Parameterizing mental model ascription across intelligent agents

Marjorie McShane
Rensselaer Polytechnic Institute, USA

Mental model ascription – also called mindreading – is the process of inferring the mental states of others, which happens as a matter of course in social interactions. But although ubiquitous, mindreading is presumably a highly variable process: people mindread to *different extents* and with *different results*. We hypothesize that human mindreading ability relies on a large number of personal and contextual features: the inherent abilities of specific individuals, their current physical and mental states, their knowledge of the domain of discourse, their familiarity with the interlocutor, the risks associated with an incorrect assessment of intent, and so on. This paper presents a theory of mindreading that models diverse artificial intelligent agents using an inventory of parameters and value sets that represent traits of humans and features of discourse contexts. Examples are drawn from Maryland Virtual Patient, a prototype system that will permit medical trainees to diagnose and treat cognitively modeled virtual patients with the optional assistance of a virtual tutor. Since real patients vary greatly with respect to physiological and cognitive features, so must a society of virtual patients. Modeling such variation is one of the goals of the overall OntoAgent program of research and development.

Keywords: mindreading; mental model ascription; agent modeling; cognitive modeling

1. Introduction

Mental model ascription, otherwise known as mindreading, can be defined as inferring features of another person or artificial agent that cannot be directly observed, such as that individual's beliefs, plans, goals, intentions, personality traits, mental and emotional states, and knowledge about the world. Thanks to mindreading, human interactions can be snappy, not requiring endless explanations and clarifications of things that people can – in most cases, effortlessly and correctly – infer.

Interaction Studies 15:3 (2014), **404–425**. DOI 10.1075/is.15.3.03mcs
ISSN 1572–0373 / E-ISSN 1572–0381

But although ubiquitous, mindreading is variable across individuals and situations. For example, a person who is typically mentally acute and adept at mindreading might fail at it, or not even attempt it, if placed in a physically or emotionally taxing situation. By contrast, a person who is less insightful overall might be able to successfully mindread given a highly familiar situation in which he experiences complete physical and emotional comfort. Since a high degree of variability is expected both within and across human individuals, a corresponding degree of variability should be incorporated into a society of intelligent agents modeled to emulate humans.

This paper discusses the individualization of mindreading potential and outcomes across intelligent agents within the OntoAgent environment. Examples are drawn from the Maryland Virtual Patient (MVP) prototype application, which seeks to provide medical trainees with the opportunity to diagnose and treat a society of virtual patients in open-ended cognitive simulations. This new work contributes to the body of published research carried out within the OntoAgent paradigm of R&D. Before proceeding to the new contributions of this paper, let us briefly describe key aspects of the OntoAgent environment, with pointers to papers that provide more comprehensive discussions of each topic.

1.1 OntoAgent

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 capable of perception, reasoning and action. The OntoAgent approach is different from many cognitive architectures (Langley, Laird & Rogers 2009) in that it centrally involves the integration of physiological and cognitive models, and emphasizes deep, semantically-oriented natural language processing. A top-level view of the architecture is shown in Figure 1.

As the figure suggests, The OntoAgent approach to configuring intelligent agents necessitates a comprehensive program of R&D that actively actively covers many areas. The main threads are briefly summarized below.

Ontologically-grounded modeling. All physiological, general cognitive and language processing capabilities of all intelligent agents rely on the same knowledge representation language and ontological knowledge substrate (McShane & Nirenburg 2012). The ontology is a formal model of the world that is organized as a multiple-inheritance hierarchical collection of frames headed by concepts (objects and events) that are named using language-independent labels (Nirenburg & Raskin 2004). It currently contains approximately 9,000 concepts. The objects and events are described using properties – attributes and relations. The properties themselves are primitives, i.e. their meaning is understood to be

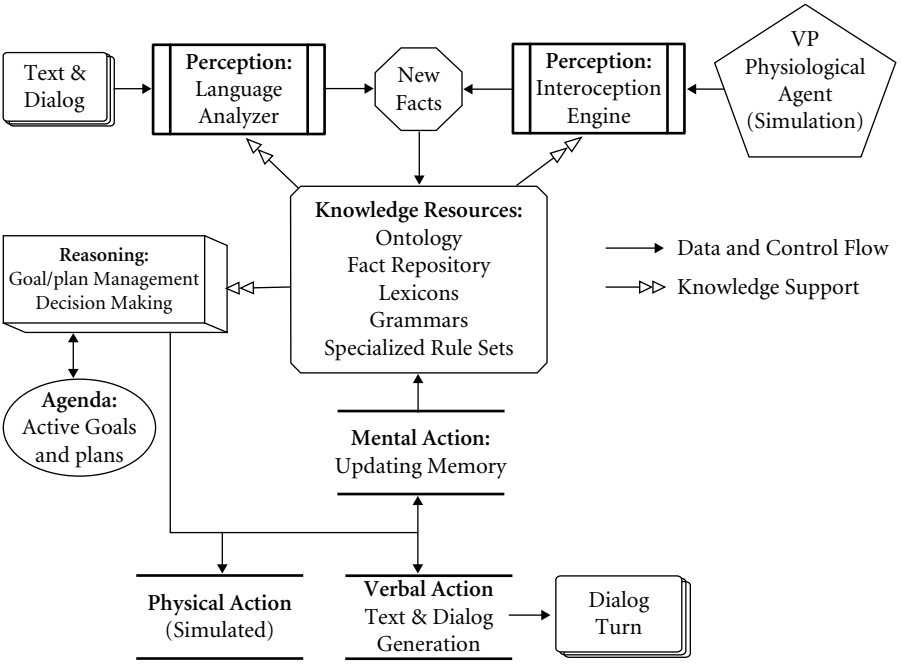

Figure 1. The architecture of OntoAgent agents. The physiological agent is represented by the upper right pentagon. The rest of the figure describes aspects of the cognitive agent.

grounded in the real world without the need for further ontological decomposition. A short excerpt from the frame for SURGERY is as follows:

SURGERY		
AGENT	default	SURGEON
	sem	PHYSICIAN
	relaxable-to	HUMAN
THEME	default	MEDICAL-PATIENT
	sem	ANIMAL
LOCATION	default	OPERATING-ROOM
	sem	MEDICAL-BUILDING
	relaxable-to	PLACE
INSTRUMENT	default	SURGICAL-INSTRUMENT
DURATION	sem	.5–8 (MEASURED-IN HOUR)
…		

Complex events are recorded using scripts of the type first introduced by Schank & Abelson 1977. Scripts represent typical sequences of events and their causal and

temporal relationships. In other words, they encode how individual events hold well-defined places in routine, typical sequences of events that happen in the world, with a well-specified set of objects filling different roles throughout that sequence. In OntoAgent, scripts both drive agent simulation and support agent reasoning. For example, the script that describes a disease (its causes, paths of progression, potential responses to interventions, etc.) permits (a) simulation of the disease in virtual patients, (b) reasoning about disease processes by the virtual medical tutor and (c) natural language dialog about the disease, since semantically-oriented natural language processing requires real-world knowledge support. In short, a theoretically and practically motivated aspect of knowledge acquisition in OntoAgent is that knowledge, once recorded, should enable the maximum number of functionalities in the maximum number of agents collaborating in the environment.

