Biologically Inspired Cognitive Architectures (2013) 3, 39–50

Available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Modeling decision-making biases

Marjorie McShane [a,*], Sergei Nirenburg [a], Bruce Jarrell [b]

[a] University of Maryland Baltimore County, United States
[b] University of Maryland School of Medicine, United States

Received 5 August 2012; received in revised form 17 September 2012; accepted 17 September 2012

KEYWORDS
Cognitive modeling;
Medical decision-making;
Intelligent agents;
Biases;
Agent reasoning

Abstract
Human decision-making can be affected by cognitive biases, and outside observers can often detect biased decision-making in others. Accordingly, intelligent agents endowed with the computational equivalent of the human mind should be able to detect biased reasoning and help people to improve their decision-making in practical applications. We are modeling bias-detection functionalities in OntoAgent, a cognitively-inspired agent environment that supports the modeling of intelligent agents with a wide range of sophisticated functionalities, including semantically-oriented language processing, decision-making, learning and collaborating with people. Within OntoAgent, different aspects of agent functionality are described using micro-theories that are realized as formal computational models. This paper presents the OntoAgent model that supports the automatic detection of decision-making biases, using clinical medicine as a sample application area. It shows how an intelligent agent serving as a clinician's assistant can follow the doctor—patient interaction and warn the doctor if it appears that his own or the patient's decisions might be unwittingly affected by biased reasoning.

Introduction

''Cognitive bias'' is a term used in the field of psychology to describe distortions in human reasoning that lead to empirically verified, replicable patterns of faulty judgment. Cognitive biases result from the inadvertent misapplication of a necessary human ability: the ability to simplify complex problems, make decisions despite incomplete information, and generally function under the real-world constraints of limited time, information, and cognitive capacity (cf. Simon's (1957) theory of bounded rationality). Contributing factors to cognitive biases include, non-exhaustively: over-reliance on one's personal experience as heuristic evidence; misinterpretations of statistics; overuse of intuition over analysis; acting from emotion; the effects of fatigue; considering too few options or alternatives; the illusion that the decision-maker has more control over how events will unfold than he or she actually does; overestimation of the importance of information that is easily obtainable over information that is not readily available; framing a problem too narrowly; and not recognizing the interconnectedness of multiple decisions (for further discussion see, e.g., Kahneman 2011; Korte 2003).

* Corresponding author.
E-mail address: marge@umbc.edu (M. McShane).

2212-683X/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bica.2012.09.001

The effects of cognitive biases on decision-making can be difficult to recognize and, if recognized, can remain difficult to overcome. As Heuer (1999, Ch. 9) writes: ''Cognitive biases are similar to optical illusions in that the error remains compelling even when one is fully aware of its nature. Awareness of the bias, by itself, does not produce a more accurate perception. Cognitive biases, therefore, are, exceedingly difficult to overcome''. However, the fact that a problem is difficult does not absolve us from responsibility for solving it: biased thinking can have highly detrimental consequences. We hypothesize that errors in judgment caused by some cognitive biases could be reduced if intelligent advisors were able to detect potentially biased decisions and generate explanatory alerts to their human collaborators. The bias-detection capabilities described in this paper have been developed for CLinician's ADvisor (CLAD), a system implementing an intelligent agent that monitors the interaction of a physician with a patient and helps to lower the physician's cognitive load by offering intelligent and timely advice.

CLAD is an application of OntoAgent (McShane, Beale, Nirenburg, Jarrell, & Fantry, 2012), a cognitive architecture whose configurable agents, general-purpose knowledge bases, reasoning engines and modeling strategies are, in principle, domain-independent. Work on specific applications seeks both to create useful, practical applications and to test and validate the overall OntoAgent approach. Although CLAD has not yet been subjected to formal evaluation, the work presented here stands on its own as a scientific contribution because: (a) the models illustrate a new *theory* of computationally modeling bias detection; (b) the models are realized by comprehensive *descriptions* — sufficiently generic to be implemented within other cognitive architectures — along with the knowledge required for implementation; and (c) these descriptions conform to the requirements of the *decision making engine* already used by CLAD. The bias-related functionalities we address, and the psychological phenomena they target, are summarized in Table 1.

Table 1 Psychologically motivated functionalities for an advisor agent in clinical medicine.

Advisor functionalities	Targeted decision-making biases
Supplying facts the clinician requests using text generation, structured presentation of knowledge (e.g., checklists), process simulation, etc.	• Depletion effects • Jumping to conclusions • Overconfidence
Automatically detecting and flagging potential **clinician biases**	• The illusion that more features is better • False intuitions • Jumping to conclusions • The small sample bias • Base-rate neglect • The illusion of validity • The exposure effect
Automatically detecting and flagging potential **patient biases**	• The framing sway • The halo effect • The exposure effect • Effects of evaluative attitudes

The remainder of the paper is organized as follows: The Background section presents a brief overview of OntoAgent and its current proof-of-concept medical systems. The Results section describes the new bias counteraction module (being introduced for the first time here), including its psychological rationale, knowledge structures and processing algorithms. The Discussion section further contextualizes the methodological choices of OntoAgent and considers issues of evaluation for cognitively-inspired agent systems.

Background

OntoAgent is a knowledge-based intelligent agent environment inspired by the traditional goals and motivations of artificial intelligence: attempting to achieve human-level behavior by modeling agents with human-like capabilities of perception, reasoning and action. The OntoAgent conception of agency is ideologically — though not methodologically — close to the belief-desire-intention (BDI) model (Bratman, 1999); but unlike the classical BDI implementations (Wooldridge, 2000), the OntoAgent approach centrally involves sophisticated language processing, physiological simulation, and a complex integration of language processing and reasoning. Fig. 1 shows the basic knowledge and functionalities of an OntoAgent agent.

As Fig. 1 shows, OntoAgent agents can undergo two types of *perception*: (1) interoception, which is the experiencing of signals generated by physiological simulation of the agent's body, and (2) language understanding, which involves a large battery of pre-semantic and semantic analysis engines. The results of processing input from both modes of perception are formal knowledge structures written in the unambiguous, ontologically grounded metalanguage described by the theory of Ontological Semantics (Nirenburg & Raskin, 2004). Depending on their content, the knowledge structures are stored in the appropriate knowledge base: ontology, for general world knowledge; fact repository for episodic memories; or lexicon, for newly learned words and phrases. Agent *reasoning* is carried out at dozens of levels, from the many processes involved in deep natural language understanding to the processes involved in memory management to the manipulation of plans and goals. Agent *action* includes mental actions, like updating memory; verbal actions, like engaging in dialog with a user; and simulated physical actions, like taking medicine or showing up for a doctor's appointment.

