Linguistics for the Age of AI

The title-level DOI for this work is:

doi:10.7551/mitpress/13618.001.0001

8

Agent Applications: The Rationale for Deep, Integrated NLU

The last chapter introduced some of the ways in which NLU is fostered by its integration in a comprehensive agent environment. In fact, it would be impossible to fully appreciate the need for ontologically grounded language understanding without taking into consideration the full scope of interrelated functionalities that will be required by human-level intelligent agents. All these functionalities rely on the availability of high-quality, machine-tractable knowledge, and this reality dwarfs the oft-repeated cost-oriented argument against knowledge-based NLU: that building the knowledge is too expensive. The fact is that agents need the knowledge *anyway*.

The second, equally compelling rationale for developing integrated, knowledge-based systems is that they will enable agents to explain their decisions in human terms, whether they are tasked with teaching, collaborating, or giving advice in domains as critical as defense, medicine, and finance. In fact, *explainable AI* has recently been identified as an important area of research. However, given that almost all recent work in AI has been statistically oriented, the question most often asked has been to what extent statistical systems can *in principle* explain their results to the human users who will ultimately be held responsible for the decision-making.

This chapter describes application areas that have served as a substrate for our program of work in developing LEIAs. As with the language-oriented chapters, the description is primarily conceptual, since specific system implementation details become ever more obsolete with each passing day. The goal of the chapter is to contextualize NLU in overall LEIA modeling without the discussion snowballing into a fundamental treatment of every aspect of cognitive systems.

8.1 The Maryland Virtual Patient System

Maryland Virtual Patient (MVP) is a prototype agent system that provides simulation-based experience for clinicians in training. Specifically, it would allow medical trainees to develop clinical decision-making skills by managing a cohort of highly differentiated virtual patients in dynamic simulations, with the optional assistance of a virtual tutor.

The benefits of simulation-based training are well-known: it offers users the opportunity to gain extensive practical experience in a short time and without risk. For example, "The evaluation of SHERLOCK II showed that technicians learned more about electronics troubleshooting [for US Airforce aircraft] from using this system for 24 hr than from 4 years of informal learning in the field" (Evens & Michael, 2006, p. 375).

Development of MVP followed the demand-side approach to system building, by which a problem is externally identified and then solved using whatever methods can be brought to bear. This stands in contrast to the currently more popular supply-side approach, in which the choice of a method—these days, almost always machine learning using big data—is predetermined, and R&D objectives are shaped to suit.

The physician-educators who conceived of MVP set down the following requirements:

1. It must expose students to virtual patients that demonstrate sophisticated, realistic behaviors, thus allowing the students to suspend their disbelief and interact naturally with them.
2. It must allow for open-ended, trial-and-error investigation—that is, learning through self-discovery—with the virtual patient's anatomy and physiology realistically adjusting to both expected and unexpected interventions.
3. It must offer a large population of virtual patients suffering from each disease, with each patient displaying clinically relevant variations on the disease theme; these can involve the path or speed of disease progression, the profile and severity of symptoms, responses to treatments, and secondary diseases or disorders that affect treatment choices.
4. It must be built on models with the following characteristics:
    a. They must be explanatory. Explanatory models provide transparency to the medical community who must endorse the system. They also provide the foundation for tutoring, since they make clear both the *what* and the *why* of the simulation.
    b. They must integrate well-understood biomechanisms with clinical knowledge (population-level observations, statistical evidence) that bridges the gaps when causal explanations are not available.
    c. They must allow these nonexplanatory *clinical bridges* to be replaced by biomechanical causal chains if they are discovered, without perturbation to the rest of the model.
    d. They must be sufficient to support automatic function and realism, but they need not include every physiological mechanism known to medicine. That is, creating useful applications does not impose the impossible precondition of creating full-blown virtual humans.
5. It must cover diseases that are both chronic and acute and both well and poorly understood by the medical community.

6. It must allow students to have control of the clock—that is, to advance the simulation to the next phase of patient management at will, thus simulating the doctor's choices about when a patient is to come for a follow-up visit.
7. It must offer optional tutoring support that can be parameterized to suit student preferences.
8. It must allow virtual patients to make all kinds of decisions that real patients do, such as when to see the doctor, whether to agree to tests and interventions, and whether to comply with the treatment protocol.

The virtual patients in MVP are *double agents* in that they display both physiological and cognitive function, as shown by the high-level system architecture in figure 8.1.[1] Physiologically, they undergo both normal and pathological processes in response to internal and external stimuli, and they show realistic responses to both expected and unexpected interventions. Cognitively, they experience symptoms, have lifestyle preferences, can communicate with the human user in natural language, have memories of language interactions and simulated experiences, and can make decisions (based on their knowledge of the world, their physical, mental, and emotional states, and their current goals and plans). An optional tutoring agent provides advice and feedback during the simulation. The other medical personnel include the agents that carry out tests and procedures and report their results.

It is noteworthy that the MVP vision and modeling strategy not only fulfill the desiderata for virtual patient models detailed in the National Research Council's 2009 joint report (Stead & Lin, 2009), but they were developed before that report was published. A short excerpt illustrates the overlap:

> In the committee's vision of patient-centered cognitive support, the clinician interacts with models and abstractions of the patient that place the raw data in context and

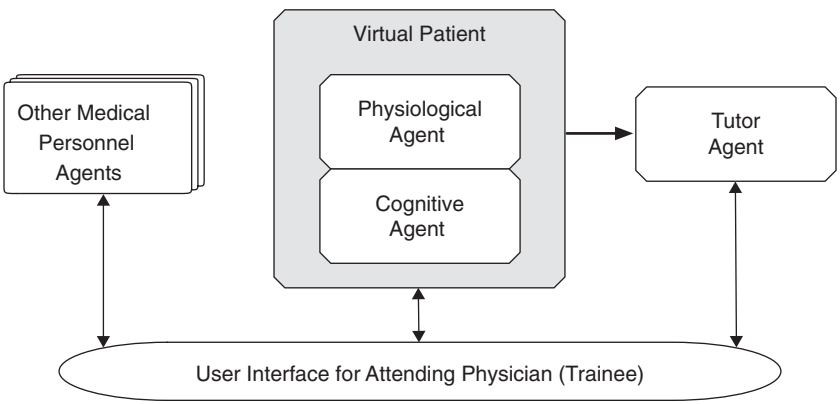

Figure 8.1
The Maryland Virtual Patient (MVP) architecture.

> synthesize them with medical knowledge in ways that make clinical sense for that patient. … These virtual patient models are the computational counterparts of the clinician's conceptual model of a patient. They depict and simulate a theory about interactions going on in the patient and enable patient-specific parameterization and multicomponent alerts. They build on submodels of biological and physiological systems and also of epidemiology that take into account, for example, the local prevalence of diseases. (p. 8)

MVP is a prototype system whose knowledge bases, software, and core theoretical and methodological foundations were developed from approximately 2005 to 2013 (e.g., McShane, Fantry, et al., 2007; McShane, Nirenburg, et al., 2007; McShane, Jarrell, et al., 2008; McShane, Nirenburg, & Jarrell, 2013; Nirenburg, McShane, & Beale, 2008a, 2008b, 2010a, 2010b). The system that was demonstrated throughout that period has not been maintained, but the knowledge bases, algorithms, methodology, and code remain available for reimplementation and enhancement. We refer to the system using the present tense to focus on the continued availability of the conceptual substrate and resources.

The obvious question is, *Why hasn't the work on MVP continued?* The reason is logistical: For a pedagogical system to be adopted by the medical community, large-scale evaluations—even of the prototype—are needed, and this is difficult to accomplish given the levels of funding typically available for research-oriented work. And without a formal evaluation of the prototype, it proved difficult to sustain sufficient funding to expand it into a deployed system. We still believe that MVP maps out an exciting and necessary path toward developing sophisticated, high-confidence, explanatory AI.

The description of MVP below includes the modeling of the virtual patient's physiology and cognition, a sample system run, the under-the-hood traces of system functioning, and a discussion of the extent to which such models can be automatically learned from the literature and extracted from domain experts. The descriptions attempt to convey the nature and scope of the work, without excessive detail that would be of interest only to experts in the medical domain.

Of course, ideally, system descriptions are preceded by demos—of which we had many during the period of development. In lieu of that, readers might find it useful to first skim through the system run described in section 8.1.4.

8.1.1 Modeling Physiology

The model of the virtual patient's physiology was developed in-house using the same ontology and metalanguage of knowledge representation as are used for NLU. Diseases are modeled as sequences of changes, over time, in the values of ontological properties representing aspects of human anatomy, physiology, and pathology. For each disease, some number of conceptual stages is established, and typical values (or ranges of values) for each property are associated with each stage. Values at the start or end of each stage are recorded

explicitly, with values between stages being interpolated. Disease models include a combination of fixed and variable features. For example, although the number of stages for a given disease is fixed, the duration of each stage is variable. Similarly, although the values for some physiological properties undergo fixed changes across patients (to ensure that the disease manifests appropriately), the values for other physiological properties are variable within a specified range to allow for different instances of virtual patients to differ in clinically relevant ways.

Roughly speaking, diseases fall into two classes: those for which the key causal chains are well understood and can drive the simulation, and those for which the key causal chains are not known. The models for the latter types of diseases rely on clinical observations about what happens and when (but not why). Most disease models integrate both kinds of modeling strategies in different proportions.

To develop computational cognitive models that are sufficient to support realistic patient simulations in MVP, a knowledge engineer leads physician-informants through the process of distilling their extensive and tightly coupled physiological and clinical knowledge into the most relevant subset and expressing it in the most concrete terms. Not infrequently, specialists are also called on to hypothesize about the unknowable, such as the preclinical (i.e., presymptomatic) stage of a disease and the values of physiological properties between the times when tests are run to measure them. Such hypotheses are, by nature, imprecise. However, rather than permit this imprecision to grind agent building to a halt, we proceed in the same way as live clinicians do: by developing *a* model that is reasonable and useful, with no claims that it is the only model possible or that it precisely replicates human functioning.[2]

The selection of properties to be included in a disease model is guided by practical considerations. Properties are included if (a) they can be measured by tests, (b) they can be affected by medications or treatments, and/or (c) they are central to a physician's mental model of the disease. In addition to using directly measurable properties, we also include abstract properties that foster the creation of a compact, comprehensible model. For example, when the property PRECLINICAL-IRRITATION-PERCENTAGE is used in scripts describing esophageal diseases, it captures how irritated a person's esophagus is before the person starts to experience symptoms. Preclinical disease states are not measured because people do not go to the doctor before they have symptoms. However, physicians know that each disease process has a preclinical stage, which must be accounted for in an end-to-end, simulation-supporting model. Inventing useful, appropriate abstract properties reflects one of the creative aspects of computational modeling.[3]

Once an approach to modeling a disease has been devised and all requisite details have been elicited from the experts, the disease-related events and their participants are encoded in ontologically grounded scripts written in the metalanguage of the LEIA's ontology.[4] MVP includes both domain scripts and workflow scripts. Domain scripts describe basic physiology, disease progression, and responses to interventions, whereas workflow scripts model the way an expert physician would handle a case, thus enabling automatic tutoring.

8.1.2 An Example: The Disease Model for GERD

GERD—gastroesophageal reflux disease—is one of the most common diseases worldwide.[5] It is any symptomatic clinical condition that results from the reflux of stomach or duodenal contents into the esophagus. In laymen's terms, acidic contents backwash into the esophagus because the sphincter between the two—called the lower esophageal sphincter (LES)—is not functioning properly. The acidity irritates the esophagus, which is not designed to withstand such acid exposure.

What follows is a summary of the model for GERD. Even if you choose to skip over the details, do notice that the modeling involves an explanatory, interpretive analysis of physiological and pathological phenomena, reflecting the way physicians think about the disease. This is not merely a compilation of factoids from the medical literature, which would not be sufficient to create an end-to-end, simulation-supporting model.