Physiological simulation (upper right pentagon in Figure 1). We model normal and pathological physiology as changes, over time, in the property values of the ontological objects and events that are used to model human beings. The depth and granularity of the models are determined by the goal of achieving of realism in our current and anticipated applications. As a rule of thumb, a feature value or process is included in the model if it can either be measured through tests, be affected by medication/interventions, or cause a change in some other clinically relevant feature value or process. Of central importance is the fact that our models can be easily modified, reflecting new medical findings.

Table 1 presents an abstract example of the kind of knowledge that contributes to disease modeling. We present an abstract, rather than an actual, example because the medical details of real examples cannot be sufficiently described in this space. For medical examples and full descriptions of disease models, see McShane, Fantry et al. 2007; McShane, Nirenburg, et al. 2007; McShane, Jarrell et al. 2008).

Table 1. Abstract example of knowledge supporting the physiological simulation of a disease

	Initial (non-disease) value	Stage 1 end value	Stage 2 end value	Stage 3 end value
Property-1	10–13	14–20	20–30	40–70
Property-2	0	.1–.3	.1–.5	.3–.9
Property-3	negative	negative	positive	positive

Diseases modeling orients around conceptual stages – i.e. stages at which clinically important events happen. Typical ranges of values for each relevant property are associated with each stage. Patient instances show particular progressions of property values over time, which are calculated using interpolation functions in the simulation engine.

The hypothetical disease shown in Table 1 is described using three properties. Property-1 is a scalar attribute with values that range from 10 to 70. The value for healthy individuals can range from 10 to 13. Property-2 is also a scalar attribute but on the abstract scale of {0,1}. Property-3 is a literal attribute, whose values are negative or positive.

When an individual patient instance is suffering from this disease, the value of Property-1 at the end of stage 1 can be anywhere from 14 to 20, and so on for the other stages. When one considers that the length of each stage of the disease varies across patients, it becomes clear that the actual graph for this property value over time across patients can be very different: e.g. whereas a patient with a rapidly progressing disease might reach the end of Stage 2, with a value of 30 for Property-1, after only 3 months, another patient might remain in the middle of Stage 1, with a value of 16 for Property-1, after a year.

The fact that properties values are variable, but only within a certain tolerance, illustrates that our physiological models are constrained enough to represent the clinically observed shape of the disease but variable enough to show significant differences across the population of patients. Diseases progress according to models like these unless medical therapies are applied, the patient's other medical conditions change, or the patient changes his/her lifestyle – processes that are described in the abovementioned references.

Interoception. Interoception is the agent's perception of its bodily signals, which are generated by the physiological agent (Nirenburg et al. 2008). The interoception submodule operates a set of demons that are programmed to notice the changes in values of specific physiological parameters; if these values move outside a certain range, they instantiate corresponding symptoms in the agent's memory.

Language understanding. Language understanding in OntoAgent involves preprocessing followed by morphological, syntactic, semantic and pragmatic analysis (see Nirenburg & Raskin 2004; McShane, Nirenburg & Beale 2005; McShane 2009; McShane & Nirenburg 2009; and references therein). The result of language analysis, when all goes well (semantically-oriented natural language understanding being, after all, a long-term challenge of artificial intelligence), is an unambiguous text meaning representation (TMR) written in the metalanguage of the OntoAgent ontology. The text meaning representation is organized as a group of frames headed by uniquely indexed instances of ontological concepts. Concepts, which are written in small caps to distinguish them from English words, are

unambiguous: e.g. DOG refers only to the domesticated canine, not to a despised person or the act of following someone. Although in the *ontology* each concept is described by a large inventory of relations and attributes, in a *text meaning representation*, each concept instance is described using only the subset of properties that is overtly mentioned in the input text. For example, all of the sentences in (1) will result in the same meaning representation.[1]

(1) a. Sallie saw ⟨caught sight of, noticed⟩ the brown puppy.
b. The brown puppy ⟨It was the brown puppy that⟩ caught Sallie's eye.
c. It was Sallie who noticed ⟨saw, caught sight of⟩ the brown puppy.

```
INVOLUNTARY-VISUAL-EVENT-1
    EXPERIENCER          HUMAN-1
    THEME                DOG-1
    TIME                 < find-anchor-time   ; before the time the text was written
HUMAN-1
    GENDER               female
    HAS-PERSONAL-NAME    Sallie
DOG-1
    COLOR                brown
    AGE                  < 1 (MEASURED-IN YEAR)
```

The text analysis system looks up each word used in the sentence in the OntoAgent lexicon, which currently contains over 30,000 word senses, and selects the senses and semantic dependencies that are most suitable for the context. For example, the words *see, catch sight of,* and *notice* all have a sense that maps to the concept INVOLUNTARY-VISUAL-EVENT; the meaning of *puppy* is recorded as DOG (AGE < 1 (MEASURED-IN YEAR)); the EXPERIENCER and THEME case roles are selected, and filled respectively, based on syntactic and semantic expectations recorded in the lexical senses of the selecting verbs. Differences in syntactic structure are handled as a matter of course by the text analyzer.

Meaning representations like this one are optimized for intelligent agent reasoning because (a) they refer to ontological concepts that provide additional information about each object and event, should that be needed for reasoning, and (b) they are unambiguous and reflect the results of extensive reasoning about language and the speech context.