The OntoAgent paradigm adheres to the following two hypotheses: (1) **functional verisimilitude** — artificial agents that model (simulate) human perception and decision-making capabilities have a better chance for success than computer programs that do not attempt such modeling; and (2) **feasibility** — it is possible at the present time to develop artificial agents (including advisors, tutors, virtual patients and more) that demonstrate useful, non-toy human-level capabilities on a realistic budget and over a realistic period of time. Feasibility is fostered by the theme-and-variations

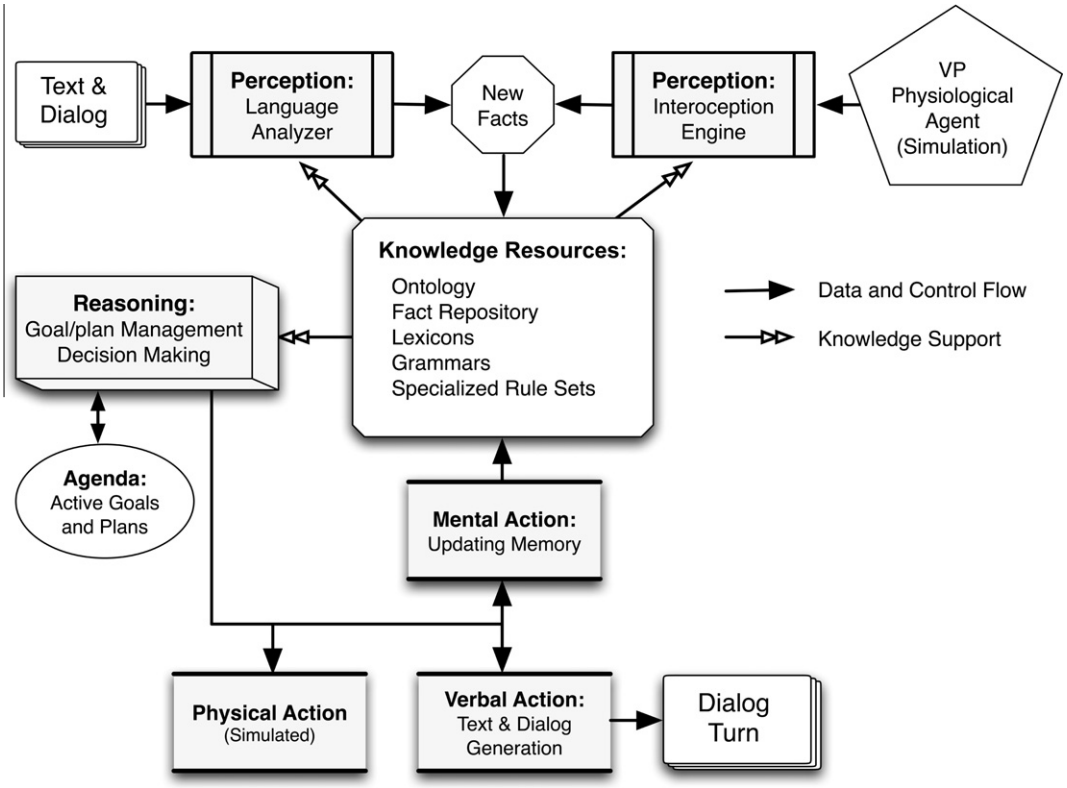

Fig. 1 Agent modeling in the OntoAgent environment.

nature of OntoAgent agents: by selecting different values for key parameters describing an agent's body or mind, we can configure a large inventory of usefully differentiated agents, such as virtual patients with different predispositions, diseases and personalities, and virtual advisors with different specializations and advice-giving strategies.

The two current proof-of-concept applications of OntoAgent are Maryland Virtual Patient (MVP) and CLinician's ADvisor (CLAD). MVP seeks to foster the development of decision-making skills in physicians in training by providing them with the opportunity to diagnose and treat a society of virtual patients in open-ended simulations, with the optional support of a virtual tutor. The virtual patients include all of the components of the generic agent shown in Fig. 1, whereas the tutor lacks a body (a physiological simulation), since it would be superfluous to its role in the system. (The tutor maintains a model of the body of the patient, which is necessary for its decision making.) CLAD uses the same knowledge bases, modeling strategies and engines as MVP but in a configuration aimed at improving the decision making of practicing clinicians. In CLAD, the core artificial agent is the advisor that provides targeted, motivated advice to a live clinician treating a live patient. As CLAD follows the doctor—patient interaction — currently, through entries to the patient chart but eventually by processing the full dialog as well — it creates an ever more specific model of both the doctor and the patient, and uses those when deciding what if any advice to give to the clinician.

The OntoAgent development approach forces us to conduct active research and development in a broad spectrum of areas including natural language processing, knowledge modeling, reasoning, decision-making, physiological simulation, agent memory management, metacognition and learning. Select publications include (McShane, Fantry, Beale, Nirenburg, & Jarrell 2007; McShane & Nirenburg, 2012; McShane et al., 2012; Nirenburg & McShane, 2012; Nirenburg, McShane, & Beale 2008; Nirenburg & Raskin, 2004); others can be found at http://www.trulysmartagents.org.

Supporting bias avoidance

In this paper we report on (1) the *theory* of modeling cognitive support for bias avoidance, which involves the selection of parameters and values to be treated (e.g., bias types), detection heuristics, decision functions, and knowledge support; and (2) the descriptive *realization* of the theory as a set of models compatible with those implemented in existing OntoAgent agents. We consider three types of bias-avoidance support (Table 1): (1) providing memory support (the need for which is predicted by the theory of decision-making biases) in response to user-initiated queries; (2) automatically detecting potential biases in clinicians' decision-making; and (3) automatically detecting potential biases in patients' decision-making.

