proceedings of MODSIM World Conference, Medical Track, 2007, Virginia Beach, Sept. 11-13 2007.

[5] McShane, M., Nirenburg, S., Beale, S., Jarrell, B. and Fantry, G. 2007. Knowledge-based modeling and simulation of diseases with highly differentiated clinical manifestations. 11th Conference on Artificial Intelligence in Medicine (AIME 07), Amsterdam, The Netherlands, July 7-11, 2007.

[6] Nirenburg, S., McShane, M., Beale, S., English, J. and Catizone, R. 2010. Four kinds of learning in one agent-oriented environment. Proceedings of the First International Conference on Biologically Inspired Cognitive Architectures (BICA), Arlington, VA, Nov. 13-14.

[7] McShane, M., Nirenburg, S. and Beale, S.. 2008. Two Kinds of Paraphrase in Modeling Embodied Cognitive Agents. In Proceedings of the Workshop on Biologically Inspired Cognitive Architectures, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7-9.

[8] Nirenburg, S., McShane, M., and Beale, S. 2008. A Simulated Physiological/Cognitive "Double Agent". In Proceedings of the Workshop on Naturally Inspired Cognitive Architectures, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7-9.

[9] Nirenburg, S., McShane, M., and Beale, S. 2009. A unified ontological-semantic substrate for physiological simulation and cognitive modeling. In Proceedings of the International Conference on Biomedical Ontology, University at Buffalo, NY, July 24-26, 2009.

[10] McShane, M., Nirenburg, S. and Beale, S. 2011. Reference-related memory management in intelligent agents emulating

humans. Proceedings of AAAI Fall 2011 Symposium on Advances in Cognitive Systems.

[11] McShane, M., Jarrell, B., Fantry, G., Nirenburg, S., Beale, S. and Johnson, B. 2008. Revealing the conceptual substrate of biomedical cognitive models to the wider community. Medicine Meets Virtual Reality 16, ed. J. D. Westwood, R. S. Haluck, H. M. Hoffman, G. T. Mogel, R. Phillips, R. A. Robb, K. G. Vosburgh, 281 – 286.

[12] Nirenburg, S., McShane, M., and Beale, S. 2010. Aspects of metacognitive self-awareness in Maryland Virtual Patient. Proceedings of the AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, Nov. 11-13, Arlington, VA.

[13] McShane, M. and Nirenburg, S. 2009. Dialog Modeling Within Intelligent Agent Modeling. Proceedings of the IJCAI-09 Workshop on Knowledge and Reasoning in Practical Dialog Systems, Pasadena, California, USA, July 12, 2009, pp. 52-59.

[14] Nirenburg, S. and Raskin, V. 2004. Ontological Semantics. Cambridge, Mass.: The MIT Press.

[15] McShane, M., Nirenburg, S. and Beale, S. Ms. Meaning-Centered Language Processing. Book-length manuscript, submitted.

[16] McLaren, B. 2006. Computational Models of Ethical Reasoning: Challenges, Initial Steps and Future Directions. IEEE Intelligent Systems, July-August.