Decision-making and learning. Apart from the reasoning and decision-making involved in language understanding, OntoAgents make decisions in many

1. Although these paraphrases convey different semantic nuances to human readers, this level of distinction is not relevant to the functioning of our intelligent agents at this time and, therefore, is not reflected in TMRs.

other realms as well. Consider just a couple of examples. (1) Memory management involves deciding whether an encountered object or event is already known, in which case new information should link to the existing anchor in memory, or whether it represents a new instance that should give rise to a new anchor (McShane 2009; McShane, Nirenburg & Beale 2011; McShane & Nirenburg 2013). (2) Collaborative dialog involves determining how to phrase, and paraphrase, information so that it will be understandable to the interlocutor – a process that involves a different aspect of mindreading than the one to be discussed here (McShane, Nirenburg & Beale 2008). (3) Planning involves tracking the goals of oneself and others: e.g. when a virtual patient decides whether or not to agree to a procedure suggested by a user, it incorporates many features of its personality, state of health, beliefs about the attending physician, etc., into that decision function (Nirenburg et al. 2008; Nirenburg and McShane 2012). (4) Providing professional advice can involve detecting cognitive decision-making biases in others (McShane et al. 2013). (5) Effectively collaborating can involve detecting when one's collaborator might not be telling the truth and deciding how to proceed in those circumstances (McShane et al. 2012). (6) Learning involves applying the agent's current state of knowledge to newly encountered experiences and information: e.g. an agent can learn about the manifestations of a disease from its own simulated experiences; it can learn about the properties of a disease, or available treatments, by being told by its physician; and it can learn about its collaborators by reasoning about them based on their actions (Nirenburg et al. 2010).

All of these aspects of agent functionality, and more, must converge to permit intelligent agents to perform with verisimilitude in interactive applications with human users.

1.2 Maryland Virtual Patient (MVP)

MVP is one prototype application being developed within the OntoAgent environment. The core agent is a virtual patient, which is a knowledge-based model and simulation of a person suffering from one or more diseases (see Figure 2).

To reiterate, the virtual patient is a "double agent" in that it models and simulates both the physiological and the cognitive functionality of a human. Physiologically, it undergoes both normal and pathological processes and responds realistically both to expected and to unexpected stimuli: e.g. a human user can launch a correct procedure to treat a disease (an expected stimulus) or he/she can launch a procedure that has nothing to do with the given disease but whose effects will nevertheless alter the subsequent simulated life of the patient. Cognitively, the virtual patient experiences symptoms, has lifestyle preferences (a model of character traits), has dynamic memory and learning capabilities, has the ability to reason

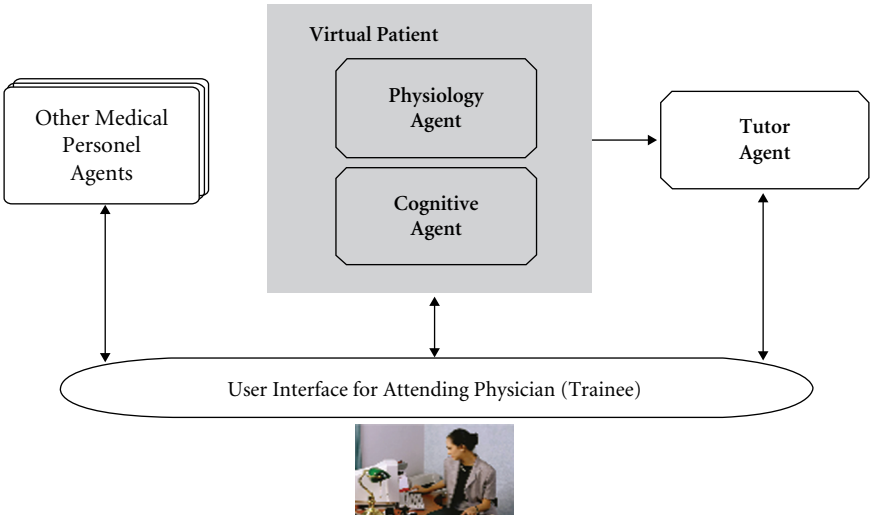

Figure 2. Architecture of the Maryland Virtual Patient prototype application

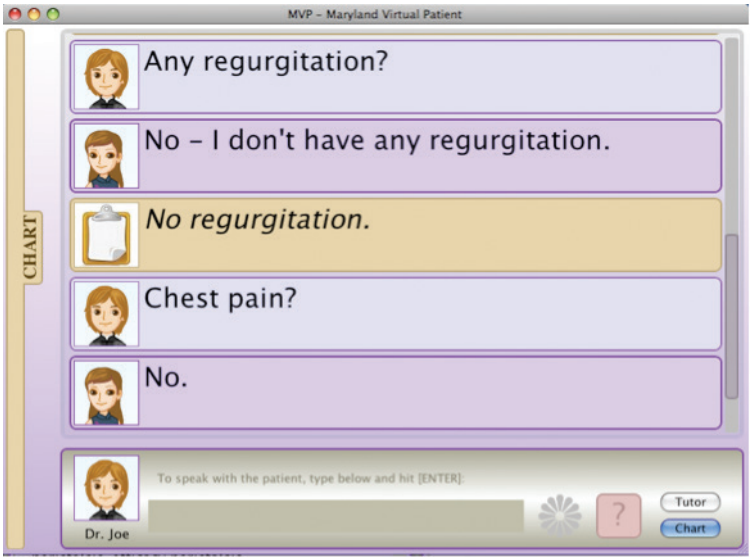

Figure 3. Sample dialog excerpt from Maryland Virtual Patient

in a context-sensitive way, and can communicate with the human user about its personal history, symptoms and preferences for treatment. Figure 3 shows a screen shot from the user interface in the middle of an interactive simulation. The first character (light background) is the doctor and the second one (darker background) is the patient. The "turn" marked by the clipboard shows what is

automatically saved to the patient chart, which can be accessed by the user using the "chart" tab to the left.

Users of MVP can interview a virtual patient; order lab tests; receive the results of lab tests from technician agents; receive interpretations of lab tests from consulting physician agents; posit diagnostic hypotheses, clinical diagnoses and definitive diagnoses; prescribe treatments, like medication and surgery; follow-up on those treatments to judge their efficacy; follow a patient's condition over an extended period of time; receive mentoring from the automatic tutor, if desired; and repeat the management of a given patient using different management strategies to compare their outcomes. The user can launch any intervention available in the system at any time during the simulation, be it clinically justified or not. In the latter case, if the user inadvertently worsens the patient's condition or initiates a new disease process, he must recover from the error in the continuing simulation by treating the new condition he has inadvertently caused. Users can also, under certain system configurations, peer "under the hood" of the simulation, directly observing the physiological models, which can serve as supplementary training tools (McShane, Jarrell et al. 2008). To date, several diseases of the esophagus have been modeled, and sufficient physiological and cognitive modeling has been carried out to support demonstrations of this proof of concept system.