The development of any model begins by selecting the properties that define it. That selection process is informed by the descriptions provided by domain experts. The description of GERD begins with its cause: one of two abnormalities of the LES. Either the LES has an abnormally low basal pressure (< 10 mmHg) or it is subject to an abnormally large number or duration of so-called *transient relaxations*. Both of these result in the sphincter being too relaxed too much of the time, which increases acid exposure to the lining of the esophagus. Clinically speaking, it does not matter which LES abnormality gives rise to excessive acid exposure; what matters is the amount of time per day this occurs. We record this feature as the property TOTAL-TIME-IN-ACID-REFLUX.

Although TOTAL-TIME-IN-ACID-REFLUX earns its place in the model as the variable that holds the results of the test called pH monitoring, it does not capture—for physicians or knowledge engineers—relative GERD severity. For that we introduced the abstract property GERD-LEVEL. The values for GERD-LEVEL correlate (not by accident) with LES pressure:

- If GERD is caused by a hypotensive (too loose) LES, then the GERD-LEVEL equals the LES pressure. So, a GERD-LEVEL of 5 indicates an LES pressure of 5 mmHg.
- If GERD is caused by excessive transient relaxations, then the GERD-LEVEL reflects the same amount of acid exposure as would have been caused by the given LES pressure. So a GERD-LEVEL of 5 indicates a duration of transient relaxations per day that would result in the same acid exposure as an LES pressure of 5 mmHg.

Key aspects of the model orient around GERD-LEVEL (rather than LES pressure, transient relaxations, or TOTAL-TIME-IN-ACID-REFLUX) because this is much easier to conceptualize for the humans building and vetting the model. For example, as shown in table 8.1, GERD-LEVEL is used to determine the pace of disease progression, with lower numbers (think "a looser LES") reflecting more acid exposure and faster disease progression. (The full list covers the integers 0–10.)

Table 8.1
Sample GERD levels and associated properties

GERD-LEVEL	TOTAL-TIME-IN-ACID-REFLUX **in hours per day**	**Stage duration in days**
10	*less than 1.2*	*a non-disease state*
8	1.92	160
5	3.12	110
3	4.08	60

The conceptual stages of GERD are listed below. Each stage is associated with certain physiological features, test findings, symptom profiles, and anticipated outcomes of medical interventions. All these allow for variability across patients.

1. Preclinical stage: Involves the nonsymptomatic inflammation of the esophagus. It is called *preclinical* because patients do not present to doctors when they have no symptoms.
2. Inflammation stage: Involves more severe inflammation of the esophagus. Symptoms begin.
3. Erosion stage: One or more erosions (areas of tissue destruction) occur in the esophageal lining. Symptoms increase.
4. Ulcer stage: One or more erosions have progressed to the depth of an ulcer. Symptoms increase even more.
5. Post-ulcer stage, which takes one of two paths:
    a. Barrett's metaplasia: A premalignant condition that progresses to cancer (an additional stage) in some patients.
    b. Peptic stricture: An abnormal narrowing of the esophagus due to changes in tissue caused by chronic overexposure to gastric acid. It does not lead to cancer.

Patients differ with respect to the end stage of GERD if it is left untreated. Some lucky individuals will never experience more than an inflamed esophagus; their disease process simply stops at stage 2. By contrast, other patients will end up with esophageal cancer. For those patients progressing to the late stage of the disease, there is a bifurcation in disease path—Barrett's metaplasia versus peptic stricture—for reasons that are unknown.

The ontological scripts that support each stage of simulation include the patient's basic physiological property changes, how the patient will respond to interventions if the user (i.e., a medical trainee) chooses to administer them, and the effects of the patient's lifestyle choices. Sparing the reader the code in which scripts are written, here is an example,

in plain English, of how GERD progresses in a particular instance of a virtual patient who is predisposed to having erosion as the end stage of disease. In this example, the disease is left untreated throughout the entire simulation.

- During PRECLINICAL-GERD, the value of the property PRECLINICAL-IRRITATION-PERCENTAGE (an abstract property whose domain is MUCOSA-OF-ESOPHAGUS) increases from 0 to 100.[6]
- When the value of PRECLINICAL-IRRITATION-PERCENTAGE reaches 100, the script for PRECLINICAL-GERD is unasserted and the script for the INFLAMMATION-STAGE is asserted.
- During the INFLAMMATION-STAGE, the mucosal layer of the esophageal lining (recorded as the property MUCOSAL-DEPTH applied to the object ESOPHAGEAL-MUCOSA) is eroded, going from a depth of 1 mm to 0 mm over the duration of the stage.
- When MUCOSAL-DEPTH reaches 0 mm, the script for the INFLAMMATION-STAGE is unasserted, with the simultaneous assertion of the script for the EROSION-STAGE.
- At the start of the EROSION-STAGE, between one and three EROSION objects are created whose DEPTH increases from .0001 mm upon instantiation to .5 mm by the end of the stage, resulting in a decrease in SUBMUCOSAL-DEPTH (i.e., the thickness of the submucosal layer of tissue in the esophagus) from 3 mm to 2.5 mm.
- When SUBMUCOSAL-DEPTH has reached 2.5 mm., the EROSION-STAGE script remains in a holding pattern since the patient we are describing does not have a predisposition to ulcer.

Over the course of each stage, property values are interpolated using a linear function, though other functions could be used if they were found to produce more lifelike simulations. So, halfway through PRECLINICAL-GERD, the patient's PRECLINICAL-IRRITATION-PERCENTAGE will be 50, and three quarters of the way through that stage it will be 75.

The length of each stage depends on the patient's TOTAL-TIME-IN-ACID-REFLUX (see table 8.1). For example, a patient with a GERD-LEVEL of 8 will have a TOTAL-TIME-IN-ACID-REFLUX of 1.92 hours a day and each stage will last 160 days.

Some lifestyle habits, such as consuming caffeine, mints, and fatty foods, increase GERD-LEVEL manifestation in patients who are sensitive to those substances. In the model, if a patient is susceptible to GERD-influencing lifestyle habits and is engaging in those habits, then the effective GERD-LEVEL reduces by one. This results in an increase in acid exposure and a speeding up of each stage of the disease. If the patient is not actively engaging in the habit—for example, he or she might be following the doctor's advice to stop drinking caffeinated beverages—the GERD-LEVEL returns to its basic level. This is just one example of the utility of introducing the abstract property GERD-LEVEL into the model.

Each test that can be run is described in the ontology by the properties it measures, the clinically relevant ranges of values it can return, and expert interpretations of the results (see table 8.6 in section 8.1.5.2). When tests are launched on the patient at any time during

the simulation, their results are obtained by the system accessing the relevant feature values from the patient's dynamically changing physiological profile.

We now turn to two aspects of physiological modeling that we incorporated into the model after its initial implementation: (a) accounting for why patients have different end stages of the disease and (b) modeling partial (rather than all-or-nothing) responses to medications. The fact that we could seamlessly incorporate these enhancements, without perturbation to the base model, is evidence of the inherent extensibility of the models developed using this methodology.

Enhancement 1. *Accounting for why patients have different end stages of GERD.* Although it is unknown why patients have different end stages of GERD if the disease is left untreated, physicians have hypothesized that genetic, environmental, physiological, and even emotional factors could play a role.[7] To capture some hypotheses that have both practical and pedagogical utility, we introduced three abstract properties into the model:

- MUCOSAL-RESISTANCE reflects the hypothesis that patients differ with respect to the degree to which the mucosal lining of the esophagus protects the esophageal tissue from acid exposure and fosters the healing of damaged tissue. A higher value on the abstract {0,1} scale of MUCOSAL-RESISTANCE is better for the patient.
- MODIFIED-TOTAL-TIME-IN-ACID-REFLUX combines MUCOSAL-RESISTANCE with the baseline TOTAL-TIME-IN-ACID-REFLUX to capture the hypothesis that a strong mucosal lining can functionally decrease the effect of acid exposure. For example, patients with an average MUCOSAL-RESISTANCE (a value of 1) will have the stage durations shown in table 8.1. Patients with an above-average MUCOSAL-RESISTANCE (a value of greater than 1) will have a lower MODIFIED-TOTAL-TIME-IN-ACID-REFLUX, whereas patients with a below-average MUCOSAL-RESISTANCE (a value of less than 1) will have a higher MODIFIED-TOTAL-TIME-IN-ACID-REFLUX. For example:
  - If a patient's TOTAL-TIME-IN-ACID-REFLUX is 3.12 hours, but the patient has a mucosal resistance of 1.2, we model that as a MODIFIED- TOTAL-TIME-IN-ACID-REFLUX of 2.5 hours (3.12 multiplied by .8), and the disease progresses correspondingly slower.
  - By contrast, if the patient's TOTAL-TIME-IN-ACID-REFLUX is 3.12 hours, but the patient has a MUCOSAL-RESISTANCE of .8, then the MODIFIED-TOTAL-TIME-IN-ACID-REFLUX is 3.75 hours (3.12 multiplied by 1.2), and disease progression is correspondingly faster.
- DISEASE-ADVANCING-MODIFIED-TOTAL-TIME-IN-ACID-REFLUX is the total time in acid reflux required for the disease to manifest at the given stage. This variable permits us to indicate the end stage of a patient's disease in a more explanatory way than by simply asserting it. That is, for each patient, we indicate how much acid exposure is necessary to make the disease progress into each stage, as shown in table 8.2. If the acid exposure is not sufficient to support disease progression into a given stage (as

Table 8.2
Computing, rather than asserting, why patients have different end stages of GERD. Column 2 indicates each patient's MODIFIED-TOTAL-TIME-IN-ACID- REFLUX per day. The cells in the remaining columns indicate the total time in acid reflux needed for GERD to advance in that stage. Cells with gray shading indicate that the disease will not advance to this stage unless the patient's MODIFIED-TOTAL- TIME-IN-ACID-REFLUX changes—which could occur, for example, if the patient took certain types of medications, changed its lifestyle habits, or had certain kinds of surgery.

Patient	MODIFIED-TOTAL-TIME-IN-ACID-REFLUX	Preclinical	Inflammation	Erosion	Ulcer	Peptic stricture
John	1.92	1.92	1.92	2.3	2.5	3.12
Fred	2.8	1.92	1.92	2	2.7	3.12
Harry	4.08	1.92	1.92	3	3.5	4.0

shown by cells with gray shading), the patient's disease will be at its end stage. For example, John is a patient whose disease will not progress past the inflammation stage, even if left untreated, because his MODIFIED-TOTAL-TIME-IN-ACID-REFLUX is not high enough to support the erosion stage of GERD. By contrast, Fred's disease will advance into the ulcer stage, and Harry's disease will advance to peptic stricture.

Enhancement 2. *Modeling complete and partial responses to medication*. In order to capture the contrast between complete and partial responses to medications, medication effects are modeled as decreases in MODIFIED-TOTAL-TIME-IN-ACID-REFLUX, as shown in table 8.3. The table indicates the decrease in acid exposure caused by each medication for each patient, along with the resulting MODIFIED-TOTAL-TIME-IN-ACID-REFLUX. Explained in plain English:

- For each day that John takes an H2 blocker, his MODIFIED-TOTAL-TIME-IN-ACID-REFLUX will be 1.42, which is not a disease state. If he already has the disease, healing will occur. The other, more potent medication regimens will also be effective for him.
- For Fred, the H2 blocker is not sufficient to promote complete healing (it brings the MODIFIED-TOTAL-TIME-IN-ACID-REFLUX down to 2.5), but it would be sufficient to not permit his disease to progress to the ulcer stage. Or, if Fred were already in the ulcer stage, the ulcers would heal to the more benign level of erosions. If Fred took a PPI once or twice daily, his MODIFIED-TOTAL-TIME-IN-ACID-REFLUX would be < 1.92, meaning that his esophagus would heal completely over time.
- For Harry, the H2 blocker would barely help at all—he would still progress right through the stricture stage. Taking a PPI once a day would heal ulcers and block late stages of disease. Taking a PPI twice a day would heal the disease completely, unless Harry had already experienced a stricture: there is no nonoperative cure for a peptic

Table 8.3
Modeling complete and partial responses to medications. The reduction in MODIFIED-TOTAL-TIME-IN-ACID-REFLUX is listed first, followed by the resulting MODIFIED-TOTAL-TIME-IN-ACID-REFLUX in brackets.