Memory support

Memory lapses — which are unavoidable in clinical medicine due to the large amount of knowledge which physicians must possess and use — underlie three types of decision-making biases: **depletion effects** (the effects of fatigue),

jumping to conclusions and **overconfidence**. All of these could, we believe, be decreased with timely, ergonomically presented reminders, cribs and checklists (cf. Gawande, 2009) that reflect selected aspects of the knowledge already available in the static resources of OntoSem's expert models. This type of cognitive assistance is user-initiated, meaning that the user must recognize his or her own potential to misremember or misanalyze something in the given situation, as might happen under conditions of sleep deprivation (Gunzelmann, Gross, Gluck, & Dinges 2009). Consider just a few illustrative situations in which CLAD's knowledge could be leveraged to counter clinician memory lapses.

Example 1. A clinician is tired and forgets some basic ontological properties of a disease or treatment. He queries CLAD with an English string such as, *What are the symptoms of GERD?* CLAD interprets this input using the OntoSem text analyzer, converting it into the following unambiguous, ontologically-grounded structure, simplified for presentation. (Ontological concepts are in small caps; numerical suffixes indicate concept instances.)

REQUEST-INFO-1	
AGENT	PHYSICIAN-1
THEME	GASTROESOPHAGEAL-REFLUX-DISEASE.CAUSES-SYMPTOM-1

This meaning representation says that this input is requesting the fillers of the CAUSES-SYMPTOM property of a concept instance of GASTROESOPHAGEAL-REFLUX-DISEASE. The agent answers the question by looking up the needed information in its ontology, the relevant portion of which is shown below:

GASTROESOPHAGEAL-REFLUX-DISEASE	
CAUSES-SYMPTOM	CHEST-PAIN, DYSPHAGIA, HEARTBURN, NAUSEA, REGURGITATE

Example 2. A clinician wants to order the test called EGD (esophagogastroduodenoscopy) but forgets what preconditions must hold to justify this. He queries CLAD with *What's needed to diagnose achalasia?* As before, CLAD converts the input into the ontological metalanguage and understands that the answer will be the filler of the property SUFFICIENT-GROUNDS-TO-DIAGNOSE in the ontological concept ACHALASIA. Table 2 shows a subset of properties of the ACHALASIA concept that relate to diagnosis and treatment. For presentation purposes, the property values in the right-hand column are presented in plain English, not in the ontological metalanguage.

Example 3. A clinician knows that the disease GERD can have different manifestations in different patients but forgets the details and asks CLAD to display the disease profile for GERD. Disease profiles are human-interpretable snapshots of the disease models that underlie the physiological simulations of virtual patients in OntoAgent. Each disease profile includes: (a) the relevant physiological properties and symptoms whose values change over time, with parametric variation across patients; (b) which diagnostic tests can be run and which physiological feature values they will detect; and (c) which treatment procedures can be launched, along with the range of outcomes expected for patients at different stages of the disease. Most disease models are presented as a series of tables with feature types in the left-hand column and their values at different time intervals shown in subsequent columns. For example, Table 3 shows just a few physiological properties and symptoms of the disease achalasia during the first few conceptual stages (t0, t1, t2) of the disease. Ranges of values show typical variation among patients, with default values in parentheses. In a computer simulation of the physiological functioning of a virtual patient, property values are interpolated linearly: e.g. a starting lower esophageal sphincter (LES) pressure of 25 torr will rise to 49 by the end of the t2 stage.

The OntoAgent disease profiles are useful pedagogically because they make transparent key aspects of the disease-modeling strategy to system users. Just as importantly, the user interface permits authorized users to create new instances of virtual patients for use in simulations.

Example 4. A patient asks about his prognosis but the clinician is too tired, or too rushed, or not familiar enough with the disease to provide a well-motivated one. CLAD can help by permitting the clinician to run a simulation of the patient by creating a virtual patient with all the known current and historically attested property values of that

Table 2 Four clinical properties of the esophageal disease ACHALASIA, with values written in plain English for readability.

PROPERTY	Values (rendered in English for readability)
SUFFICIENT-GROUNDS-TO-DIAGNOSE	All three of the following conditions: 1. Either bird's beak (a visual test finding) or hypertensive lower esophageal sphincter (LES) 2. Aperistalsis 3. Negative esophagogastroduodenoscopy (EGD) for cancer (i.e., a pertinent negative)
SUFFICIENT-GROUNDS-TO-SUSPECT	Either: 1. Dysphagia (difficulty swallowing) to solids and liquids 2. Regurgitation
SUFFICIENT-GROUNDS-TO-TREAT	Definitive diagnosis
PREFERRED-ACTION-WHEN-DIAGNOSED	Either: 1. HELLER-MYOTOMY (a surgical procedure) 2. PNEUMATIC-DILATION (an endoscopic procedure)

Table 3 An excerpt from the disease model for achalasia.

	Property	Start value	t0	t1	t2	...
PHYSIO	Stage duration (in days)	180—720 (360)	180—720 (360)	180—720 (360)	180—720 (360)	...
	Ratio of contracting to relaxing neurons in the distal esophagus	100/100	80/100	60/100	40/100	...
	Basal LES pressure (in torr)	0—40 (25)	P_{start} + 8	P_{t0} + 16	P_{t0} + 24	...
SYMPT	Difficulty swallowing distal solid (on the abstract scale of 0—1)	0	0	.5—1 (.5)	1—2 (2)	...
	Do liquids stick?	No	No	No	No	...
	Chest pain (scale: 0—1)	0	0	0—.3 (.1)	0—.5 (.3)	...

particular human patient. This simulation can show the likely progression of the disease if left untreated as well as suggest the potential effectiveness of available interventions if administered at different times. More specifically, CLAD will use whatever property values are known about the patient to maximally constrain the profile of the virtual patient corresponding to the human patient, and it will leave unknown property values underspecified, by listing a range of values statistically representing the population. The simulation will be run using the selected prognosis date as the final point.

The above four examples should suffice to convey our main point: the knowledge structures and simulation capabilities *already developed for the MVP and CLAD applications* can be directly reused to help clinicians to counteract memory lapses due to the biases called **depletion effects, jumping to conclusions** and **overconfidence**. For this category of phenomena, developing models that take into account biases involves anticipating the requests of users and optimizing the presentation of the already-available knowledge to make it easily interpretable by clinicians. The initiative for seeking this class of bias-avoidance support lies in the hands of users. By contrast, solutions for the remaining two groups of phenomena will proactively seek to detect decision-making biases on the part of both participants in clinician—patient interactions.