2. Paramaterizing mindreading

From this broad view of the OntoAgent and MVP programs of R&D, let us now zoom in to the new contribution of this article: describing a cognitive model that will permit different instances of virtual patients to display different mindreading behaviors in different contexts. We will focus on two aspects of mindreading that are particularly central to the full interpretation of language input: the interpretation of *indirect speech acts* (e.g. *I'm cold* can be said if the speaker wants the interlocutor to close the window) and attending to the *goal* of an utterance over its content (e.g. if your brother asks, *Do you have gas*? you can respond, *Yeah, but you can't borrow my car*, based on your interpretation that borrowing the car was his ultimate goal).

As an organizing principle, the discussion will trace the algorithm presented in Figure 4.

It must be emphasized from the outset that this model, like all OntoAgent models, attempts to achieve functional verisimilitude within applications. There is no direct scientific evidence for this complex and highly interconnected aspect of human functioning, nor is there reason to believe that there ever will be, despite glimpses into measurable aspects of human cognition that are being pursued through psychological experimentation. We intentionally reduce complexity

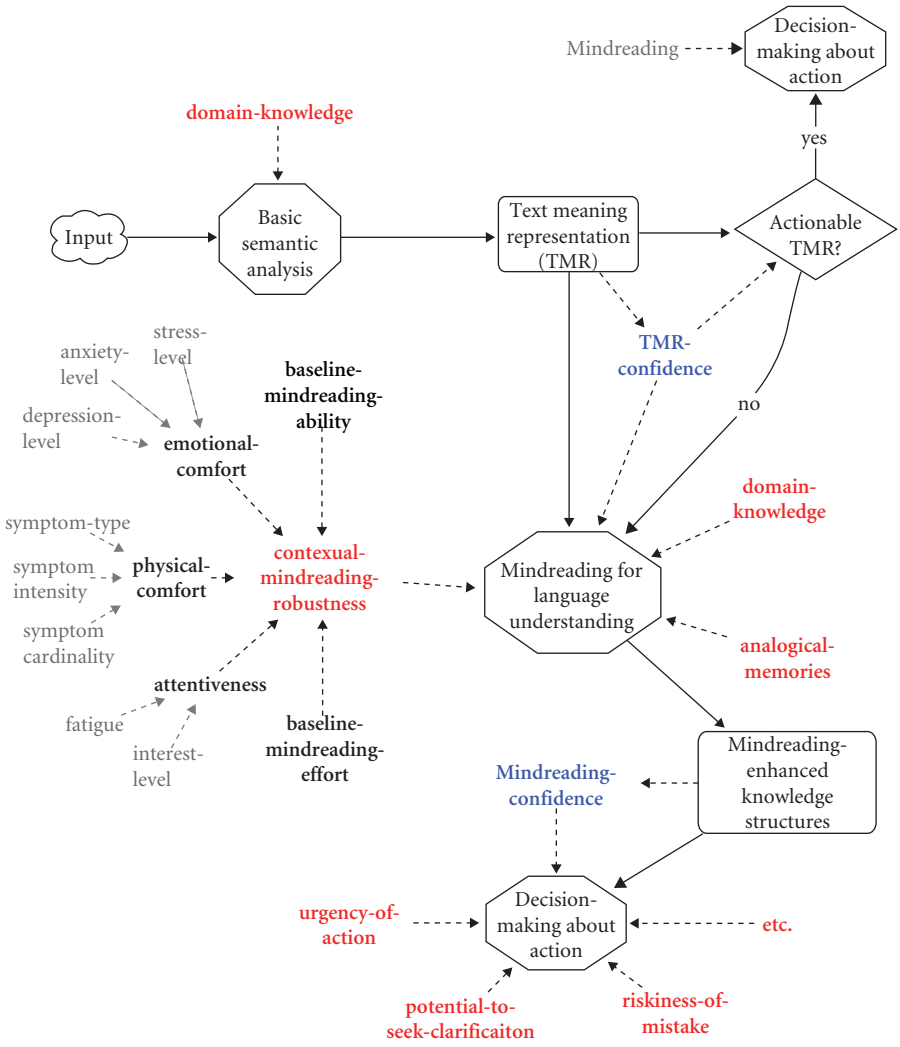

Figure 4. The model of mindreading in support of language understanding for OntoAgents

where possible, following the observation of Kahneman (2011) that simpler models tend to work better than complex, excessively feature-rich models.

Let us trace through the algorithm in Figure 4 starting at the top left, where the agent receives a language input. For purposes of illustration, we will orient around Example (3), which is an input from a human user of the MVP system to the virtual patient.

(3) I need to know your weight.

The virtual patient carries out basic semantic analysis, including reference resolution (i.e. anchoring entities and events to memory), to arrive at the text meaning representation (TMR) below. Comments after semi-colons indicate the lexical senses selected to generate the meaning representation: e.g. *need-v2* is the 2nd verbal sense of the word *need*, whose meaning is realized as obligative modality scoping over the main proposition, headed by KNOW.

MODALITY-1		*; from need-v2*
TYPE	OBLIGATIVE	
VALUE	1	*; the highest value on the abstract scale {0,1}*
ATTRIBUTED-TO	PHYSICIAN-1	*; from reference procedures recorded in I-n1*
SCOPE	KNOW-1	
KNOW-1		*; from know-v1*
AGENT	PHYSICIAN-1	
THEME	WEIGHT-1	
WEIGHT-1		*; from weight-n2*
DOMAIN	PATIENT-1	*; from the compositional analysis of 'your'*

Once the agent generates a text meaning representation, it must evaluate its confidence in its accuracy (cf. "TMR confidence" in Figure 4). This judgment takes into account such things as the number of senses recorded for each word of input in the lexicon (the more senses, the greater the chance for error), whether or not the word is in the lexicon to begin with (unknown words greatly reduce overall confidence), and how tight the ontological constraints in the dependency structure are (e.g. the concept INGEST expects its THEME to be some food or drink, which is a tight constraint; by contrast, the concept VOLUNTARY-VISUAL-EVENT permits its THEME to be any physical object, which is a loose constraint).