Patient	MODIFIED-TOTAL-TIME-IN-ACID-REFLUX	H2 blocker	PPI once daily	PPI twice daily
John	1.92	– .5 [1.42]	– 1.25 [.67]	– 1.5 [.42]
Fred	2.8	– .3 [2.5]	– 1 [1.8]	– 2.25 [.55]
Harry	4.08	– .1 [3.98]	– .8 [3.28]	– 2.2 [1.88]

stricture, a detail that we will not pursue at length here but which is covered in the model (the STRICTURE object generated by the simulation remains a part of the patient's anatomy).

To recap, these enhancements to the original GERD model permit each patient's end stage of disease progression to be calculated rather than asserted, and they permit medications to have varying degrees of efficacy.

One important point remains before we wrap up this overview of disease modeling. Any disease that has known physiological preconditions will arise any time those preconditions are met. For example, say a virtual patient is authored to have the disease achalasia, which is caused by a hypertensive LES (the opposite of GERD). And say a system user chooses to treat the achalasia using a surgical procedure that cuts the LES, changing it from *hyper*tensive to *hypo*tensive. Then the disease processes of GERD will automatically begin because the LES-oriented precondition has been met. There is no need for the person authoring the achalasia patient to say anything at all about GERD. This example illustrates why physiological models should be as causally grounded as possible, particularly as more and more interventions are added to the environment, making available all kinds of side effects outside those pertaining to the given disease.

8.1.3 Modeling Cognition

Virtual patients need many cognitive capabilities. Their language understanding capabilities have already been amply described. Their language generation involves two aspects: generating the content of what they will say, and generating its form. The content derives from reasoning and is encoded in ontologically grounded meaning representations. The form is constructed by templates, which proved sufficient for the prototype stage of this application but would need to be enhanced for a full-scale application system. Two other necessary cognitive capabilities of virtual patients are (a) learning new words and concepts through language interaction and (b) making decisions about action. We consider these in turn.

8.1.3.1 Learning new words and concepts through language interaction Learning is often a prerequisite to decision-making. After all, no patient—real or virtual—should agree to a medical procedure without knowing its nature and risks. Table 8.4 shows a brief dialog, which was demonstrated in the application system, between a virtual patient (P) and the human user playing the role of doctor (D). This dialog features the learning of ontology and lexicon through language interaction in preparation for the patient's decision-making about its medical treatment.

When the virtual patient processes each of the doctor's utterances, it automatically creates text meaning representations that it then uses for reasoning and learning. The text meaning representation for the first sentence is

ACHALASIA-1
  EXPERIENCER HUMAN-1 ("the virtual patient")

Table 8.4
Learning lexicon and ontology through language interaction

Dialog	Ontological knowledge learned	Lexical knowledge learned
D: You have achalasia.	The concept ACHALASIA is learned and made a child of DISEASE.	The noun *achalasia* is learned and mapped to the concept ACHALASIA.[8]
P: Is it treatable? D: Yes.	The value for the property TREATABLE in the ontological frame for ACHALASIA is set to *yes*.	
D: I think you should have a Heller myotomy.	The concept HELLER-MYOTOMY is learned and made a child of MEDICAL-PROCEDURE. Its property TREATMENT-OPTION-FOR receives the filler ACHALASIA.	The noun *Heller myotomy* is learned and mapped to the concept HELLER-MYOTOMY.
P: What is that? D: It is a type of esophageal surgery.	The concept HELLER-MYOTOMY is moved in the ontology tree: it is made a child of SURGICAL-PROCEDURE. Also, the THEME of HELLER-MYOTOMY is specified as ESOPHAGUS.	
P: Are there any other options? D: Yes, you could have a pneumatic dilation instead, …	The concept PNEUMATIC-DILATION is learned and made a child of MEDICAL-PROCEDURE.	The noun *pneumatic dilation* is learned and mapped to the concept PNEUMATIC-DILATION.
D: … which is an endoscopic procedure.	PNEUMATIC-DILATION is moved from being a child of MEDICAL-PROCEDURE to being a child of ENDOSCOPY.	
P: Does it hurt? D: Not much.	The value of the property PAIN-LEVEL in PNEUMATIC-DILATION is set to .2 (on a scale of 0–1).	

The patient knows to make ACHALASIA a child of DISEASE in the ontology because the lexical sense it uses to process the input "You have X" asserts that X is a DISEASE. This sense is prioritized over other transitive meanings of the verb *have* because the discourse context is a doctor's appointment and the speaker is a doctor. A similar type of reasoning suggests that a Heller myotomy is some sort of MEDICAL-PROCEDURE. Our short dialog also shows two examples of belief revision: when the virtual patient learns more about the nature of the procedures HELLER-MYOTOMY and PNEUMATIC-DILATION, it selects more specific ontological parents for them, thereby permitting the inheritance of more specific property values.[9]

8.1.3.2 Making decisions about action Virtual patients carry out dynamic decision-making in a style that approximates human decision-making—at least to the degree that we can imagine how human decision-making works. For example, whenever a decision needs to be made, the virtual patient first determines whether it has sufficient information to make it—an assessment that is based on a combination of what it actually knows, what it believes to be necessary for making a good decision, and its personality traits. If it lacks some knowledge it needs to make a decision, it can posit the goal of obtaining this knowledge, which is a metacognitive behavior that leads to learning.

Formally speaking, a goal is an ontological instance of a property, whose domain and range are specified. Goals can appear on the agent's goal agenda in four ways:

- *Perception via interoception*. The moment the patient perceives a symptom, the symptom appears in its short-term memory. This triggers the addition of an instance of the goal BE-HEALTHY onto the agenda. We assume that achieving the highest possible value of BE-HEALTHY (1 on the abstract scale {0,1}) is a universal goal of all humans, and in cases in which it seems that a person is not fulfilling this goal, he or she is simply prioritizing another goal, such as EXPERIENCE-PLEASURE.
- *Perception via language*. Any user input that requires a response from the virtual patient (e.g., a direct or indirect question) puts the goal to respond to it on the agenda.
- *A precondition of an event inside a plan is unfulfilled*. For example, most patients will not agree to an intervention about which they know nothing. So, one of the events inside the plan of decision-making about an intervention is finding out values for whichever features of it are of interest to the individual.
- *The required period of time has passed since the last instances of the events BE-DIAGNOSED or BE-TREATED have been launched*. This models regular checkups and scheduled follow-up visits for virtual patients.

The goal BE-HEALTHY is put on the agenda when a virtual patient begins experiencing a symptom. It remains on the agenda and is reevaluated when (a) its intensity or frequency (depending on the symptom) reaches a certain level, (b) a new symptom arises, or (c) a

certain amount of time has passed since the patient's last evaluation of its current state of health, given that the patient has an ongoing or recurring symptom or set of symptoms: that is, "I've had this mild symptom for too long. I should see a doctor."

When making decisions about its health care, the virtual patient considers the following types of features, which are used in the decision-making evaluation functions described below.

1. Its physiological state (particularly the intensity and frequency of symptoms), which is perceived via interoception and remembered in its memory. It is important to note that neither the patient nor the virtual tutor in the MVP system have omniscient knowledge of the patient's physiological state. The simulation system knows this, but the intelligent agents functioning as humans do not.
2. Certain character traits: TRUST, SUGGESTIBILITY, and COURAGE. The inventory can, of course, be expanded as needed.
3. Certain physiological traits: PHYSIOLOGICAL-RESISTANCE, PAIN-THRESHOLD, and the ABILITY-TO-TOLERATE-SYMPTOMS. These convey how intense or frequent symptoms have to be before the patient feels the need to do something about them.
4. Certain properties of tests and procedures: PAIN, UNPLEASANTNESS, RISK, and EFFECTIVENESS. PAIN and UNPLEASANTNESS are, together, considered typical side effects when viewed at the population level. The patient's personal individual experience of them is described below.
5. Two time-related properties: the FOLLOW-UP-DATE, that is, the time the doctor told the patient to come for a follow-up, and the CURRENT-TIME of the given interaction.

Most of these properties are scalar attributes whose values are measured on the abstract scale {0,1}.[10] All subjective features are selected for each individual virtual patient by the patient author. That is, at the same time as a patient author selects the physiological traits of the patient—such as the patient's response to treatments if they are administered—he or she selects certain traits specific to the cognitive agent, as well as the amount of relevant world knowledge that the patient has in its ontology. Two evaluation functions, written in a simple pseudocode, will suffice for illustration.

Evaluation function 1. *SEE-MD-OR-DO-NOTHING.* This function decides when a patient goes to see the doctor, both initially and for follow-up visits.

IF FOLLOW-UP-DATE is not set
  AND SYMPTOM-SEVERITY > ABILITY-TO-TOLERATE-SYMPTOMS
  THEN SEE-MD
; *This triggers the first visit to the doctor.*

ELSE IF FOLLOW-UP-DATE is not set
AND SYMPTOM-SEVERITY < ABILITY-TO-TOLERATE-SYMPTOMS
AND the SYMPTOM has persisted > 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
ELSE DO-NOTHING
; *There was a big increase in symptom severity from low to high, exceeding the patient's ability to tolerate these symptoms. This triggered an unplanned visit to the doctor.*

ELSE IF there was a previous visit
AND at the time of that visit SYMPTOM-SEVERITY is between .3 and .7
AND currently SYMPTOM-SEVERITY >.9
AND [SYMPTOM-SEVERITY − ABILITY-TO-TOLERATE-SYMPTOMS] > 0
THEN SEE-MD
ELSE DO-NOTHING
; *There was a big increase in symptom severity from medium to very high, triggering an unplanned visit to the doctor.*

ELSE IF there was a previous visit
AND at the time of that visit SYMPTOM-SEVERITY >.7
AND currently SYMPTOM-SEVERITY >.9
THEN DO-NOTHING
; *Symptom severity was already high at the last visit—do not do an unplanned visit to the doctor because of it.*

ELSE IF the TIME reaches the FOLLOW-UP-TIME
THEN SEE-MD
; *Go to previously scheduled visits.*

ELSE DO-NOTHING

As should be clear, patients with a lower ability to tolerate symptoms will see the doctor sooner in the disease progression than patients with a higher ability to tolerate symptoms, given the same symptom level. Of course, one could incorporate any number of other character traits and lifestyle factors into this function, such as the patient's eagerness to be fussed over by doctors, the patient's availability to see a doctor around its work schedule,

and so on. But even this inventory allows for considerable variability across patients—plenty, in fact, to support rigorous training of future physicians.

Evaluation function 2. *AGREE-TO-AN-INTERVENTION-OR-NOT.* Among the decisions a patient must make is whether or not to agree to a test or procedure suggested by the doctor, since many interventions carry some degree of pain, risks, side effects, or general unpleasantness. Some patients have such high levels of trust, suggestibility, and courage that they will agree to anything the doctor says without question. All other patients must decide whether they have sufficient information about the intervention to make a decision and, once they have enough information, they must decide whether they want to (a) accept the doctor's advice, (b) ask about other options, or (c) reject the doctor's advice. A simplified version of the algorithm for making this decision (which suffices for our purposes) is as follows:

IF a function of the patient's TRUST, SUGGESTIBILITY, and COURAGE is above a threshold OR the RISK associated with the intervention is below a threshold (as for a blood test)
THEN the patient agrees to intervention right away.

ELSE [*] IF the patient feels it knows enough about the RISKS, SIDE-EFFECTS, and UNPLEASANTNESS of the intervention (as a result of evaluating the function DETERMINE-IF-ENOUGH-INFO-TO-EVALUATE)
AND a call to the function EVALUATE-INTERVENTION establishes that the above risks are acceptable
THEN the patient agrees to the intervention.

ELSE IF the patient feels it knows enough about the RISKS, SIDE-EFFECTS, and UNPLEASANTNESS of the intervention
AND a call to the function EVALUATE-INTERVENTION establishes that the above risks are not acceptable
THEN the patient asks about other options.
IF there are other options
THEN the physician proposes them and control is switched to [*].
ELSE the patient refuses the intervention.