Detecting clinician biases

Diagnosing a patient typically consists of two stages: based on a patient interview and physical examination, the clinician posits a hypothesis, then he or she attempts to confirm

Table 4 Select test results and symptoms for the disease achalasia.

		t1	t2	t3	t4
TESTS	EGD	dilated esophagus & no tumor at the GI junction		dilated esophagus & ***narrowing and pop upon entering LES (= hypertensive LES)*** & retained debris & ***no tumor at the GI junction***	
	Esophageal Manometry	IR-LES* & high-normal LES pressure	IR-LES & high-normal or hypertensive LES & intermittent peristalsis	IR-LES ***& hypertensive LES & aperistalsis***	
	Barium Swallow	delayed emptying		***bird's beak*** & dilated esophagus & retained debris & retained barium	
SYMP	Dysphagia, solids	.5 – 1 (.7)	1-2 (1.5)	2-3 (2.5)	3-4 (4)
	Do liquids stick?	No	yes	yes	yes
	Chest pain	0 - .3 (.1)	0 - .5 (.3)	.3 - .8 (.5)	.5 – 1 (.7)

*IR-LES is incomplete relaxation of the LES

it through medical testing or trial therapy (e.g., medication, lifestyle change). Confirming a hypothesis by testing leads to a definitive diagnosis, while confirming a hypothesis by successful therapy leads to a clinical diagnosis. Unintentionally biased decision-making by the clinician can happen at any point in this process.

The ''need more features'' bias. When people, particularly domain experts, make a decision, they tend to think that it will be beneficial to include more variables to personalize or narrowly contextualize the decision. As Kahneman (2011, p. 224) writes, ''...Experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity. Simple combinations of features are better''. This observation was first made in Paul Meehl's highly influential work of 1954 (reprinted as Meehl, 1996) that compared statistical predictions to clinical judgments and found the former to consistently outperform the latter. A recent review of Meehl's work (Grove & Lloyd, 2006) concludes that his findings have stood up to the test of time.

A key point at which clinicians might erroneously — and at great expense — believe that more feature values are necessary is during diagnosis: e.g., they might not recognize that they already have sufficient knowledge to diagnose a disease. For many diseases, clear diagnostic criteria exist, like that shown in the first row of Table 2. If the patient chart shows sufficient evidence to diagnose a disease, but the clinician has not posited the diagnosis and has ordered more tests, CLAD can issue an alert about the possible oversight.

Jumping to conclusions. The opposite of seeking too many features is jumping to conclusions: e.g., diagnosing a disease without sufficient evidence. Typically, each disease has a constellation of findings that permit a clinician to definitively diagnose it. For example, the disease achalasia can be definitely diagnosed by the italicized test results shown in the t3 and t4 stages of Table 4. The additional values in those columns — like the patient's symptoms — provide still further evidence for this disease but cannot be used in lieu of the definitive constellation of findings. Positing a diagnosis prior to obtaining the full set of *definitive* values could be incorrect. Whenever a clinician posits a diagnosis, CLAD will double check the patient's chart for the known property values and issue an alert if not all expected property values are attested.

False intuitions. Without entering into the nuanced debate about the nature and formal validation of expert intuition — as pursued, e.g., in Kahneman & Klein, 2009 — we define skilled intuition as the recognition of constellations of highly predictive parameter values based on sufficient past experience. Nobody can have reliable intuitions about unknowable situations *or* in the absence of reliable feedback *or* without sufficient experience.

We can operationalize the notion of ''intuition'' in at least two ways. The simpler way is to leverage only and exactly the knowledge recorded in tables like 2 and 4; this assumes that the expert clinician expertise recorded in these tables covers the full spectrum of intuition-based expectations, which is clearly an oversimplification. A more sophisticated approach would be to incorporate CLAD's knowledge of the past history of the physician into its decision-making about the likelihood that the clinician is acting on the basis of false intuition. If a clinician has little past experience, then CLAD will be justified to flag seemingly false moves. However, if a clinician who has vast past experience with patients of a similar profile starts to carry out what appears to be an unsubstantiated move, CLAD might better query him or her about the reason for the move and potentially learn this new constellation of findings and their predictive power. This is an aspect of ''system-initiated learning by being told'' that is one of the foci of our work.

The illusion of validity. The illusion of validity describes a person's clinging to a belief despite evidence that it is unsubstantiated. Kahneman (2011, p. 211) reports that the discovery of this illusion occurred as a result of his practical experience with a particular method of evaluating candidates for army officer training. A study demonstrated that the selected method was entirely non-predictive — i.e., the results of the evaluation had no correlation with the candidate's ultimate success in officer training — but the evaluators still clung to the idea that the method was predictive because they believed that it *should* be predictive.

The illusion of validity can be found in clinical medicine when a physician refuses to change an early hypothesis despite sufficient counterevidence — as by rerunning tests, or continuing a failed medication trial. The definition of sufficiency depends on (a) the strength of the constellation of features suggesting the diagnosis, as indicated by the knowledge in tables like 2 and 4; (b) the strength of the constellation of features suggesting a *different* diagnosis, recorded in corresponding tables for other diseases; and (c) the trustworthiness of tests, whose error rates are recorded in the ontology. CLAD detects overzealous pursuit of a hypothesis using decision functions that combine these three factors. The decision functions are implemented using the method described in Nirenburg et al., 2008.

Base-rate neglect is a type of decision-making bias that, applied to clinical medicine, can refer to losing sight of the expected probability of a disease for a given type of patient in a given circumstance. For example, a patient presenting to an emergency room in New York is highly unlikely to have malaria, whereas that diagnosis would be very common in sub-Saharan Africa. Although physicians are trained to think about the relative likelihood of different diagnoses, *remembering* all of the relative probabilities given different constellations of signs and symptoms can be quite challenging. CLAD can help with this by flagging situations in which a doctor is pursuing a diagnostic hypothesis that is unlikely given the available data.