Although this TMR includes the results of disambiguating all the words of input and carrying out reference resolution for *I* and *you*, it does not *directly* indicate what is expected of the agent as a response. To understand why, consider what the agent that is playing the role of the virtual patient knows how to do: respond to questions, respond to requests for action (including accepting/rejecting advice), and remember/learn new things. In the simplest case, questions are formulated as questions (*What do you weigh?*), requests are formulated as imperatives (*Get on the scale, please.*), and information-providing utterances are formulated as declarative sentences (*I know that you weigh 160 pounds.*) However, in natural language there is no strict mapping between the form of utterances and they speech acts they can convey. For example, the statement *I need to know your weight* can function as a request for information (*What do you weigh?*), as a request for action (*Please get on the scale I'm pointing to*), or simply as information, as when a doctor prepares a patient for an upcoming request (*I need to know your weight. When we finish talking, we'll step out and get you on the scale.*). In terms of preparing agents

to function in a dialog application, arguably the most difficult types of inputs are statements since they commonly cloak indirect speech acts and, therefore, require mindreading for full interpretation.

To summarize, because language *form* does not always indicate discourse *function,* the agent has to decide whether or not any given text meaning representation is immediately "actionable", or whether it requires mindreading first.We define "actionable" as either of the following:

1. The TMR includes an instance of the concept REQUEST-INFO because a question was phrased as a question, *and* the agent's confidence in the language analysis is high, *and* the agent is prepared to field this kind of question. For example, our virtual patients can directly respond to *Do you have heartburn?* because they can generate a confident analysis of this question and know how to respond to questions about symptoms. By contrast, even though our virtual patients could generate a confident TMR for *Why is the sky blue*?, they don't know yet know how to answer 'why' questions about entities outside of the application domain, making the corresponding TMR not actionable and requiring the agent to pursue mindreading.
2. The TMR includes an instance of REQUEST-ACTION because a command was phrased as a command, *and* the agent's confidence in the language analysis is high, *and* the agent is prepared to respond to this kind of request/advice.

If the TMR is directly actionable, then the agent can proceed to deciding what to do, with no need for language-processing-oriented mindreading. Of course, the decision-making process itself *can* involve mindreading, but it is not of the language-oriented type we focus on here (it is of the broader type mentioned in Section 1).

If, by contrast, the TMR is *not* directly actionable, then the agent can optionally engage in mindreading. The decision about how much effort to expend on mindreading, if any at all, and which evidence to bring to bear depends upon features of the agent, the speech context and the extralinguistic context. The current model employs the following four parameters, some of which are considered basic and others of which are derived from other parameters. Figure 4 illustrates how parameters contribute to metaparameters.

1. **TMR confidence:** If the automatically generated confidence score of the TMR is high, then the agent can confidently proceed to trying to understand what is being asked of it. If the score is low – e.g. because the agent didn't know some words of the input – then the agent must judge whether or not it understood any aspect of the TMR sufficiently to act upon it. For example, some physicians provide long and complicated descriptions of a disease, followed by a

simple question to the patient. If the virtual patient incompletely understands the long description but understands the question, it can choose to answer the question without further clarification of the other information.

2. **Domain knowledge:** Domain knowledge reflects the ontologically recorded knowledge structures the agent can rely on for reasoning. For example, say the agent doesn't understand why, during history-taking, the doctor asks the elliptical "And your appendix?" because it does not have any memories about its APPENDIX at all – no related symptoms, no past tests, etc. All the agent understands from the surface form of the question is that something is being asked about its APPENDIX. If the only thing the agent knows about its APPENDIX (thanks to its ontology) is that, being a descendant of ANIMAL-ORGAN, it can be the location of PAIN, then the agent can assume – via mindreading – that the doctor is asking about pain and generate a rejoinder like, *It doesn't hurt* or *Are you asking if it hurts?* If, by contrast, the agent has ontological knowledge that one's APPENDIX can be removed (APPENDIX THEME-OF REMOVE-ORGAN), then its option space for reasonable rejoinders is expanded to include things like, *It was not removed* or *Are you asking if it was removed?* We are not suggesting that reasoning like this occurs for free thanks to having an ontology – of course not. Agents have to be configured to search their knowledge bases for relevant features of known entities in the language input, then formulate responses based on those features. However, although not free, such reasoning is also not prohibitively expensive to implement, particularly in narrow-domain applications where the cost of an error is relatively low.
3. **Analogical reasoning:** When an agent is trying to understand what is expected of it in response to an utterance, past experience with similar utterances can guide the reasoning. Let's consider the *And your appendix?* example again from this perspective. Say the agent receives this input after previously (recently or not recently) having been asked, *And your tonsils?* The agent would "remember" this by searching its past logs of dialog interactions for utterances with similar features – e.g. same speaker, same sentence structure, close ontological distance between key concepts. If a close match is found, the agent can trace how that interaction concluded, even if it involved clarification questions or an initial misinterpretation by the agent followed by being corrected by the interlocutor. In short, what worked before can suggest, by analogy, what might work now, both in support of processing unknown inputs and in support of mindreading in response to any input.
4. **Contextual mindreading robustness** is a metafeature that we introduce primarily to simplify the process of modeling: it is easier to manipulate one feature value in a function than to manipulate a dozen that contribute to it. However, the constellation of combined features is not random, it reflects a

cognitively-based rationale: agents can show the same mindreading-oriented behavior for different reasons. The behavior itself is reflected by the value of contextual-mindreading-robustness; the reasons behind it can be traced back to the contributing parameter values.

Five features directly contribute to contextual mindreading robustness, three of which are, themselves, metafeatures:

- the agent's **baseline mindreading ability**, which is a fixed character trait whose values range from 0 (an uninsightful individual) to 1 (an incredibly insightful individual)
- the agent's **baseline mindreading effort**, which is a fixed character trait whose values range from 0 (the agent never makes any effort to mindread) to 1 (the agent attempts to understand the motivation behind every utterance)
- the agent's **emotional comfort level** which is a contextually-calculated function of stress level, anxiety level and depression level; 0 indicates emotional breakdown whereas 1 indicates perfect emotional comfort
- the agent's **physical comfort level**, which is a contextually-calculated function that incorporates the number, type and intensity of symptoms it is experiencing; 0 indicates agony whereas 1 indicates a lack of symptoms
- the agent's **attentiveness**, which is a function of tiredness and interest level (e.g. an agent can care or not care about the details of its medical condition); 0 indicates a complete lack of attentiveness whereas 1 indicates rapt attention

The metaparameter **contextual-mindreading-robustness** is measured on the scale {0,1}. If the value is ≤.25, then the agent will not even attempt mindreading; if the value is >.25 and ≤.7, then the agent will attempt mindreading but constrain it to simpler, higher-confidence cases; and if the value is >.7, then the agent will exploit all of its resources to optimize and maximize mindreading. The point is that *an agent can arrive at any of these values of contextual-mindreading-robustness for many different reasons.* Table 2 shows some comparative constellations of feature values. This particular mathematical calculation is one of many possibilities; we will hone the numbers as development and testing proceed.