ELSE IF the patient does not feel it knows enough about the intervention (as a result of evaluating the function DETERMINE-IF-ENOUGH-INFO-TO-EVALUATE)
THEN the patient asks for information about the specific properties that interest it, based on its character traits (e.g., a cowardly patient will ask about RISKS, SIDE-EFFECTS, and UNPLEASANTNESS, whereas a brave but sickly person might only ask about SIDE-EFFECTS).
IF a call to the function EVALUATE-INTERVENTION establishes that the above RISKS are acceptable

THEN the patient agrees to the intervention.
ELSE the patient asks about other options
IF there are other options
THEN the physician proposes them and control is switched to [*].
ELSE the patient refuses the intervention.

This evaluation function makes use of two functions that we do not detail here, EVALUATE-INTERVENTION and DETERMINE-IF-ENOUGH-INFO-TO-EVALUATE (see Nirenburg et al., 2008a). These details are not needed as our point is to illustrate (a) the kinds of decisions virtual patients make, (b) their approach to knowledge-based decision-making, and (c) the kinds of dialog that must be supported to simulate the necessary interactions.

8.1.4 An Example System Run

To illustrate system operation, we present a sample interaction between a medical trainee named Claire and a virtual patient named Michael Wu. *Sample* is the key word here, as there are several substantially different paths, and countless trivially different paths, that this simulation could take based on what Claire chooses to do. She could intervene early or late with clinically appropriate or inappropriate interventions, or she could do nothing at all; she could ask Mr. Wu to come for frequent or infrequent follow-ups; she could order appropriate or inappropriate tests; and she could have the tutor set to intervene frequently, only in cases of imminent mistakes, or not at all. However, since Mr. Wu is a particular *instance* of a virtual patient, he has an inventory of property values that define him, which put some constraints on the available outcomes of the simulation. His physiological, pathological, psychological, and cognitive profile is established before the session begins, using the patient-creation interface described in section 8.2.5.1.

Psychological traits: trust [.2], suggestibility [.3], courage [.4]

Physiological traits: physiological resistance [.9], pain threshold [.2], ability to tolerate symptoms [.4]

Knowledge of medicine: minimal, meaning that the patient does not know the features of any interventions the user might propose

Disease(s) explicitly authored for this patient:[11] achalasia

Duration of each stage of the disease: preclinical [7 months], stage 1 [7 months], stage 2 [8 months], stage 3 [8 months], stage 4 [9 months]

Response to treatments if they are launched: BoTox [effective, wearing off over 12 months], pneumatic dilation [effective with regression], Heller myotomy [effective permanently]

Claire does not have direct access to any of this information and must learn everything about Mr. Wu through dialog, tests, and procedures.

When Claire launches the simulation, she must wait for Mr. Wu to present to the office. He makes this decision using the decision function in section 8.1.3.2. We use numbers in square brackets to indicate the key points of this simulation run.

[1] Mr. Wu presents with the chief complaint "difficulty swallowing." This is day 361 of the progression of his disease, which includes the preclinical stage and a portion of the first symptomatic stage. (Claire, of course, will not know this temporal information.) Mr. Wu has had symptoms for some time but until now the evaluation function SEE-DOCTOR-OR-DO-NOTHING has returned the answer DO-NOTHING.

[2] When Mr. Wu presents at the office, this initiates the first dialog with Claire. She types in unconstrained English text (note the extensive use of elliptical expressions). Mr. Wu (being a virtual patient) analyzes it into TMRs, makes a decision about how to answer, and generates a response. The first interaction runs as follows:

Claire : So, you have difficulty swallowing?
Mr. Wu : Yes.
Claire : Do you have difficulty swallowing solid food?
Mr. Wu : Yes.
Claire : Liquids?
Mr. Wu : No.
Claire : Do you have chest pain?
Mr. Wu : Yes, but it's mild.
Claire : Any heartburn?
Mr. Wu : No.
Claire : Do you ever regurgitate your food?
Mr. Wu : No.
Claire : How often do you have difficulty swallowing?
Mr. Wu : Less than once a week.
Claire : It's too early to take any action. Please come back in 9 months.
Mr. Wu : OK.

As we see, Claire decides to do nothing—an important kind of decision in clinical medicine, and one that is difficult to teach since a doctor's natural response to a patient asking for help is to do *something*.

[3] After nine months (on day 661 of the disease progression) Mr. Wu comes back for his follow-up. The cognitive simulation engine has regularly been running the evaluation function SEE-DOCTOR-OR-DO-NOTHING (since he is still symptomatic), but it has always returned DO-NOTHING—that is, do not schedule a new appointment before the scheduled

follow-up. Claire again asks Mr. Wu about his difficulty swallowing, chest pain, and regurgitation, using paraphrases of the original formulations (for variety and, in system demonstrations, to show that this is handled well by the NLU component). Mr. Wu responds that he has moderate chest pain, experiences regurgitation a few times a week, and has difficulty swallowing solids daily and liquids occasionally. Note that the progression of difficulty swallowing from solids to liquids is a key diagnostic point that the user should catch: this suggests a motility disorder rather than an obstructive disorder.

[4] Claire posits the hypothesis that Mr. Wu has a motility disorder and advises Mr. Wu to have a test called an EGD (esophagogastroduodenoscopy). Mr. Wu evaluates whether he will accept this advice using the function EVALUATE-INTERVENTION, described in section 8.1.3.2. Since he is concerned about the risks, he asks about them. When Claire assures him that they are extremely minimal, he agrees to the procedure.

[5] A lab technician agent virtually runs the test and delivers the results. This involves querying the physiological model underlying the simulation at the given point in time. A specialist agent returns the results with the interpretation: "Narrowing of LES with a pop upon entering the stomach. No tumor in the distal esophagus. Normal esophageal mucosa." These results include both positive results and pertinent negatives.

[6] Claire reviews the test results, decides that it is still too early to intervene, and schedules Mr. Wu for another follow-up in four months.

[7] When Mr. Wu presents in four months and Claire interviews him, the symptom that has changed the most is regurgitation, which Mr. Wu now experiences every day. Note that throughout the simulation the patient chart is automatically populated with responses to questions, results of tests, and so on, so Claire can compare Mr. Wu's current state with previous states at a glance.

[8] Claire suggests having another EGD and Mr. Wu agrees immediately, not bothering to launch the evaluation function for EGD again since he agreed to it the last time.

[9] Then Claire suggests having two more tests: a barium swallow and esophageal manometry. Mr. Wu asks about their risks (that remains his only concern about medical testing), is satisfied that they are sufficiently low, and agrees to the procedures. Lab technicians and specialist agents are involved in running the tests and reporting results, as described earlier. The barium test returns "Narrowing of the lower esophageal sphincter with a bird's beak," and the manometry returns "Incomplete relaxation of the LES, hypertensive LES, LES pressure: 53."

[10] Claire decides that these test results are sufficient to make the diagnosis of achalasia. She records this diagnosis in Mr. Wu's chart.

[11] Claire suggests that Mr. Wu have a Heller myotomy. He asks about the risks and pain involved. Claire responds that both are minimal. Mr. Wu agrees to have the procedure. Claire tells him to come back for a follow-up a month after the procedure.

[12] Mr. Wu has the procedure.

[13] Mr. Wu returns in a month, Claire asks questions about symptoms, and there are none. She tells Mr. Wu to return if any symptoms arise.

8.1.5 Visualizing Disease Models

If a cognitive modeling strategy and the applications it supports are to be accepted by researchers, educators, and domain experts, it is important that the knowledge substrate be transparent. We cannot expect professors at medical schools to adopt technologies based on opaque knowledge when they are responsible for the competence of the physicians they train. In MVP, the need for transparency was addressed in three ways: (a) by encapsulating each disease model used to author instances of patients with that disease; (b) by organizing in human-readable, tabular form the types of knowledge that extend beyond what is captured by the patient-authoring interface; and (c) by graphically displaying traces of system functioning for purposes of system demonstration. We consider these visualization capabilities in turn. (Note that although all of the visualizations to be described were implemented in interactive interfaces, the reproduction quality of those screenshots was suboptimal, making it preferable to convey the material here using other expressive means. Examples of actual interfaces are available in McShane, Jarrell, et al. (2008) and Nirenburg et al. (2010a), as well as at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html).

8.1.5.1 Authoring instances of virtual patients The virtual patients that users interact with are *instances* that are spawned from a single, highly parameterizable ontological model of patients experiencing the given disease. Authors of patient instances—who could be professors in medical schools, system developers, or even students preparing practice cases for their study partners—create patient instances by selecting particular values for variable features in the model. All patient models include basic information including name, age, gender, height, weight, and select personality traits. Beyond that, the nature of the model depends on the nature of the disease being modeled.

To illustrate the patient-authoring process, we use the esophageal disease called achalasia, introduced earlier. As a reminder, it involves the opposite physiological abnormality as GERD—namely, a *hyper*tensive LES. We switch from GERD to achalasia for two reasons. First, this provides a glimpse into a different class of disease: unlike GERD, achalasia has an unknown etiology, so the disease model derives from population-level clinical observations. Second, the achalasia model is more easily encapsulated using visualizations.

Each patient-authoring session opens with a short description of the disease to refresh the memory of patient authors. Methods of progressive disclosure—for example, displaying a portion of explanatory texts in a relatively small window with scroll bars offering the rest—permit users of different profiles to interact with the interface efficiently. The explanatory texts are not only remedial reminders. Instead, they describe key aspects of

the modeling strategy to everyone interacting with the interface—including, importantly, specialists whose mental model might be different from the one implemented in the system.

Table 8.5 shows patient-authoring choices involving stage duration, physiological properties, and symptoms for achalasia. Property values in plain text are fixed across patient instances, whereas those in square brackets are variable. The actual value shown in each set of brackets is the editable default. In the dynamic interface, the legal range of values was shown by rolling over the cell. This amount of variability allows for a wide range of patient profiles while still ensuring that disease progression remains within clinically observed patterns. However, given other teaching goals or new clinical evidence, the choice of variable versus fixed features could be changed with no need to alter the simulation engine. There is more variability in symptom profiles than in the physiological model itself, reflecting the clinical observation that different patients can perceive a given, test-confirmed physiological state in very different ways.

The patient author also indicates—using pull-down menus and editable text fields presented in tables similar to table 8.5—how the virtual patient will respond to treatments, should they be administered at each stage of the disease. Details aside, two of the treatment options—pneumatic dilation and Heller myotomy—have three potential outcomes: unsuccessful, successful with regression, and successful with no regression. In the case of success with regression, the rate of regression is selected by the patient author. The third treatment option, BoTox, always involves regression, but the rate can vary across patients. These authoring options are provided with extensive explanations because only specialists are expected to remember all the details.

To summarize, the patient-authoring interface for each disease provides patient authors with an encapsulated version of the disease model along with the choice space for patient parameterization. It does not repeat all the information about each disease available in textbooks or attempt to elucidate every detail of implementing the simulation engine. The grain size of description—including which aspects are made parameterizable and which physiological causal chains are included in the model—is influenced by the judgment calls of the domain experts participating in system development.

8.1.5.2 The knowledge about tests and interventions Throughout this chapter, we have been presenting ontological knowledge about clinical medicine in tables, which are a method of knowledge representation understandable by domain experts, knowledge engineers, and programmers alike. Specifically, such tables are readable enough to be vetted by domain experts and formal enough to be converted into the ontological metalanguage by knowledge engineers for programmers. Among the kinds of table-based knowledge that are not displayed in the patient-creation interface are those shown in tables 8.6–8.8.