For example, esophageal carcinoma can result from gastroesophageal reflux disease (GERD) but typically only if GERD is not sufficiently treated for a long time and if the person smokes, drinks alcohol, lives or works in an industrial environment or has had exposure to carcinogenic materials. These likelihood conditions are recorded in the ontology as complex fillers for the property SUFFICIENT-GROUNDS-TO-SUSPECT for the disease ESOPHAGEAL-CARCINOMA, as pretty-printed below:

```
ESOPHAGEAL-CARCINOMA
   SUFFICIENT-GROUNDS-TO-SUSPECT
   Both
   — (GERD (EXPERIENCER PATIENT-1) (DURATION (>5 (MEASURED-IN YEAR)))
   — At-least-one-of
      — (PATIENT-1 (AGENT-OF SMOKE))
      — (PATIENT-1 (AGENT-OF (DRINK (THEME ALCOHOL) (FREQUENCY
        (>.3)))))
      — (PATIENT-1 (AGENT-OF (RESIDE (LOCATION INDUSTRIAL-PLACE))))
      — (PATIENT-1 (AGENT-OF (WORK (LOCATION INDUSTRIAL-PLACE))))
      — (PATIENT-1 (EXPERIENCER-OF (EXPOSE (THEME CARCINOGEN)
        (FREQUENCY (>.3)))))
   Other conditions...
```

If a clinician hypothesizes esophageal carcinoma for a 20-year old person with a 3-month history of GERD, CLAD will issue a warning that there appears to be insufficient evidence for this hypothesis, and it will show the clinician the conditions under which the hypothesis is typically justified.

The small sample bias. A person's understanding of the frequency or likelihood of an event can be swayed from objective measures by the person's own experience, and by the ease to which an example of a given type of situation — even if objectively rare — comes to mind (Kahneman 2011, p. 129). The small sample bias can lead to placing undue faith in personal experience. For example, if the widely preferred medication for a condition happens to fail one or more times in a physician's personal experience, the physician is prone to give undue weight to those results — effectively ignoring population-level statistics — and prefer a different medication instead. This is where the ''art'' of medicine becomes fraught with complexity: while personal experience should not be discounted, its importance should not be inflated since it could be idiosyncratic. As Kahneman (2011, p. 118) writes, ''The exaggerated faith in small samples is only one example of a more general illusion — we pay more attention to the content of messages than to information about their reliability, and as a result end up with a view of the world around us that is simpler and more coherent than the data justify''.

CLAD will automatically detect the small sample bias in clinicians' decisions by comparing three things: (1) the clinician's current clinical decision, (2) CLAD's memory of the clinician's past decisions when dealing with the particular disease, and (3) the objective, population-level preference for the selected decision compared to other options. For example, suppose three of the clinician's recent achalasia patients did not respond to the standard procedure for treating achalasia or developed complications. If the clinician then stops recommending that procedure and, instead, opts for a less preferred one, CLAD will issue a reminder of the population-level preference for the first procedure and point out there is a danger of a small-sample bias (i.e., reliance on the availability heuristic). (Of course, the actual reason for the switch in procedures might be due a difference in practitioner quality for the two procedures — an eventuality we are modeling as well.)

The exposure effect. The exposure effect describes people's tendency to believe frequently-repeated statements even if they are false because, as Kahneman (2011, p. 62) says, ''familiarity is not easily distinguished from truth''. This is biologically grounded in the fact that if you have encountered something many times and are still alive, it is probably not dangerous (ibid, p. 67). CLAD will detect potential cases of the exposure effect using a function whose arguments will include the following:

- A new ontological property, HYPE-LEVEL, applied to interventions (drugs and procedures), whose values reflect the amount of advertising, drug company samples, etc., to which a clinician is exposed; if unknown for a particular clinician, a population-level value will be used, based upon the amount of overall advertising and sample distribution practices.
- The objective ''goodness'' of an intervention, as compared with alternatives, at the level of population, which is a function of its relative efficacy, side effects, cost, etc.
- The objective ''goodness'' of an intervention, as compared with alternatives, for the specific patient, which adds patient-specific features, if known, to the above calculation.
- The actual selection of an intervention for this patient in this case.
- The doctor's past history of prescribing — or not prescribing — this intervention in relevant circumstances. For example, a doctor might: (1) be continuing to prescribe an old medication instead of a better new one due to engrained past experience; (2) insist on a name-brand if a generic has been made available; or (3) prefer one company's offering over a similar offering from another company despite high additional costs to the patient; and so on.

At present, OntoAgent, supports decision functions either directly acquired as if-then rules or using influence diagrams. To illustrate the former, the rule presented below in pseudocode determines whether a virtual patient will go to see the doctor at a given time (for more such functions and discussion see Nirenburg et al., 2008):

```
If FOLLOW-UP-DATE is not set
   And SYMPTOM-SEVERITY > ABILITY-TO-TOLERATE-SYMPTOMS
   Then SEE-MD      ; this triggers the first visit to the MD
Else if FOLLOW-UP-DATE is not set
   And SYMPTOM-SEVERITY < ABILITY-TO-TOLERATE-SYMPTOMS
   And SYMPTOM-DURATION > 6 months
   Then SEE-MD      ; a tolerable symptom has been going
                    ; on for too long
Else if there was a previous visit
   And at the time of that visit SYMPTOM-SEVERITY ⩽ .3
   And currently SYMPTOM-SEVERITY > .7
   And SYMPTOM-SEVERITY-ABILITY-TO-TOLERATE-SYMPTOMS > 0
   Then SEE-MD      ; there was a big increase in symptom
                    severity from low to
                    ; high, triggering an unplanned visit to MD
Else DO-NOTHING
```

Fig. 2 CLAD's relative preferences for each of three available procedures for a given patient.

The second type of decision function employs influence diagrams (Howard & Matheson, 2005), which are a convenient graphical representation facilitating the creation of Bayesian networks and/or decision trees that support reasoning under uncertainty. Such reasoning is particularly useful for clinical decisions, such as choosing the best treatment for a given patient. The current implementation of CLAD represents the results of such reasoning as well as rationale behind the decision. A sample message automatically generated when the clinician seeks advice about which of three available procedures to select for a patient with the disease achalasia is: *Given the current data, we strongly recommend Heller Myotomy. The fact that the patient's overall health is good and that the patient has health insurance policy A were especially relevant to this decision.* The relative preference for each available procedure is shown using a bar chart like the one shown in Fig. 2.

Both types of decision functions will be used, as applicable, to support bias detection and other enhancements to OntoAgent agents.

Detecting patient biases

The gold standard of modern medical care is patient-centered medicine, defined as a collaboration in which the physician helps the patient to meet the latter's goals in managing his or her states of health and disease. In the patient-centered paradigm, the physician does not impose a single solution upon the patient but, rather, instructs, advises and listens to the patient with the purpose of jointly arriving at an optimum solution. The patient's goals might be summarized as, ''Talk to me, answer my questions, and solve my problem in a way that suits my body, my personal situation and my preferences''. The doctor's goals might be summarized as, ''Make an accurate diagnosis; have a compliant patient who is informed about the problem and his or her options, and makes responsible decisions; and launch an effective treatment''.