Artie has the baseline personality traits of being a good mindreader (.8) and being motivated to carry out mindreading (.8), but his very low values for emotional comfort, physical comfort and attentiveness (.2) mean that he is currently in very bad shape (perhaps bleeding profusely in an emergency room) and is not up to expending the effort to mindread. Similarly, mindreading will not be attempted by Bertha but for different reasons: although she is in better shape than Artie, she is still suffering physically and emotionally, and has baseline character traits that are not very supportive of insightful mindreading. Chuck and Dick, by contrast,

Table 2. Constellations of feature values that determine how deeply agents pursue mindreading

	Artie	Bertha	Chuck	Dick	Ellie
baseline mindreading ability	0.8	0.4	0.4	0.3	1
baseline mindreading effort	0.8	0.4	0.4	1	1
emotional comfort	0.2	0.6	1	0.8	0.7
physical comfort	0.2	0.6	0.8	0.7	1
attentiveness	0.2	0.5	1	1	0.8
TOTAL	2.2	2.5	3.6	3.8	4.5
Mindreading?	Not attempted		Attempted with moderate effort		Pursued full-on

have a more positive constellation of feature values that support a moderate degree of mindreading. Chuck, like Bertha, is not stellar in terms of mindreading-oriented personality traits, but he is in very good shape (experiencing only minor physical discomfort), so all of his resources are available to devote to mindreading. Dick, like Chuck, is in acceptable physical and emotional shape, and although he is not a good mindreader by nature (.3), he tries awfully hard (1). Ellie stands alone with exceptionally robust contextual mindreading potential: she has inherently excellent mindreading ability, puts out great effort, and is experiencing only slight discomfort that does not impede her efforts.

To summarize, given a language input that requires interpretation beyond generation of a basic text meaning representation, an agent can – for various reasons – not attempt to mindread, attempt to mindread but constrain its efforts to the simpler cases, or go for all-out, super-insightful mindreading. But even if an agent attempts to mindread, it is not guaranteed success since the three other factors discussed earlier all play a role. So, if our crack mindreader Ellie poorly understands the input (TMR confidence), knows little about the domain (domain knowledge) and can't use any comparable memories for guidance (analogical memories), then she will still fail at mindreading.

Let us concretize the discussion by returning to our example of processing the indirect speech act, *I need to know your weight*, and consider how our agents would treat it. Artie and Bertha will not engage in mindreading. The content of the TMR will be the sole evidence used as input to their decision about how to respond. Chuck and Dick will carry out basic procedures for the detection of indirect speech acts, which involve testing the input against an inventory of recorded indirect-speech-act-detection rules. One such rule is the following, in pseudocode:

If all of the following are true of the TMR:
- KNOW is scoped over by MODALITY of type OBLIGATIVE or VOLITIVE
- that MODALITY is ATTRIBUTED-TO the speaker
- the VALUE of that MODALITY is >.5 *; i.e. the person does have to or want to do it*
- the SCOPE of that MODALITY is an ATTRIBUTE whose DOMAIN is the agent
- the agent has knowledge about that ATTRIBUTE in its memory

then reinterpret the TMR as REQUEST-INFO (AGENT speaker) (THEME *THAT-ATTRIBUTE*.RANGE) (BENEFICIARY our-agent).

Although this might seem like an overly specific rule for interpreting indirect speech acts, it actually covers dozens of paraphrases of the non-attribute part (e.g. *I would like to know, I need to know, I need to find out, I'll need to find out, it is necessary to know, I would like to understand, I'm afraid I'll need to find out…*) combined with any of the hundreds of attributes available in the lexicon and ontology. This breadth of coverage surely justifies spending a couple of minutes writing this rule by hand. The agents who expend a moderate amount of effort on mindreading, like Chuck and Dick, rely exclusively on the inventory of recorded rules, so they will readily and with high confidence be able to generate the same meaning representation for *I need to know your weight* as they would for *How much do you weigh?* Ellie, too, will arrive at this "easy" answer and, not being paranoid, will accept it as a sufficient interpretation.

But let us raise the bar and present our agents with a more difficult analysis task. Imagine the virtual patient, during a patient interview, receives the input, *Have you done any traveling lately?* (This example is not selected at random: it was featured in an unpublished OntoAgent demo system that focused on implementing and graphically displaying agent goal tracking, though the models supporting that system were less developed than the one presented here.) All of our agents will readily detect that this input is a question. Everyone apart from Ellie will search their memories for recent TRAVEL-EVENTS and report about them. These could include a day trip to visit Mom in Philly, a jaunt to the beach, or a shopping trip across the river in New Jersey. Ellie, by contrast, pursues mindreading in earnest, striving to be a highly efficient collaborator. She recognizes that TRAVEL-EVENT, unlike WEIGHT, is not part of her DOCTORS-VISIT or MEDICAL-EVENT scripts; so she will try to determine the doctor's goal in asking about travel. She searches her ontology for relationships between TRAVEL-EVENT and the condition for which she presented to the doctor – say, STOMACH-PAIN. The only close relationships she finds involve (a) TRAVEL-EVENTS with the INSTRUMENT AIRPLANE, and (b) TRAVEL-EVENTS with the DESTINATION 3RD-WORLD-COUNTRY. Ellie has no memories of having traveled in an airplane or having visited a 3rd world country, so she reasons that she has not engaged in any *relevant* travel events. She can select

various ways of formulating a response (the next step of processing), but its content will reflect the fact that she is responding to what she believes is the point of the question rather its literal content.

Of course, a doctor could have many different goals in asking about travel, including taking a break from his hectic workday by listening to a travel story from a patient; building trust with the patient through small talk; following up with a patient who had previously reported extreme travel-oriented anxiety; or finding out if the patient's chief complaint might be due to environmental contaminants. The string in isolation – even when situated in a doctor's office scenario – does not unambiguously convey an intention. But agents like Ellie prefer to err on the side of overanalyzing, whereas our beginning-of-the-alphabet individuals forego the extra effort.

At this point, our agents have generated the "mindreading-enhanced knowledge structures" at the bottom right of Figure 4. In some cases, those structures differ from the TMR generated earlier; in other cases – if mindreading was not attempted or failed – they are identical.