Table 8.6 includes the test name (in some cases, multiple tests can measure the same property), sample results, and a specialist's interpretation of those results. Results are

Table 8.5
Patient-authoring choices for the disease achalasia

	Start	t0	t1	t2	t3	t4
Stage duration (in months)		[12]	[12]	[12]	[12]	[12]
Physiological Properties						
Ratio of relaxing to contracting neurons in the distal esophagus	100/100	80/100	60/100	40/100	20/100	10/100
Basal LES pressure (torr)	[25]	[start-value] + 8	[start-value] + 16	[start-value] + 24	[start-value] + 32	[start-value] + 40
Residual LES pressure (torr)	0	8	16	24	32	40
Residual LES diameter (cm)	2.0	1.5	1.0	.5	0.0	0.0
Amplitude of contraction during peristalsis	80	65	40	30	20	10
Efficacy of peristalsis	peristalsis	peristalsis	peristalsis	Intermittent	aperistalsis	aperistalsis
Diameter of distal esophagus (cm)	2	2.8	3.6	4.2	5	6
Retained esophageal content (on the abstract scale {0,1})	0	[.1]	[.3]	[.55]	[.85]	[1]
Emptying delay (min)	0	1	5	10	30	35,000[12]
Symptoms						
Difficulty swallowing distal	0	0.1	[1]	[2]	[3]	[4]
Do solids stick?	No	no	Yes	Yes	yes	yes
Do liquids stick?	No	no	No	Yes	yes	yes
Weight loss	0	0	[0]	[0]	[.1]	[.2]
Chest pain	0	0	[.1]	[.3]	[.5]	[.7]
Regurgitation (times/month)	0	0	[0]	[10]	[40]	[70]

Table 8.6
Examples of ontological knowledge about tests relevant for achalasia

Test	Sample results (presented informally)	Specialist's interpretation
EGD or BARIUM-SWALLOW	LES diameter <= .5	"Narrowing of the LES with a pop upon entering the stomach"
EGD or BARIUM-SWALLOW	LES diameter .5–1.5	"Subtle narrowing of the LES"
EGD or BARIUM-SWALLOW	Diameter of lower body of esophagus: 3–4 cm	"Slightly dilated esophagus"
EGD or BARIUM-SWALLOW	Diameter of lower body of esophagus: >= 4 cm	"Moderately dilated esophagus"
ESOPHAGEAL-MANOMETRY	LES pressure at rest: > 45 torr	"Hypertensive LES"
ESOPHAGEAL-MANOMETRY	LES pressure at rest: 35–45 torr	"High-normal LES pressure"
BARIUM-SWALLOW	Duration of swallowing: 1–5 mins	"Slight delay in emptying"
BARIUM-SWALLOW	Duration of swallowing: > 5 mins	"Moderate-severe delay in emptying"

Table 8.7
Examples of knowledge that supports clinical decision-making about achalasia, which is used by the virtual tutor in the MVP system

PROPERTY	Values (presented in plain English for readability)
SUFFICIENT-GROUNDS-TO-DIAGNOSE	All three of the following conditions: 1. Either a bird's beak (a visual test finding) or a hypertensive LES 2. Aperistalsis 3. Negative esophagogastroduodenoscopy (EGD) for cancer (i.e., a pertinent negative)
SUFFICIENT-GROUNDS-TO-TREAT	Definitive diagnosis

Table 8.8
MVP. Knowledge about the test results expected at different stages of the disease achalasia. Used by the tutoring agent in MVP. The test results in italics are required to definitively diagnose the disease.

Test Name	t1	t2	t3	t4
EGD	Dilated esophagus & no tumor at the GI junction		Dilated esophagus & *narrowing and pop upon entering LES (i.e., a hypertensive LES)* & retained debris & *no tumor at the GI junction*	
Esophageal Manometry	Incomplete relaxation of the LES & high-normal LES pressure	Incomplete relaxation of the LES & high-normal or hypertensive LES & intermittent peristalsis	Incomplete relaxation of the LES & *hypertensive LES* & *aperistalsis*	

Table 8.9
Inventory of under-the-hood panes that are dynamically populated during MVP simulation runs

Physiology	Interoception	Thoughts	Knowledge learned	TMRs
A list of disease-relevant property value pairs, with values being highlighted every time they change during the simulation. Reflects an omniscient view of the patient's physiology.	A list of the virtual patient's perceived symptoms as property value pairs. Every time a symptom appears or changes, a new entry is posted.	Dynamically populated traces of the patient's decision functions, rendered in plain English for readability. E.g., "I don't know the risks of EGD. I'd better ask about them."	Traces of words and concepts learned through dialog. The words are mapped to concepts, and the concepts are placed appropriately in the ontological hierarchy.	Text meaning representations of the virtual patient's interpretations of the user's inputs during the simulated doctor-patient interactions.[13]

expressed informally for readability. The actual results are written in the ontological metalanguage.

Our point in presenting these tables is to emphasize that it is important to make the collaboration between the domain experts, knowledge engineers, and programmers explicit, organized, and easily modifiable. In other words, the knowledge representation must serve many masters.

8.1.5.3 Traces of system functioning Dynamic traces of system functioning are shown in what we call the under-the-hood panes of MVP. The inventory of panes is shown in table 8.9, along with brief descriptions of what they contain. The panes are presented as columns since this is how they are rendered in the demonstration system—that is, all of them can be viewed at the same time. (As a reminder, sample screenshots are shown at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html.)

The under-the-hood panes of the MVP environment are key to showing that the simulations are real—there is no hand-waving; nothing is hidden. They can also be used to pedagogical ends, for example, by allowing students to view the physiological changes during disease progression and the effects of medical interventions.

8.1.6 To What Extent Can MVP-Style Models Be Automatically Learned from Texts?

In the current climate of big data and machine learning, a natural question is, *To what extent can models like these be automatically learned from texts?*[14] The answer: Only very partially. Full models cannot be automatically learned, or even cobbled together by diligent humans, from the literature because they do not exist in the literature. However, we think that some *model components* could be automatically extracted. We define model components as ontologically grounded property value pairs that contribute to full models. Learnable properties have the following characteristics:

- They are straightforward and concrete, such as LES-PRESSURE (measurable by a test) and SENSITIVITY-TO-CAFFEINE (knowable based on patient reports). Learnable properties cannot be abstract, like our MODIFIED-TOTAL-TIME-IN-ACID-REFLUX or MUCOSAL-RESISTANCE, because abstract properties will certainly have no equivalents in published texts.
- They are known to be changeable over time, based on our ontological knowledge of the domain. For example, since we know that new medications and tests are constantly being invented, we know that the properties TREATED-BY-MEDICATION and ESTABLISHED-BY-TEST must have an open-ended inventory of values. By contrast, we do not expect to have to change the fact that heartburn can be a symptom of GERD or that HEARTBURN-SEVERITY is best modeled as having values on the abstract scale {0,1}.
- They describe newly discovered causal chains that can replace clinical bridges in a current model. By contrast, if the model already includes causal chains that fully or partially overlap, their modification is likely to be too complex to be learned automatically without inadvertently perturbing the model.[15]

Table 8.10 shows some examples of properties—associated with their respective concepts—whose values we believe could be learned from the literature.

In order for a model component to be fully learned, the property and the fillers for its domain and range must be ontological entities, not words of language. LEIAs can, in principle, produce these using NLU. For example, all of the following text strings, and many more, will result in text meaning representations that include the knowledge GASTROESOPHAGEAL-REFLUX-DISEASE (HAS-TREATMENT PROTON-PUMP-INHIBITOR):

- A proton pump inhibitor treats <can treat, can be used to treat, can be prescribed to treat, is often prescribed to treat> GERD.
- GERD is <can be> treated <cured> by (taking) a proton pump inhibitor.
- Doctors <Your doctor may> recommend <prescribe> (taking) a proton pump inhibitor to treat GERD symptoms.
- If you have GERD, you might <may> be advised to take a proton pump inhibitor.

Table 8.10
Examples of properties, associated with their respective concepts, whose values can potentially be automatically learned from the literature

Concept	Properties
DISEASE	HAS-EVENT-AS-PART, AFFECTS-BODY-PART, CAUSED-BY, HAS-SYMPTOMS, HAS-DIAGNOSTIC-TEST, HAS-TREATMENT
DIAGNOSTIC-TEST	MEASURES-PROPERTY, NORMAL-RESULT, ABNORMAL-RESULT, SIDE-EFFECTS, PAIN-INDUCED
MEDICAL-TREATMENT	HAS-EVENT-AS-PART, EFFICACY, HAS-RISKS, PAIN-INDUCED

Establishing the functional equivalence of these strings would not be done by listing. Instead, it would be done by combining our general approach to natural language understanding with methods for paraphrase detection and ontologically grounded reasoning.[16]

Let us consider just three examples of how natural language understanding could support the automatic learning of disease model components. Assume that the LEIA is seeking to automatically learn or verify the correctness of the previously discussed fact GASTROESOPHAGEAL-REFLUX-DISEASE (HAS-TREATMENT PROTON-PUMP-INHIBITOR). As we said, all the inputs above provide this information, albeit some more directly than others. The input *GERD is treated by a proton pump inhibitor* perfectly matches the lexical sense for the verb *treat* that is defined by the structure DISEASE *is treated by* MEDICATION, and the analyzer can generate exactly the text meaning representation we are seeking: GASTROESOPHAGEAL-REFLUX-DISEASE (HAS-TREATMENT PROTON-PUMP-INHIBITOR).

In other cases, the basic text meaning representation includes additional information that does not affect the truth value of the main proposition. For example, the potential modality scoping over the proposition *GERD can be treated by a proton pump inhibitor* does not affect the truth value of the main proposition, which is the same as before and matches the expectation we seek to fill.

In still other cases, the meaning we are looking for must be inferred from what is actually written. For example, the input *Your doctor may recommend a proton pump inhibitor* does not explicitly say that a proton pump inhibitor treats GERD, but it implies this based on the general ontological knowledge that a precondition for a physician advising a patient to take a medication is DISEASE (HAS-TREATMENT MEDICATION). Because a LEIA's language understanding system has access to this ontological knowledge, it can be taught to make the needed inference and fill in our slot as before. It should be noted that these types of reasoning rules are not spontaneously generated—they must be recorded, like any other knowledge. However, once recorded, they can be used for any applicable reasoning need of the agent.

When we were investigating what information could be extracted from medical texts in service of disease-model development, we focused on two genres that offer different opportunities for knowledge extraction: case studies and disease overviews.

Case studies do not present all disease mechanisms. Instead, they typically begin with a broad overview of the disease to serve as a reminder to readers who are expected to be familiar with it. Then they focus on a single new or unexpected aspect of the disease as manifest in one or a small number of patients. For example, Evsyutina et al.'s (2014) case study reports that a mother and daughter both suffer from the same rare disease, achalasia, and suggests that this case supports previous hypotheses of a genetic influence on disease occurrence. The new findings are typically repeated in the abstract, case report, and discussion sections, offering useful redundancy to improve system confidence.

A LEIA could set to the task of comparing the information in a case study with the ontologically grounded computational model as follows. First it could semantically analyze

the case study, focusing on the TMR chunks representing the types of learnable property values listed above. (This focusing means that the system need not achieve a perfect analysis of every aspect of the text: it knows what it is looking for.) Then, it could compare the learned property values with the values in the model. Continuing with our example of mother-daughter achalasia, our current model of achalasia has no filler for the value of CAUSED-BY since, when we developed the model, the cause was not definitively known (it still is not; the genetic influence remains to be validated). Automatically filling an empty slot with a new filler can be carried out directly, with no extensive reasoning necessary. However, the nature of that slot filler must be understood: in the context of a case study, it represents an instance, not a generic ontological fact. The system has two sources of evidence that this information is an instance: (a) the individuals spoken about are instances, so the features applied to them are also instances (compare this with assertions about people in general), and (b) the genre of case study sets up the expectation that reported information will be at the level of an instance.

Such analysis could, for example, be folded into an application to alert clinicians to new findings in a snapshot formalism like the one shown below (invented for illustration):

Journal article: "Meditation as medication for GERD"

Contribution type:	Case study
Author:	Dr. Joseph Physician
Date:	*Some future date*

GERD Therapies:
Non-medical: lifestyle modifications, **MEDITATION-new**
Mild: H2 blocker, PPI QD
Severe: PPI BID

This presentation style encapsulates the following expectations:

1. Clinicians know, without explanation, that one of the ontological properties of diseases is that they have therapies.
2. When providing new information, it is useful to provide old information as the backdrop, with a clear indication of whether the new information adds to or overwrites the old information.
3. Clinicians understand that information provided in case studies represents instances and not across-the-board generalizations.
4. Modern-day users understand that entities can be clicked on for more information (e.g., which lifestyle modifications are being referred to).
5. Terseness is appreciated by busy people operating within their realm of specialization.