To best serve the patient, the clinician should be aware of psychological effects on decision-making that might negatively impact the patient's decisions. If a patient makes a decision that the doctor considers suboptimal, the doctor can attempt to understand why by modeling what he believes the patient knows, believes, fears, prioritizes, etc., and hypothesizing a decision function that might have led to the given decision. For example, say the doctor suggests that the patient take a medication that the doctor knows to be highly effective and that has infrequent, mild side effects about which the doctor informs the patient. In response to the doctor's suggestion, the patient refuses, saying he does not want to take that kind of medication. When the doctor asks why, the patient responds in a vague manner, saying that he just has a bad feeling about it. Rather than try to force the patient or badger him for a better explanation, the doctor — in the role of psychologist gumshoe — can break down the decision process into inspectable parts and constructively pursue them in turn. Let us consider the process in more detail.

A person who is considering advice to take a medication will likely consider parameters including the following: the list of potential benefits, risks and side effects and, for each, its intensity (e.g., *how* beneficial is it?), importance, and likelihood; the cost, in terms of money, time, emotional drain, etc.; the patient's trust in the doctor's advice; and the patient's beliefs in a more general sense — about medication use overall, being under a physician's care, etc.

Returning to our example, suppose the drug that the doctor recommended was hypothetical drug X, used for headache relief. Suppose also that the doctor describes the drug to the patient as follows: ''It is very likely that this drug will give you significant relief from your headaches and it might also improve your mood a little. The most common side effect is dry mouth, and there is a small chance of impotence. Unfortunately, the drug has to be injected subcutaneously twice a day''. From this, the patient — or the system modeling the patient's knowledge — will have the following information for its decision function. We include conditional flags (described below) in the structures as italicized comments.

HEADACHE-RELIEF	[intensity: high, likelihood: high]	
MOOD-LIFT	[intensity: low, likelihood: moderate]	
IMPOTENCE	[likelihood: low]	*FLAG for male patients*
DRY-MOUTH	[likelihood: high]	*FLAG for wind instrument players*
COST-EFFORT	[high]	*FLAG because injectable*
COST-EMOTIONAL	[potentially high]	*FLAG if needle-phobic*

In addition, the patient and doctor know that the following can affect the decision, though no values have been explicitly discussed:

COST-FINANCIAL	*FLAG if no health insurance*
TRUST-IN-DOCTOR	*FLAG if the doctor feels the patient doesn't trust him*
MEDICATION—AVERSION	*FLAG if the patient belongs to certain socio-ethnic groups*

Finally, the doctor knows that the patient can be affected by various decision-making biases like the following, each of which can be considered a standing (always available) flag for the physician as he attempts to understand the patient's thought processes:

- **The exposure effect**. Modern-day patients are barraged by drug information on the internet and in TV and radio ads, with the latter rattling off potential side effects at a pace. From this, the patient's impression of the medication might involve a vague but lengthy inventory of side effects *that the doctor did not mention*, and these might serve as misinformation in the inventory of parameters used in the patient's decision function.
- **The effect of small samples**. The patient might know somebody who took this medication and had a bad time with it, and generalized from that that it is a bad drug, despite the doctor's description.
- **The effect of evaluative attitudes**. The patient might not like the idea of taking medication at all; or he might not like the idea of some class of medications due to a perceived stigma (e.g., against antidepressants); or he might be so opposed to a given type of side effect that its potential overshadows any other aspect of the drug.
- **Depletion effects**. The patient might be tired or distracted when making his decision and he might consider saying 'no' to be the least risk option; or his fatigue might have caused lapses in attention so that he misremembered the doctor's description of the medication.

CLAD can assist the physician in understanding the patient's decision-making by making explicit the flags that are relevant in the given situation and possible in any situation. (This will, naturally, require the full NLP capabilities of OntoAgent, supplemented by speech recognition technology that we intend to import wholesale.) For example, if our patient is a male with good health insurance and a medical history of having given himself allergy injections for years, it is possible that the impotence side effect is an issue, but unlikely that the financial cost or fear of injections is a detractor. Since CLAD will have access to a patient's online medical records, it can make such contextual judgment calls and give the clinician advice about what features might be best to pursue first. Even things like the patient's trust in the doctor can, we believe, be detected to some degree by the doctor—patient dialog: e.g., if the patient argues with the doctor, or asks a lot of questions, or frequently voices disagreement, it is possible that low trust is affecting his decision. To summarize, we are working on understanding a patient's decision-making using underspecified decision functions, defined as functions comprised of some features whose values are known and some features whose values can only be constrained to a population-informed range. (Compare this use of underspecification with the underspecified virtual patient models that CLAD uses for prognosis-oriented simulations.)

Another decision-making effect that might affect a patient is the **halo effect**, which is the tendency to make an overall positive or negative assessment of a person on the basis of a small sample of known positive or negative features. For example, if you know a person is kind and successful you might also assume that he or she is generous, even though you know nothing about this aspect of his or her character. As Kahneman (2011, p. 83) says, ''...The halo effect increases the weight of first impressions, sometimes to the point that subsequent information is mostly wasted''. We will suggest that an extended notion of the halo effect — in which it can extend to objects and events as well — can undermine good decision-making by patients. On the one hand, the patient might like the doctor so much as to agree to the latter's advice before learning a sufficient amount about it to make a responsible, informed decision; on the other hand, the patient might dislike the doctor so much as to refuse advice that would actually be beneficial because of a halo effect-induced generalization that a ''bad'' doctor must be giving bad advice. Extending the halo effect to events, a patient might be so happy that a procedure has few risks that he or she assumes that it will not involve any pain and will have no side effects — both of which might not be true. By contrast, a patient might be so influenced by the knowledge that the procedure will hurt that he or she loses sight of its potential benefits. Physicians should detect halo effects in order to ensure that patients are making the best, most responsible decisions for themselves. It is no better for a patient to blindly undergo surgery because he likes a doctor than it is for a patient to refuse life-saving surgery because he is angry with him.

In order to operationalize the automatic detection of halo effects, we are constructing ''halo property nests'' like the ones shown in Table 5. These are inventories of properties that form a constellation with respect to which a person might evaluate another person, thing or event.