Next comes decision-making about what action to undertake in response to the input. This, too, can be influenced by many features. For example, there can be risks associated with misinterpreting an input, be they personal (*I'll be embarrassed if I jump to the wrong conclusion*) or objective (*Is the doctor telling me to take my medicine 5 times a* day *or 5 times a* week?); the agent may or may not have the option of seeking clarification before responding; the temporal pressure to respond might be great or the agent might be at liberty to postpone taking action at all; and so on.

Although we will not undertake to detail that decision-making process here, let us recall a key aspect of agent modeling discussed earlier: an end user has access to the inner life of an agent only through its actions. If our goal is to create a cohort of interestingly differentiated agents, we need to ensure that end users will recognize that differentiation. It would make little practical sense to hide Ellie's mindreading prowess behind such a fearful personality that she would never act upon the results of mindreading, instead retreating to the minimal-risk (under one interpretation of 'risk') strategy of perpetually asking the user for clarification. In short, the constellation of personal and contextual feature values driving agent action need to be correlated to ensure that agents display a range of behaviors that will optimally serve the application.

This broad range of plausible eventualities offers an interesting opportunity for agent development over time: it offers a mechanism that can buffer the effects of language processing errors that are an inevitable part of building sophisticated agent systems. As we saw, people don't demonstrate crack mindreading ability in every situation – misinterpreting the speaker's intention, or not guessing at it at all,

is quite normal, especially if one is stressed or in an unfamiliar situation. The MVP application concentrates on exactly such situations. As such, if virtual patients do not achieve the greatest heights of mindreading from the outset, that will, we hypothesize, seem entirely plausible. Of course, as the environment becomes ever more sophisticated it will be optimal to have the full range of mindreading-enhanced agents; but for a start, we can configure agents to have personality traits and be in mental and physical states that tolerate some degree of error. Maintaining users' engagement – suspension of disbelief, if you will – will be the litmus test of success.

3. Discussion

Building realistic, multifunctional, language-endowed intelligent agents is a long-term program of work whose utility relies on making agents sufficiently realistic to uphold a user's engagement with the given application. The proposed model of mindreading, as incorporated into the prototype MVP application, addresses two aspects of this requirement. First, it introduces a cognitively-inspired mechanism for generating a wide variety of behaviors across a population of agents experiencing different life circumstances. Second, it offers a plausible mechanism for deemphasizing cognitive insufficiencies demonstrated by agents: since many of the agents are experiencing medically-oriented stresses, they are not expected to be at their best. We are not suggesting that this escape hatch will account for all agent errors; however, a reasoning glitch by a virtual patient will, we think, cause less perturbation to the experience of a human user than, say, a reasoning glitch by an agent playing the role of a domain expert. Moreover, such a glitch may, in fact, faithfully reflect the behavior of a person under stress, which is a useful pedagogical feature of the MVP system.

Since human cognitive abilities are not directly measurable or inspectable, our modeling of complex cognitive processes derives from observation, introspection, and the goal of functional verisimilitude, following the spirit of abductive reasoning – that is, reasoning to the best available explanation. All of our models are human-inspectable and human-interpretable, meaning that if they fall short of functional sufficiency, they can be readily amended. This eliminates the risk of reaching a ceiling of results then having to jettison all work to date and start from scratch.

To place it in the larger context of scientific investigation, OntoAgent development can be classified as needs-driven, since we introduce whatever strategies we need in order to endow the agent with whatever functionalities it needs by means of explanatory models. By contrast, current psycholinguistic and fMRI-based

psychological work is data-driven: experimenters measure what they can, then interpret the results using hypotheses about human cognition. In both paradigms, interpretation is involved since cognition is not directly observable – a matter that is central to agent modeling work but tends to be downplayed in the data-driven paradigm.

Different aspects of mindreading have been pursued in different research paradigms. For example, belief revision (Bridewell & Bello 2013), mental simulation (Bello & Guarini 2010), emotion modeling (Hudlicka 2008; Plutchik 2001), and the ability to "ground" language utterances in a discourse context (Visser et al. 2012) all involve different aspects of mental model ascription. What sets the OntoAgent approach apart, we believe, is the overarching goal of better understanding how diverse agent functionalities can be integrated within a single knowledge environment.

Being in the prototype stages of development, MVP has not yet been formally evaluated, though many component parts have been implemented in demonstration systems. For example, the OntoAgent approach to language understanding has been under development for over 25 years and has been evaluated in various natural language processing applications; the interactive physiological simulations of several esophageal diseases have been inspected and enthusiastically endorsed by a number of expert physicians; one version of the MVP system was positively reviewed by a group of medical students who tested it under the supervision of system developers; and several other versions of the MVP system have been configured to demonstrate combined physiological and cognitive function, along with "under the hood", real-time reasoning not directly observable by system users.

For readers not familiar with the trajectory of work on knowledge-based systems, developing algorithms then implementing them in stepwise fashion over time, as resources permit, is natural. In building computational systems, the traditional estimate is that if the cost of developing a proof of concept system is X, then it will cost 10X to develop a prototype and 100X to deploy a commercial-level product. Building even proof of concept systems that implement knowledge-based explanatory models of agency is a complex and multifaceted undertaking. In this work one incurs high upfront costs. This is in contrast with the typical outlays of so-called "knowledge lean" systems, which are much faster and cheaper to configure and evaluate, but which do not pursue a corresponding level of agent sophistication; as a result, costs are incurred by system users whose needs are not, in the typical case, fully met. The main goal of this paper has been to introduce ideas about agent modeling that can be useful in full or in part to other agent systems, irrespective of the approach taken in their implementation.

Finally, let us explain why we are not dissuaded by the oft-repeated bogeyman of "the knowledge bottleneck", which has squelched the field's enthusiasm

for pursuing the original, ambitious goals of artificial intelligence. There actually is no bottleneck: if we want agents to perform sophisticated tasks, we have to prepare them to do so, and encoding high-quality knowledge suitable for machine reasoning is part of that work. By way of analogy, it is interesting to note that within the field of computer science, nobody is decrying the need for programmers to manually write programs. There are no broad-scale machine learning efforts (that we know of) aimed at replacing programmers. The only difference between writing computer programs and recording machine-tractable knowledge (which, in fact, is also known as knowledge programming) is that the latter has – for societal rather than scientific reasons – been the target of unproven conjectures about the potential for automation. However, history shows that the only way to avoid manually recording knowledge is to avoid developing applications that seek to emulate truly sophisticated human-like behavior. Considering that such applications have the potential to be of great service to society, we are happy to devote the necessary effort to recording the requisite knowledge.