Let us turn now to the other genre from which model information can be extracted: disease overviews. Disease overviews typically present a stable inventory of properties of interest, often even introduced by subheadings, such as causes of the disease, risk factors, physiological manifestations, symptoms, applicable tests and procedures, and so on. Not surprisingly, these categories align well with the knowledge elements we seek to extract from texts, shown in table 8.10. The natural language processing of disease overviews would proceed as described for case studies. However, we envision applications for this processing to be somewhat different. For example, an application could respond to a clinician's request for a thumbnail sketch of a disease by reading overviews, populating the inventory of key property values, and presenting them in a semiformal manner, such as a list of concept-property-value triples.

To wrap up this section on learning components of disease models, note how different the sketched approaches are from statistically oriented knowledge extraction. Our goal would be to speed up, and dynamically enhance, cognitively inspired disease models, not extract uninterpreted text strings into templates that have no connection to ontologies or related cognitive models.

8.1.7 To What Extent Can Cognitive Models Be Automatically Elicited from People?

Text processing is only one of the available methods of reducing the role of knowledge engineers in the process of domain modeling. Another is to guide domain experts through the process of recording components of disease models using a mixed-initiative computer system.[17] The results can then seed the collaborative process between the experts and knowledge engineers.[18]

The strategy for the methodology we describe below, OntoElicit, was informed by two things: lessons learned from developing the first several disease models for MVP through unstructured and semistructured interviews with domain experts, and our past work on a mixed-initiative knowledge elicitation system in a different domain—machine translation.

Let us present just a passing introduction to the latter. The Boas system (McShane et al., 2002; McShane & Nirenburg, 2003) was designed to quickly gather machine-tractable knowledge about lesser-studied languages from native speakers of those language without the assistance of linguists or system developers. The results of the knowledge-elicitation process had to directly feed into a machine translation system from that language into English. By "directly feed into" we mean that, once the user supplied the requested information, he or she pushed a button, waited a minute, and ended up with a translation system. Since developers were completely out of the loop once they delivered the environment to the user, the elicitation process and associated interface had to ensure that the necessary knowledge would get recorded in the right way and that the environment itself provided users with sufficient pedagogical support. The informants, for their part, were not expected to have any formal linguistic knowledge, just the ability to read and write the language in question, as well as a functional knowledge of English. The automatically elicited knowledge

was not identical to what could be crafted if knowledge engineers were involved in the process, but it was sufficient to enable basic machine translation capabilities to be configured in this way. Change the domain from language to medicine, the experts from native speakers to physicians, and the goal from machine translation to seeding ontological models for clinical medicine, and Boas smoothly morphs into OntoElicit.

The knowledge elicitation methods of OntoElicit, shown below, will look familiar to readers as they share much in common with the patient-authoring interfaces and knowledge representation schemes described earlier.

In OntoElicit, domain experts are asked to divide the disease into any number of conceptual stages correlating with important events, findings, symptoms, or the divergence of disease paths among patients. They are also asked to indicate the typical duration of each stage as a range (x–y in table 8.11) with a default value (d). Next, they are led through the process of describing the relevant physiological and symptom-related properties during each stage. They can either record all information directly in a table like table 8.11 or be led through a more step-by-step process that results in a summary like table 8.11. Following a practice we invented for Boas, we call the former the *fast lane* and the latter the *scenic route*. Both paths offer links explaining the *why* and *how* of the associated decision-making, as well as examples.

In describing tests and their results, the expert indicates the test name, alternative names, which physiological properties are measured, clinically relevant ranges of results, the specialist's interpretation of those ranges (e.g., "Suggestive of disease X"), clinical guidelines regarding ordering the test, and diseases for which the test is appropriate.

For interventions, including medications, the expert indicates which properties and/or symptoms are affected by the intervention, the possible outcomes of the intervention, possible side effects, and, if known, the percentage of the population expected to have each outcome and side effect.

Table 8.11
Fast-lane elicitation strategy for recording information about physiology and symptoms

	Properties	Start value	Stage 1	Stage 2	…
Duration			x–y (d)	x–y (d)	
Physiology	P1	x–y (d)	x–y (d)	x–y (d)	
	P2	x–y (d)	x–y (d)	x–y (d)	
	…				
Symptoms	S1	x–y (d)	x–y (d)	x–y (d)	
	S2	x–y (d)	x–y (d)	x–y (d)	
	…				

Note: "x–y" indicates the acceptable range of values; (d) indicates the default. If "x" and "y" are nonnumerical, they are presented as a list.

As concerns recording knowledge about clinical practices—that is, the knowledge to support automatic tutoring—two different functionalities must be supported: checking the validity of a clinical move, which is relatively simple and relies on the knowledge of *preconditions of good practice*, and advising what to do next, which can range from simple to very complex.

The knowledge about preconditions of good practice is readily encoded using ontological properties. For example, for each disease, we record values for properties such as SUFFICIENT-GROUNDS-TO-SUSPECT, SUFFICIENT-GROUNDS-TO-DIAGNOSE, and SUFFICIENT-GROUNDS-TO-TREAT (e.g., clinical diagnosis or definitive diagnosis). Similar inventories of properties are used for tests, treatments, making definitive diagnoses, and so on. The content of this knowledge is both broader and deeper than that available in published "best practices" guides. OntoElicit uses tables for eliciting this information (see table 8.12), with the experts providing prose descriptions of property fillers. These descriptions are then converted—like all other aspects of acquired knowledge—into formal, ontologically grounded structures by knowledge engineers and programmers.

As concerns clinical knowledge about what to do next, things can get complicated quickly. Many clinical moves must be decided (a) in the face of competing conditions, (b) with different preferences of different stakeholders (e.g., the patient, the physician, the insurance company), and (c) using incomplete knowledge of relevant property values. For those cases, we have experimented with the use of Bayesian networks that are constructed with the help of influence diagrams.[19] The knowledge encoded in influence diagrams represents an expert's opinion about the utility scores (i.e., the preference level, or "goodness") of different combinations of property values associated with each possible decision. One of the main reasons why we chose to work with influence diagrams is that the kind of information required of experts is of a nature that they can readily conceptualize. In essence, they are asked: *Given this combination of property values, how good is solution Y? Given this other combination of property values, how good is solution Z?* and so on. The properties and values are familiar to our experts because they are the same

Table 8.12
Sample precondition of good practice. Domain experts supply the descriptive fillers and knowledge engineers convert it into a formal representation.

DISEASE	ACHALASIA
PROPERTY	SUFFICIENT-GROUNDS-TO-SUSPECT
Descriptive filler	solid and liquid dysphagia or regurgitation
Formal encoding	(or (and ((SOLIDS-STICK HUMAN YES) (LIQUIDS-STICK HUMAN YES)) (REGURGITATION-FREQUENCY HUMAN (> 0)))))

ones used to build the other models in the system. Knowledge engineers help experts to organize the problem space into subproblems, as applicable, and to develop a case-specific methodology of filling out the utility tables in the most efficient way.

Although the nature of information required of experts in an influence-diagram-driven methodology is straightforward, one problem is that the number of features involved in making a complex decision can be large, easily driving the number of feature-value permutations into the tens or hundreds of thousands. As in all aspects of modeling, we approach this problem using realistic strategies including the following:

1. We organize the knowledge optimally—for example, covering as many variables as possible using local decisions whose output contributes to a more general decision.
2. We simplify the problem space and judge whether the results are sufficient to yield realistic, accurate functioning—for example, not including every property we can think of but, instead, focusing on those considered to have the most impact by clinicians.
3. We work toward automating the process of knowledge acquisition—for example, using functions to provide values for many of the feature-value combinations once a pattern of utility scores has been recognized.[20]

As regards incorporating aspects of influence diagram creation into OntoElicit, our thinking is that experts could, in fact, be led through the process of decomposing the problem into the main variables in the decision versus the variables in local decisions.

We have not yet experimented with how far we can push a mixed-initiative elicitation strategy in the domain of clinical medicine. However, considering that we covered a lot of ground with the Boas predecessor, and considering that the realm of language description is arguably no easier than clinical medicine, we believe that this approach has great potential to be useful.

This wraps up our discussion of the MVP application which was, as we mentioned, implemented at a prototype level. We now move to a model that has not yet been implemented but relies largely on the same knowledge substrate as MVP.

8.2 A Clinician's Assistant for Flagging Cognitive Biases

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 necessary human abilities: the ability to simplify complex problems, make decisions despite incomplete information (called decision-making under uncertainty), and generally function under the real-world constraints of limited time, information, and cognitive capacity (cf. Simon's [1957] theory of bounded rationality). Factors that contribute to cognitive biases include, nonexhaustively: overreliance 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.)

Even if one recognizes that cognitive biases could be affecting decision-making, their effects can be difficult to counteract. As Heuer (1999, Chapter 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 detrimental consequences, particularly in a high-stakes domain like clinical medicine. We hypothesize that at least some errors in judgment caused by some cognitive biases could be reduced if LEIAs serving as clinician advisors were able to detect potentially biased decisions and generate explanatory alerts to their human collaborators. Even such partial solutions to very difficult problems have the potential to offer rewards at the societal level.

The bias-related functionalities we will address and the psychological phenomena they target are summarized in table 8.13.[21]

Table 8.13
Functionalities of a bias-detection advisor in clinical medicine

Advisor functionalities	Targeted decision-making biases
Memory support: Supplying facts the clinician requests using text generation, structured presentation of knowledge (e.g., checklists), process simulation, and so on	• Depletion effects
Detecting and flagging potential clinician biases	• Illusion that more features are better • False intuitions • Jumping to conclusions • Small sample bias • Base-rate neglect • Illusion of validity • Exposure effect
Detecting and flagging potential patient biases	• Framing sway • Halo effect • Exposure effect • Effects of evaluative attitudes

In discussing each class of bias, we will present (a) a theory of how to model cognitive support to avoid the bias, which involves the selection of properties and values to be treated (e.g., bias types), detection heuristics, decision functions, and knowledge support, and (b) the descriptive realization of the theory as a set of models compatible with LEIA modeling overall.

8.2.1 Memory Support for Bias Avoidance

Memory lapses are unavoidable in clinical medicine due to not only the large amount of knowledge that physicians must manipulate but also *depletion effects*—that is, the effects of fatigue. We believe that depletion effects could be decreased with timely, ergonomically presented reminders, cribs, and checklists (Gawande, 2009) that reflect particular aspects of the knowledge already available in a LEIA's expert models. This type of cognitive assistance would be 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 et al., 2009). Consider just a few situations in which a LEIA's knowledge could be leveraged to counter clinician memory lapses. Let us use as our example primary care physician Dr. Allegra Clark.

Example 1. It's the end of the day, Dr. Clark is tired, and she forgets some basic ontological properties of a disease or treatment. She queries the LEIA with an English string such as, *What are the symptoms of achalasia?* The LEIA semantically analyzes this input, converting it into the following text meaning representation.

```
REQUEST-INFO-1
    AGENT        PHYSICIAN-1
    THEME        ACHALASIA.CAUSES-SYMPTOM-1
```

This TMR says that this input is requesting the fillers of the CAUSES-SYMPTOM property of ACHALASIA. The LEIA can answer the question by looking up the needed information in its ontology, the relevant portion of which is shown below.

```
ACHALASIA
    CAUSES-SYMPTOM    sem    CHEST-PAIN, DYSPHAGIA, REGURGITATE, WEIGHT-LOSS
```

Example 2. Dr. Clark wants to order the test called EGD (esophagogastroduodenoscopy) but forgets what preconditions must hold to justify this. She queries the LEIA with *What's needed to diagnose achalasia?* As before, the LEIA translates the input into a TMR and understands that the answer will be the filler of the property SUFFICIENT-GROUNDS-TO-DIAGNOSE in the ontological concept ACHALASIA. Table 8.14 shows a subset of properties of the disease ACHALASIA that relate to diagnosis and treatment. For presentation purposes, the property values in the right-hand column are presented in plain English rather than the ontological metalanguage.