Each value or range of values for a property has a positive-halo, negative-halo or neutral-halo score. If a patient knows about a given property value that has a positive-halo score (e.g., low risk) but does not know about any of the other property values in the nest, it is possible that he or she will assume the values of the other properties to have the same ''polarity of halo'' score (e.g., low pain, low side-effects, high benefits). This can explain why a patient who knows little about a procedure might accept or decline it out of hand. Understanding this potential bias can help a clinician to tactfully continue the knowledge-providing conversation until the patient actually has all the information he or she needs to make a good decision. CLAD's role in the process is to trace the hypothetical decision-making process of the patient, determine whether or not he or she knows enough feature values to make a good decision and, if not, flag the clinician.

The final class of decision-making biases to which a patient might be subject pertains to the nature of the doctor—patient dialog. The way a situation is presented or a question is asked can strongly impact a person's perception of it and subsequently affect related decision making. For example, if a person is asked ''Doesn't something hurt right now?'' he or she will have a tendency to seek corroborating evidence — something that actually hurts (**the confirmation bias**). If a person is asked, ''Your pain is very bad, isn't it?'' he or she is likely to overestimate his pain because he has been primed with a high pain level (**the priming effect**). And if a person is told, ''There is a 20% chance this will fail,'' his perception of it will be more negative than if he had been told, ''There's an 80% chance this will succeed'' (**the framing sway**).

CLAD will help doctors to be aware of, and learn to avoid, negative consequences of such effects by automatically

Table 5 Halo property nests.

Object or event	Nest of properties
MEDICAL-PROCEDURE	RISK, PAIN, SIDE-EFFECTS, BENEFITS
PHYSICIAN	INTELLIGENCE, SKILL-LEVEL, AFFABILITY, KINDNESS, TRUSTWORTHINESS

Table 6 Examples of utterances and their respective bias-oriented features.

Example	UTTERANCE-FEATURE
— You don't smoke, do you? — I assume you don't eat before sleeping.	SEEK-CONFIRMATION
Do you have sharp pain in your lower abdomen?	SUGGESTIVE-YES/NO
Do you drink between 2 and 4 cups of coffee a day?	PRIME-WITH-RANGE
There's a 10% chance the procedure will fail.	NEGATIVE-FRAMING-SWAY
There's a 90% chance the procedure will succeed.	POSITIVE-FRAMING-SWAY

detecting and flagging relevant situations. The detection methods are very similar to those being used in OntoAgent for the language phenomena of indirect speech acts and ellipsis.[1] Table 6 contains examples of utterances and the related bias-oriented features that CLAD will detect.

It is relatively straightforward to detect these kinds of utterance features based on the semantic representations generated by the OntoAgent text understander. The utterance feature values are, in turn, incorporated into rules for good clinical negotiation that look just like rules for the detection of indirect speech acts and ellipsis. For example, if the clinician is attempting to convince the patient to agree to a procedure, he will be more a more effective negotiator if he frames the side effects, risks, etc., using a positive framing sway rather than a negative one. Similarly, if the clinician wants the patient to provide maximally objective ratings of his symptoms, then symptom-related questions should be neutral, ''Do you have any chest pain?'' rather than SUGGESTIVE-YES/NO or PRIME-WITH-RANGE. CLAD can match the most desired utterance types with its assessment of the clinician's goal in the given exchange, using the tracking of hypothesized goals and plans (e.g., ''convince patient to undergo procedure'').

[1] Speech acts correlate with the goal of the person in producing an utterance: e.g., if a person is cold and the window is open, saying *I'm cold* can be a request to the person near the window to close it. We coin the term ''utterance-feature'' in order to avoid the need to formally define speech acts vs. other speech-related features.

Discussion

When considering the utility of CLAD in advising a clinician in patient-communication matters, it is important to remember that the psychological effects we have been discussing are typically not recognized by people during decision making, conversation, etc. As such, it is not that CLAD is expected to *discover* anything the clinician does not already know or could not learn in principle: clearly, a clinician could memorize all of the decision-making biases presented in this paper in a very short time. Instead, CLAD will *point out* to the clinician certain things that the latter might not remember or might fail to notice in the given situation. We expect CLAD's help to be particularly useful for clinicians who have less experience overall, who have little experience with a particular constellation of findings, who are under pressures of time and/or fatigue, or who are dealing with difficult non-medical aspects of a case, such as a non-compliant patient.

The research methodology of OntoAgent is geared toward developing theories, technologies and resources for solving real-world problems at a level of quality approaching that of a human expert. Training clinicians and providing them with real-time decision support are excellent examples of real-world problems for which modeling human cognitive capabilities is the most promising approach. Task orientation promotes the development of application systems based on hybrid approaches, with diverse microtheories contributing to overall decisions.

With respect to evaluation, OntoAgent applications face the same challenges as any system that seeks to model multiple complex aspects of human cognition: how to ensure that the evaluation is fair and representative. No extant evaluation paradigms can capture the scientific contributions of modeling advances of the kind implemented in OntoAgent while simultaneously fulfilling the standard expectations for formal evaluations, such as broad coverage, statistical comparisons with previous work and immediate utility of publicly available application systems. In fact, we believe that these kind of evaluation requirements threaten scientific progress in computational cognitive modeling by imposing unachievable short-term demands on inherently long-term efforts.

A review of the literature shows at least four strategic answers to the problem of evaluating inherently long-term scientific and technological endeavors. First, one can select a small, even toy, domain and carry out targeted evaluations of aspects of it. This has been done, for example, in the work of James Allen and collaborators (e.g., Allen & Chambers, 2007; Allen et al., 2006; Ferguson & Allen, 1998), whose agent-oriented systems show marked similarities to those of OntoAgent. Second, one can wear two hats at the same time: carry out cognitively-inspired, descriptive work in one's role as scientist while in one's role as technologist, using a different, simplified approach to building applications that can be evaluated using the traditional metrics. This appears to be the choice of the dialog specialist David Traum (compare Traum (1994) for scientific work with Nouri, Artstein, Leuski, and Traum (2011) for application-oriented work). Third, one can frame work as computationally-oriented but not engage in system building at all. This approach, which is typical, e.g., of NLP-oriented

formal semanticists, has been criticized on the grounds that NLP must involve actual computation (see, e.g., Wilks, 2011); however, its motivation lies in the promise of contributing to future system building. The fourth approach, which is the one we take, involves aspects of all three of the above. Like the first approach, we constrain the domain but not to the point of avoiding the treatment of realistic complexities: e.g., our language understanding uses a full-scale lexicon with all its inherent ambiguity, not a domain-specific lexicon that erases lexical ambiguity. Like the second approach, our model-building proceeds at a faster pace than our implementations, and our implementations can and do incorporate simplifications of the models; however, we do not attempt to launch applications that completely exclude cognitive modeling. Like the third approach, we do not claim to be building systems for end users, as at this time this requires far-reaching simplification of system functionalities that will result in unacceptably low system quality, to say nothing about reducing the exploratory scientific impact of our approach.