Acknowledgments

Many thanks to the OntoAgent team, and particularly to Sergei Nirenburg for useful discussions of the material described in this paper. This research was supported in part by Grant N00014-09-1-1029 from the U.S. Office of Naval Research, which is not responsible for the paper's contents.

References

Bridewell, W., & Bello, P. (2013). Changing minds by reasoning about belief revision: A challenge for cognitive systems. *Proceedings of the 2nd Annual Conference on Advances in Cognitive Systems* (ACS), Baltimore, MD.

Bello, P., & Guarini, M. (2010). Introspection and mindreading as mental simulation. In S. Ohlsson, & R. Catrambone(Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Cognitive Science Society (pp. 2022–2028).

Hudlicka, E. (2008). What are we modeling when we model emotion? *Proceedings of the AAAI Spring Symposium: Emotion, Personality, and Social Behavior* (pp. 52–59).

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux. DOI: 10.1086/674372

Langley, P., Laird, J.E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160. DOI: 10.1016/j.cogsys.2006.07.004

McShane, M. (2009). Reference resolution challenges for an intelligent agent: The need for knowledge. *IEEE Intelligent Systems*, *24*(4), 47–58, July/Aug. DOI: 10.1109/mis.2009.79

McShane, M., Nirenburg, S., & Beale, S. (2005). An NLP lexicon as a largely language independent resource. *Machine Translation 19*(2), 139–173. DOI: 10.1007/s10590-006-9001-y

McShane, M., Fantry, G., Beale, S., Nirenburg, S., & Jarrell, B. (2007). Disease interaction in cognitive simulations for medical training. *Proceedings of the MODSIM World Conference*, Medical Track. Virginia Beach, Sept. 11–13.

McShane, M., Nirenburg, S., Beale, S., Jarrell, B., & Fantry, G. (2007). Knowledge-based modeling and simulation of diseases with highly differentiated clinical manifestations. In R. Bellazzi, A. Abu-Hanna, & J. Hunter (Eds.), *Proceedings of the 11th Conference on Artificial Intelligence in Medicine* (AIME 07) (pp. 34–43). Amsterdam, The Netherlands. Berlin, Heidelberg: Springer-Verlag. DOI: 10.1007/978-3-540-73599-1_4

McShane, M., Jarrell, B., Fantry, G., Nirenburg, S., Beale, S., & Johnson, B. (2008). Revealing the conceptual substrate of biomedical cognitive models to the wider community. In J.D. Westwood, R.S. Haluck, H.M. Hoffman, G.T. Mogel, R. Phillips, R.A. Robb et al. (Eds.), *Medicine meets virtual reality 16*, (pp. 281–286). Amsterdam, Netherlands: IOS Press.

McShane, M., Nirenburg, S., & Beale, S. (2008). Two kinds of paraphrase in modeling embodied cognitive agents. In A.V. Samsonovich (Ed.), *Biologically inspired cognitive architectures*. Papers from the AAAI Fall Symposium (pp. 162–167). Washington, DC, Nov. 7–9, 2008. AAAI Technical Report FS-08-04. Menlo Park, CA: AAAI Press.

McShane, M., & Nirenburg, S. (2009). Dialog modeling within intelligent agent modeling. In A. Jönsson, J. Alexandersson, D. Traum, & I. Zukerman (Eds.), *Proceedings of the IJCAI-09 Workshop on Knowledge and Reasoning in Practical Dialog Systems* (pp. 52–59). Pasadena, CA, USA.

McShane, M., Nirenburg, S., & Beale, S. (2011). Reference-related memory management in intelligent agents emulating humans. In P. Langley (Ed.), *Advances in cognitive systems*: *Papers from the AAAI fall symposium*. AAAI technical report FS-11-01, 232–239. Menlo Park, CA: AAAI Press.

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. DOI: 10.1016/j.artmed.2012.04.005

McShane, M., & Nirenburg, S. (2012). A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing, 06*(1), 3–23. DOI: 10.1142/s1793351x12400016

McShane, M., & Nirenburg, S. (2013). Use of ontology, lexicon and fact repository for reference resolution in Ontological Semantics. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New trends of research in ontologies and lexical resources* (pp. 157–185). Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-31782-8_9

McShane M, Nirenburg S, & Jarrell B. (2013). Modeling decision-making biases. *Biologically-Inspired Cognitive Architectures* (*BICA*) *Journal*, *3*, 39–50. DOI: 10.1016/j.bica.2012.09.001

Nirenburg, S., & McShane, M. (2012). Agents modeling agents: Incorporating ethics-related reasoning. *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". *Proceedings of the Workshop on Naturally Inspired Cognitive Architectures*, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7–9.

Nirenburg, S. & Raskin, V. (2004). *Ontological semantics*. Cambridge, Mass.: The MIT Press. DOI: 10.1007/s11023-008-9100-z

Nirenburg, S., McShane, M., Beale, S., English, J., & Catizone, R. (2010). Four kinds of learning in one agent-oriented environment. In A. V. Samsonovich, K.R. Jóhannsdóttir, A. Chella, & B. Goertzel. (Eds.), *Proceedings of the First International Conference on Biologically Inspired Cognitive Architectures* (pp. 92–97). Arlington, VA, Nov. 13–14, 2010. Amsterdam, Netherlands: IOS Press.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, *89*, 344–350. DOI: 10.1511/2001.28.739

Schank, R.C., & Abelson, R.P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: L. Erlbaum. DOI: 10.2307/412850

Visser, T., Traum, D., DeVault, D., & Akker, R. op den. (2012). Toward a model for incremental grounding in spoken dialogue systems. *Proceedings of the Workshop on Real-Time Conversations with Virtual Agents* (RCVA 2012), Santa Cruz, California.

Author's address

Marjorie McShane
Rensselaer Polytechnic Institute
Cognitive Science Department
Carnegie 108
110 8th Street
Troy, NY 12180
USA

mcsham2@rpi.edu

Author's biography

Marjorie McShane is an Associate Professor in the Cognitive Science Department of Rensselaer Polytechnic Institute. She works on knowledge-based approaches to configuring intelligent agents, with a particular interest in endowing agents with the ability to carry out reasoning-intensive aspects of language understanding, such as recovering elided material, grounding referring expressions in agent memory, and – as described in this paper – employing mindreading to foster collaboration with people.