Table 8.14
Four clinical properties of the esophageal disease achalasia, with values written in plain English for readability

PROPERTY	Values
SUFFICIENT-GROUNDS-TO-DIAGNOSE	All three of the following: 1. Either a bird's beak (a visual test finding) or a 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)

Example 3. Dr. Clark knows that the disease achalasia can have different manifestations in different patients but forgets the details and asks the LEIA to display the ontologically grounded disease model for achalasia. As explained earlier, all disease models are available in the human-inspectable formats shown in section 8.1.5.1, which the LEIA displays.
Example 4. A patient asks Dr. Clark for a prognosis, but she is too tired, too rushed, or not familiar enough with the disease to provide a well-motivated answer. The LEIA could help by permitting her to run one or more simulations of virtual patients that are constrained by the known features of the human in question. This will make the sample simulations as predictive as possible given the coverage and accuracy of the underlying models.

The above four examples should suffice to convey our main point: the knowledge structures and simulation capabilities already developed for the MVP application can be directly reused to help clinicians to counteract memory lapses or knowledge gaps. For this category of phenomena, developing models that take into account biases involves anticipating the requests of clinicians and optimizing the presentation of already available knowledge to make it easily interpretable by them. The initiative for seeking this class of bias-avoidance support lies in the hands of the clinician-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.

8.2.2 Detecting and Flagging Clinician Biases

Diagnosing a patient typically begins with a patient interview and a physical examination. Next, the clinician posits a hypothesis and then attempts to confirm it through medical testing or trial therapy (e.g., lifestyle changes or medication). Confirming a hypothesis by testing leads to a definitive diagnosis, whereas 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 it. 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."[22] One point at which clinicians might erroneously—and at great expense—believe that more feature values are necessary is during diagnosis: they might not recognize that they already have sufficient information to diagnose a disease. For many diseases, clear diagnostic criteria exist, like that shown in the first row of table 8.14. If the patient chart shows sufficient evidence to diagnose a disease, but the clinician has not posited the diagnosis and has ordered more tests, a LEIA could issue an alert about the possible oversight.

Jumping to conclusions. The opposite of seeking too many features is jumping to conclusions, as by 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 combination of italicized test results shown in the third and fourth disease stages shown in table 8.15. Positing a diagnosis prior to obtaining the full set of definitive values could be incorrect. Whenever a clinician posits a diagnosis, a LEIA could 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, for example, in Kahneman and Klein (2009)—we define skilled intuition as the recognition of constellations of highly predictive property values based on sufficient past experience. Nobody can have reliable intuitions (a) about unknowable situations, (b) in the absence of reliable feedback, or (c) 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 the ones above, which would assume that they exhaust valid medical knowledge. A more sophisticated approach would be to incorporate a LEIA'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 the LEIA will be justified in flagging

Table 8.15
Knowledge about expected test results during progression of achalasia

	Stage 1	Stage 2	Stage 3	Stage 4
EGD	Dilated esophagus & no tumor at the GI junction		Dilated esophagus *& narrowing and pop upon entering LES (i.e., hypertensive LES)* & retained debris *& no tumor at the GI junction*	
Esophageal manometry	Incomplete relaxation of the LES & high-normal LES pressure	Incomplete relaxation of the LES & high-normal or hypertensive LES & intermittent peristalsis	Incomplete relaxation of the LES *& hypertensive LES* *& aperistalsis*	
Barium swallow	Delayed emptying		*Bird's beak* & dilated esophagus & retained debris & retained barium	

Note: The combination of italicized results is sufficient to posit a definitive diagnosis of this disease.

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, the LEIA might better query him about the reason for the move and potentially learn this new constellation of findings and their predictive power. This aspect of *system-initiated learning by being told* is a core functionality of LEIAs.

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 nonpredictive—that is, 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. (He or she might, for example, rerun tests or continue a failed medication trial.) The definition of *sufficient* counterevidence depends on (a) the strength of the constellation of features suggesting the diagnosis; (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 must be recorded in the ontology. A LEIA could detect overzealous pursuit of a hypothesis using decision functions that combine these three factors.

Base-rate neglect. 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 of given different constellations of signs and symptoms can be quite challenging. A LEIA could help with this by flagging situations in which a clinician 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 MEDICAL-PATIENT-1) (DURATION (> 5 (measured-in YEAR)))
            - At least one of
              • (MEDICAL-PATIENT-1 (AGENT-OF SMOKE))
              • (MEDICAL-PATIENT-1 (AGENT-OF (DRINK (THEME ALCOHOL) (FREQUENCY (> .3)))))
              • (MEDICAL-PATIENT-1 (AGENT-OF (RESIDE (LOCATION INDUSTRIAL-PLACE))))
              • (MEDICAL-PATIENT-1 (AGENT-OF (WORK (LOCATION INDUSTRIAL-PLACE))))
              • (MEDICAL-PATIENT-1 (EXPERIENCER-OF (EXPOSE (THEME CARCINOGEN) (FREQUENCY
                (> .3)))))
```

If a clinician hypothesizes esophageal carcinoma for a twenty-year-old person with a three-month history of GERD, the LEIA should 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 with 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 (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."

A LEIA could automatically detect the small sample bias in clinicians' decisions by comparing three things: (a) the clinician's current clinical decision, (b) the LEIA's memory of the clinician's past decisions when dealing with the particular disease, and (c) the objective, population-level preference for the selected decision compared to other options. For example, suppose that three of a clinician's recent patients with a particular disease did not respond sufficiently to the preferred treatment or developed complications from it. If the clinician then stops recommending that treatment and, instead, opts for a less preferred one, the LEIA can issue a reminder of the population-level preference for the originally selected treatment and point out there is a danger of a small sample bias. Of course, the actual reason for the switch in treatment preferences might be legitimate. For example, if the treatment involves a procedure carried out by a specialist, then perhaps a highly skilled specialist was replaced by a less skilled one—which is an eventuality that must be modeled 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 (p. 67). The LEIA can detect potential cases of the exposure effect using a function whose arguments include the following.

- A new ontological property, HYPE-LEVEL, that applies to interventions—drugs and procedures. Its values reflect the amount of advertising, drug company samples, and so on to which a clinician is exposed. If this is unknown for a particular clinician, a population-level value will be used, based on the amount of overall advertising and sample distribution.
- 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, and so on.
- 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 clinician's past history of prescribing, or not prescribing, this intervention in relevant circumstances. For example, a clinician might (a) be continuing to prescribe an

old medication instead of a better new one due to engrained past experience, (b) insist on a name brand if a generic has been made available, or (c) prefer one company's offering over a similar offering from another company despite high additional costs to the patient; and so on.

8.2.3 Detecting and Flagging Patient Biases

The gold standard of modern medical care is patient-centered medicine. In the patient-centered paradigm, the physician does not impose a single solution on 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 makes responsible decisions. Launch an effective treatment."

To best serve the patient, the doctor 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 or she believes the patient knows, believes, fears, prioritizes, and so on and by hypothesizing the decision function that might have led to the given decision. For example, imagine that a doctor suggests that a patient, Matthew, take a medication that the doctor knows to be highly effective and that has infrequent, mild side effects about which the doctor informs Matthew. In response to the doctor's suggestion, Matthew refuses, saying he doesn't want to take that kind of medication. When the doctor asks why, Matthew responds in a vague manner, saying that he just has a bad feeling about it. Rather than try to force Matthew 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 things like the following: the list of potential benefits, risks, and side effects; the cost, in terms of money, time, emotional drain; 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 doctor's care, and so on.

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 Matthew 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, Matthew will have the following

information to inform his decision-making. 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, both patients and doctors know that the following can affect health care decisions:

COST-FINANCIAL	*FLAG if no health insurance*
TRUST-IN-PHYSICIAN	*FLAG if the doctor feels the patient doesn't trust him*
MEDICATION-AVERSION	*FLAG for certain individuals and socio-ethnic groups*

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

- *The exposure effect*. People 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 a medication might involve a vague but lengthy inventory of side effects that the doctor did not mention, and these might serve as misinformation for the patient's decision-making.
- *The effect of small samples*. The patient might know somebody who took this medication and had a bad time with it, thus generalizing that it is a bad drug, despite the doctor's description of it.
- *The effect of evaluative attitudes*. The patient might not like the idea of taking any medication, or some class of medications, due to a perceived stigma (e.g., against antidepressants). Or the patient might be so opposed to a given type of side effect that its potential overshadows any other aspect of the medication.
- *Depletion effects*. The patient might be tired or distracted when making a decision and therefore decide that refusing a proposed intervention is the least-risk option. Or fatigue might have caused lapses in attention so that the patient misremembers the doctor's description of the medication.

A LEIA could assist the physician in trying to understand the patient's decision-making by making the relevant flags explicit. For example, if our patient, Matthew, has 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 it is unlikely that the financial cost

or fear of injections is a detractor. Since the LEIA will have access to Matthew's online medical records, it can make such contextual judgment calls and give the doctor advice about which features might be best to pursue first. Even things like a patient's trust in the doctor can, we believe, be detected to some degree by the doctor-patient dialog. For example, if Matthew argues with the doctor, or asks a lot of questions, or frequently voices disagreement, it is possible that low trust is affecting his decision-making.

Another factor 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, you might believe that a person who is kind and successful will also be generous, even though you know nothing about this aspect of the person's 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 apply also to objects and events—can undermine good decision-making by patients. On the one hand, our patient Matthew might like his doctor so much he agrees to the latter's advice before learning a sufficient amount about what is recommended to make a responsible, informed decision. On the other hand, he might dislike his doctor so much that he refuses advice that would actually be beneficial. Extending the halo effect to events, Matthew might be so happy that a procedure has few risks that he assumes that it will not involve any pain and will have no side effects—both of which might not be true. By contrast, Matthew might be so influenced by the knowledge that the procedure will hurt that he loses sight of its potential benefits. Doctors should detect halo effects in order to ensure that patients are making the best, most responsible decisions for themselves. It would be no better for Matthew to blindly undergo surgery because he likes his doctor than for him to refuse lifesaving surgery because he is angry with him or her.

In order to operationalize the automatic detection of halo effects, we can construct *halo-property nests* like the ones shown in table 8.16. 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

Table 8.16
Example of halo-property nests

OBJECT OR EVENT	Nest of PROPERTIES
MEDICAL-PROCEDURE	RISK, PAIN, SIDE-EFFECTS, BENEFITS
PHYSICIAN	INTELLIGENCE, SKILL-LEVEL, AFFABILITY, KINDNESS, TRUSTWORTHINESS

(e.g., low risk) but doesn't know about any of the other property values in the nest, it is possible that he or she will assume that the values of the other properties have the same halo-polarity 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 doctor to tactfully continue a knowledge-providing conversation until the patient actually has all the information needed to make a good decision. The agent'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, notify the doctor.

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 impact a person's perception of it and subsequently affect related decision-making. For example, if someone is asked, "I imagine you hurt right now?" they will have a tendency to seek corroborating evidence by noticing something that hurts, even if just a little (*the confirmation bias*). If someone is asked, "Your pain is very bad, isn't it?" they are likely to overestimate the perceived pain, having been primed with a high pain level (*the priming effect*). And if someone is told, "There is a 20% chance that this will fail," they are likely to interpret it more negatively than if they were told, "There's an 80% chance that this will succeed" (*the framing sway*).

The agent could help doctors be aware of, and learn to avoid, the negative consequences of such effects by automatically detecting and flagging relevant situations. The detection methods involve recording constructions in the lexicon that can predictably lead to biased thinking. Table 8.17 shows some examples.

The semantic descriptions of such constructions (recorded in the lexicon) must include the information that the DISCOURSE-FUNCTION of the construction is the listed bias (e.g., SEEK-CONFIRMATION). So, for example, the semantic representation of tag-question constructions like "You don't VP, do you?" will be

Table 8.17
Examples of constructions that can lead to biased thinking

Example	Associated bias
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

REQUEST-INFO		
AGENT	HUMAN	("speaker")
BENEFICIARY	HUMAN	("hearer")
THEME	^main-clause	(the meaning of the main clause)
DISCOURSE-FUNCTION	SEEK-CONFIRMATION	

The values of DISCOURSE-FUNCTION can be incorporated into rules for good clinical negotiation. For example, a doctor is more likely to convince a patient to agree to a lifesaving procedure by framing the side effects, risks, and so on using a positive framing sway rather than a negative one. Similarly, a doctor is more likely to get a patient to provide maximally objective ratings of symptom severity by asking neutral questions ("Do you have any chest pain?") rather than questions framed as SUGGESTIVE-YES/NO or PRIME-WITH-RANGE. The agent can match the most desired utterance types with its assessment of the doctor's goal in the given exchange using the tracking of hypothesized goals and plans (e.g., "convince patient to undergo procedure").

When considering the utility of LEIAs in advising doctors, it is important to remember that the psychological effects we have been discussing are typically not recognized by people in the course of normal interactions. So it is not that we expect LEIAs to discover anything that doctors do not already know or could not learn in principle. Instead, we think that LEIAs could point out aspects of decision-making and interpersonal interactions that, for whatever reason, the doctor is unaware of in the heat of the moment. We think that LEIAs could be particularly useful to doctors who have less experience overall, who have little experience with a particular constellation of findings, who are under the pressures of time and/or fatigue, or who are dealing with difficult nonmedical aspects of a case, such as a noncompliant patient.

8.3 LEIAs in Robotics

It is broadly recognized that progress in social robotics is predicated on improving robots' ability to communicate with people. But there are different levels of communication. While robots have, for example, been able to react to vocal commands for quite some time, this ability does not invoke the kind of fundamental, broad-coverage NLU described throughout this book. Instead, the language utterances understood by robots have been tightly constrained, with most research efforts focused on enabling robots to learn skills through demonstration (e.g., Argall et al., 2009; Zhu & Hu, 2018). The robotics community has not willfully disregarded the promise of language-endowed robots; rather, it has understandably postponed the challenge of NLU, which, in an embodied application, must also incorporate extralinguistic context (what the robot sees, hears, knows about the domain, thinks about its interlocutor's goals, and so on). Integrating the language capabilities of LEIAs into robotic systems is the obvious next step forward.

Typical robots have some inventory of physical actions they can perform, as well as objects they can recognize and manipulate. A LEIA-robot hybrid can acquire a mental model of these actions and objects through dialog with human collaborators. That is, people can help LEIA-robots to understand their world by naming objects and actions; describing actions in terms of their causal organization, prerequisites, and constraints; listing the affordances of objects; and explaining people's expectations of the robots. This kind of understanding will enhance LEIA-robots' ability to understand their own actions and the actions of others and to become more humanlike collaborators overall. Clearly, this kind of learning relies on semantically interpreting language inputs, and it mirrors a major mode of learning in humans—learning through language. In this section we describe our work on integrating a LEIA with a robot in an application system.

The system we describe is a social robot collaborating with a human user to learn complex actions. The experimental domain is the familiar task of furniture assembly that is widely accepted as useful for demonstrating human-robot collaboration on a joint activity. Roncone et al. (2017) report on a Baxter robot supplied with high-level specifications of procedures implementing chair-building tasks, represented in the hierarchical task network (HTN) formalism (Erol et al., 1994). In that system, the robot uses a rudimentary sublanguage to communicate with the human in order to convert these HTN representations into low-level task planners capable of being directly executed by the robot. Since the robot does not have the language understanding capabilities or the ontological knowledge substrate of LEIAs, it cannot learn by being told or reason explicitly about the HTN-represented tasks. As a result, those tasks have the status of uninterpreted skills stored in the robot's procedural memory.

We undertook to develop a LEIA-robot hybrid based on the robot just described. The resulting system was able to

- learn the semantics of the initially uninterpreted basic actions;
- learn the semantics of operations performed by the robot's human collaborator from natural language descriptions of them;
- learn, name, and reason about meaningful groupings and sequences of actions;
- organize those sequences of actions hierarchically; and
- integrate the results of learning with knowledge stored in the LEIA-robot's semantic and episodic memories.

To make clear how all this happens, we must start from the beginning. The LEIA-robot brings to the learning process the functionalities of both the LEIA and the robot. Its robotic side can (a) visually recognize parts of the future chair (e.g., the seat) and the tools to be used (e.g., screwdriver) and (b) perform basic programmed actions, which are issued as non-natural-language commands such as GET(LEFT-BRACKET), HOLD(SCREWDRIVER), RELEASE(LEFT-BRACKET). The hybrid system's LEIA side, for its part, can generate

ontologically grounded meaning representations (MRs) from both user utterances and physical actions.[23] The interactive learning process that combines these capabilities is implemented in three modules.

Learning module 1: *Concept grounding*. The LEIA-robot learns the connection between its basic programmed actions and the meaning representations of utterances that describe them. This is done by the user verbally describing a basic programmed action at the same time as launching it. For example, he or she can say, "You are fetching a screwdriver" while launching the procedure GET(SCREWDRIVER). The LEIA-robot generates the following TMR while physically retrieving the screwdriver.