To evaluate the system, we presented the virtual patient simulations to a group of clinicians, who uniformly confirmed the verisimilitude of the simulations.

Building systems to fulfill externally delineated desiderata — such as the need for decision support for overworked clinicians — involves front-loading the research program so that the full problem space, necessary component technologies, and scope of modeling strategies can be understood. It is only through understanding the overall system architecture that we can understand how best to build the often overlapping component modules. Moreover, the insights derived from work on a given angle of cognitive modeling can, we believe, be useful across research paradigms even before every *i* has been dotted and *t* crossed, whether those insights be imported wholesale, in part, or whether they simply serve as food for thought in this vast enterprise of realizing artificial intelligence.

Conclusions

This paper is the initial report of our work on developing intelligent agents that can help people to recognize and avoid cognitively biased decision-making. We hypothesize that since people can often recognize biased decision-making in others, we can and should build human-inspired intelligent agents with the same capabilities. The implementation of our theory of bias recognition within the OntoAgent cognitive architecture serves as proof of concept of the theory's computability.

The literature devoted to intelligent medical advisors includes extensive discussion of what physicians want in advisor systems and their willingness — or unwillingness — to incorporate new technologies into their workflow (see, e.g., Ely, Osheroff, Chambliss, Ebell, & Rosenbaum, 2005). The vision of, and goals for, CLAD are aimed solely at providing non-binding assistance to physicians who choose to use the system. Moreover, the research findings and technologies being developed have direct applications to other areas, such as training, where expert input is unquestionably appropriate.

The modeling strategies illustrated here have been developed for, and tested on, a half dozen diseases of the esophagus. These are not the most complex diseases known to medicine, but we believe that the modeling strategies we developed will be sufficient to cover a wide range of more complex diseases and disease interactions. The core prerequisite for modeling the more complex cases is that expert informants be able to explicitly formulate sufficient knowledge about physiology and clinical medicine to support deterministic modeling. To facilitate this, we have developed a dedicated methodology for acquiring knowledge from experts. As the field of medicine discovers more about disease processes and interactions, and develops better clinical management strategies, all of these can be incorporated into the OntoAgent knowledge bases, thus expanding the scope of cases about which OntoAgent advisors can confidently offer advice.

References

Allen, J. F., Chambers, N. et al. (2007). PLOW: A collaborative task learning agent. *National Conference on Artificial Intelligence (AAAI)*. Vancouver, BC.

Allen, J., Ferguson, G., Blaylock, N., Byron, D., Chambers, N., Dzikovska, M., et al (2006). Chester: Towards a personal medication advisor. *Journal of Biomedical Informatics, 39*, 500—513.

Bratman, M. (1999). *Faces of intention*. Cambridge University Press.

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H., & Rosenbaum, M. E. (2005). Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association, 12*(2), 217—224.

Ferguson, G. & Allen, J. (1998). TRIPS: An integrated intelligent problem-solving assistant. In *Proceedings of the national conference on artificial intelligence*.

Gawande, A. (2009). *The checklist manifesto*. NY: Henry Holt and Company.

Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology, 115*(2), 192—194.

Gunzelmann, G., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep deprivation and sustained attention performance. Integrating mathematical and cognitive modeling. *Cognitive Science, 33*(5), 880—910.

Heuer, R. J. Jr. (1999). *Psychology of intelligence analysis. Center for the study of intelligence*. Central Intelligence Agency <http://www.au.af.mil/au/awc/awcgate/psych-intel/>.

Howard, R. A., & Matheson, J. E. (2005). Influence diagrams. *Decision Analysis, 2*(3), 127—143.

Kahneman, D. (2011). *Thinking: Fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515—526.

Korte, Russell F. (2003). Biases in decision making and implications for human resource development. *Advances in Developing Human Resources, 5*(4), 440—457.

McShane, M., Beale, S., Nirenburg, S., Jarrell, B. & Fantry, & G. (2012). Inconsistency as a diagnostic tool in a society of intelligent agents. *AI in Medicine, 55*(3), 137—148.

McShane, M., Fantry, G., Beale, S., Nirenburg, S. & Jarrell, B. (2007). Disease interaction in cognitive simulations for medical

training. In *Proceedings of MODSIM world conference, medical track, 2007, September 11—13*. Virginia Beach.

McShane, M., & Nirenburg, S. (2012). A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing, 06*(1).

Meehl, P. (1996). *Clinical vs. statistical predictions: A theoretical analysis and a review of the evidence*. Northvale, NJ: Jason Aronson, Originally published 1954.

Nirenburg, S. & McShane, M. (2012). Agents modeling agents: Incorporating ethics-related reasoning. In *Proceedings of the symposium moral cognition and theory of mind at the AISB/ IACAP world congress 2012*. Birmingham, UK.

Nirenburg, S., McShane, M., & Beale, S. (2008). A simulated physiological/cognitive ''double agent''. In *Proceedings of the workshop on naturally inspired cognitive architectures, AAAI 2008 fall symposium, November 7—9*. Washington, DC.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Cambridge, Mass.: The MIT Press.

Nouri, E., Artstein, R., Leuski, A. & Traum, D. (2011). Augmenting conversational characters with generated question-answer pairs. In *Proceedings of the AAAI symposium on question generation*.

Simon, H. (1957). *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York: Wiley.

Traum, D. (1994). *A computational theory of grounding in natural language conversation*, TR 545 and Ph.D. Thesis, Computer Science Dept., U. Rochester, December.

Wilks, Y. (2011). Computational semantics requires computation. In: C. Boonthum-Denecke, Philip M. McCarthy & T. Lamkin (Eds.), *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*. (pp. 1—8). IGI Global.