```
CHANGE-LOCATION-1
    THEME          SCREWDRIVER-1
    AGENT          ROBOT-1
    TIME           find-anchor-time
    EFFECT         BESIDE-1
    textstring     fetch
    lex-sense      get-v1

BESIDE-1
    DOMAIN         SCREWDRIVER-1
    RANGE          HUMAN-1
```

Learning module 2: *Learning legal sequences of known basic actions*. The robot learns legal sequences of known basic actions by hierarchically organizing the TMRs for sequential event descriptions. It recognizes these sequences as new complex actions (ontological events), which it names and records in its ontology. Since the full process of chair assembly is far too long to present here (see Nirenburg & Wood, 2017, for details), we illustrate this process (in table 8.18) by tracing the robot's learning how to assemble the third of the four chair legs.

Learning module 3: *Memory management for newly acquired knowledge*. Newly learned process sequences (e.g., ASSEMBLE-RIGHT-BACK-LEG) and objects (e.g., RIGHT-BACK-LEG) must be incorporated in the LEIA-robot's long-term semantic and episodic memories. For each newly learned concept, the memory management module first determines whether this concept should be (a) added to the LEIA-robot's semantic memory or (b) merged with an existing concept. To make this choice, the agent uses an extension of the concept-matching algorithm reported in English and Nirenburg (2007) and Nirenburg et al. (2007). This algorithm is based on unification, with the added facility for naming concepts and determining their best position in the ontological hierarchy. Details aside, the matching algorithm works down through the ontological graph—starting at the PHYSICAL-OBJECT or PHYSICAL-EVENT node, as applicable—and identifies the closest match that does not violate any recorded constraints. Nirenburg et al. describe the eventualities that this process can encounter.

To recapitulate, the system described here concentrates on robotic learning through language understanding. This learning results in extensions to and modifications of the three

Table 8.18
Learning while assembling the right back leg

The user says,	"We are building the right back leg."
The LEIA-robot carries out a mental action:	It generates a TMR for that utterance.
The user says,	"Get another foot bracket."
The user launches the associated robotic action by inputting	GET(BRACKET-FOOT)
The LEIA-robot carries out a sequence of physical actions:	First, it undertakes the asserted GET(BRACKET-FOOT) action. Then it carries out the action it typically performs next: RELEASE(BRACKET-FOOT).
The LEIA-robot carries out a mental action:	It learns to associate this complex event with the TMR for "Get another foot bracket."
The user says,	"Get the right back bracket."
The user launches the associated robotic action by inputting	GET(BRACKET-BACK-RIGHT)
The LEIA-robot performs the associated physical and learning actions, as before.	
The user says,	"Get and hold another dowel."
The user launches the associated robotic actions by inputting	GET(DOWEL), HOLD(DOWEL)
The LEIA-robot performs the associated physical and learning actions.	
The user says,	"I am mounting the third set of brackets on a dowel."
The LEIA-robot carries out a mental action:	It generates a meaning representation of this utterance.
The user carries out a physical action:	He affixes the foot and the right back brackets to the dowel.
The LEIA-robot learns through demonstration:	It observes this physical action and generates a meaning representation of it.
The user says,	"Finished."
The LEIA-robot carries out a mental action:	It generates a meaning representation of this utterance.
The user says,	"Release the dowel."
The user launches the associated robotic action by inputting	RELEASE(DOWEL)
The LEIA robot performs the associated physical and learning actions.	
The user says,	"Done assembling the right back leg."
The LEIA robot carries out a sequence of mental actions:	(a) It generates a meaning representation for that utterance. (b) It learns the action subsequence for ASSEMBLE-RIGHT-BACK-LEG. (c) It learns the following ontological concepts in their meronymic relationship: RIGHT-BACK-LEG (HAS-OBJECT-AS-PART BRACKET-FOOT, BRACKET-BACK-RIGHT, DOWEL) (d) It learns that RIGHT-BACK-LEG fills the HAS-OBJECT-AS-PART slot of CHAIR.

kinds of memory in a LEIA-robot: explicit semantic memory (i.e., ontology); explicit episodic memory (i.e., a recollection of what happened during the learning session); and the implicit (skill-oriented) procedural memory. We expect these capabilities to allow the robot to (a) perform complex actions without the user having to spell out a complete sequence of basic and complex actions; (b) reason about task allocation between itself and the human user; and (c) test and verify its knowledge through dialog with the user, avoiding the need for the large number of training examples often required when learning is carried out by demonstration only.

The work on integrating linguistically sophisticated cognitive agents with physical robots offers several advantages over machine learning approaches. First, LEIA-robots can explain their decisions and actions in human terms, using natural language. Second, their operation does not depend on the availability of big-data training materials; instead, we model the way people learn, which is largely through natural language interactions. Third, our work overtly models the LEIA-robot's memory components, which include the implicit memory of skills (the robotic component), the explicit memory of concepts (objects, events, and their properties), and the explicit memory of concept instances, including episodes, which are represented in our system as hierarchical transition networks. The link established between the implicit and explicit layers of memory allows the robot to reason about its own actions.

Scheutz et al. (2013) discuss methodological options for integrating robotic and cognitive architectures and propose three "generic high-level interfaces" between them—the perceptual interface, the goal interface, and the action interface. In our work, the basic interaction between the implicit robotic operation and explicit cognitive operation is supported by interactions among the three components of the memory system of the LEIA-robot.

There are several natural extensions to this work. After the robot's physical actions are grounded in ontological concepts, the robot should be able to carry out commands or learn new action sequences by acting directly on the user's utterances, without the need for direct triggering of those physical actions through software function calls. In addition, the incorporation of text generation and dialog management capabilities would allow the robot to take a more active role in the learning process (as by asking questions) as well as enrich the verisimilitude of interactions with humans during joint task performance. Yet another direction of work, quite novel for the robotics field, would be to enable the robot to adapt to particular users, leveraging the sort of mindreading discussed earlier in this chapter.

8.4 The Take-Home Message about Agent Applications

We expect that some readers will have skipped over the details of the applications presented in this chapter. That is fine as long as the point behind all those details is not lost.

The main argument against developing deep NLU systems, particularly using knowledge-based methods, has been that it requires deep and broad, high-quality knowledge, which

is expensive to acquire. Yes, it *is* expensive to acquire, but it is needed not only for language processing but also to enable virtual and robotic agents to function intelligently in many kinds of applications. That is, the knowledge problem is not restricted to matters of language; it is at the core of many of AI's most imposing challenges.

The program of knowledge-based AI presented in this book has not infrequently been dubbed ambitious—a term that tends to carry at least some degree of skepticism. That skepticism is not surprising: if very few people are pursuing one of the most compelling problems science has to offer, there must be some reason why. We hypothesize that a substantial contributing factor is that people simply don't enjoy, and/or don't receive sufficient personal and professional benefits from, doing the kinds of knowledge engineering we illustrate. However, personal preferences and personal cost-benefit analyses should not be confused with more objective assessments of the potential for knowledge engineering to foster progress in the field of AI at large.

We do not expect that this book will turn every reader into an optimistic champion of knowledge-based NLU or, more broadly, knowledge-based AI. However, as we have shown in this chapter and those that precede it, it would be unsound to dismiss our commitment to this paradigm as deriving from unrealistic ideas about how much work it requires. Only time will tell how AI will unfold over the decades to come, but we are making our bets with eyes wide open.