Cognitive Systems: Toward Human-Level Functionality

Sergei Nirenburg

■ *The term cognitive systems may mean different things to different people. This article argues that the core desiderata of an artificial intelligent system are properties that make it more humanlike in its abilities to understand, learn, and explain.*

Whatever area or application domain of AI we work in, when we choose an approach to solving a problem we face, we willy-nilly make a few high-level methodological choices. One such choice is between developing systems that aim to supplant humans — cognitive prostheses — and systems that aim to enhance human performance — cognitive orthotics. The distinction between the two is clear on the example of machine translation (MT).

Although prosthetic systems aim to supplant humans by independently matching human performance on a task, most prosthetic systems still have to rely on people to yield a high-quality final result. Thus, results of Google Translate must be edited by a person to yield a high-quality translation. The practice of postediting the results of machine translation has been employed for over half a century. It is clear that today's fully automatic MT systems yield much better raw translations than systems of yore, thus making the job of a

posteditor easier and making the use of such systems commercially more attractive. The reason automatic MT systems are considered prosthetic, even though humans are needed to achieve high-quality results, is that such systems operate fully independently before humans — in a separate, optional, process — posted-it the results.

Orthotic systems, by contrast, are intended to collaborate with humans on carrying out tasks, serving as high-functioning members of a society populated by a mixture of humans and artificial intelligent agents. As the intelligent agents in this society, orthotic systems must both perform tasks and communicate at a human level. When applied to translation, an orthotic system will help a human translator by providing the best translation it can confidently derive automatically and, instead of stopping there, it will guide the translator through the difficult spots (for example, convoluted syntactic constructions or residual lexical ambiguities) presenting options for human review. When such a system is not sufficiently confident to produce even an initial translation independently, it will engage the human translator interactively as an oracle to rule on those difficult spots. Such a system can also, on its own initiative, decipher opaque references; for example those to Evelyn Waugh's *Brideshead Revisited* in "Once a project is up and running, I hang around my subject like Charles Ryder hung around Sebastian Flyte ...." (*Times Literary Supplement,* June 27, 2017). Experienced translators will agree that such information is of great help in coming up with the right translation of the original. A good orthotic system will also learn from experience, with respect to both its own decision making during initial translation and the nature and intensity of communication desired by a particular translator. (The latter capability is, of course, not essential for a prosthetic system since a human participates in the process only after the system has completed its work.)

As just illustrated, a key feature of AI systems — be they prosthetic or orthotic — is the ability to automatically estimate self-confidence in the results of processing and then proceed accordingly. In orthotic systems, this type of metacognition will foster joint-initiative interactive problem solving, making it faster to achieve high-quality results while continually lowering the human's cognitive load. Prosthetic systems can also include a module for gauging confidence. Indeed, this capability was arguably the most spectacular achievement of IBM's *Jeopardy!*-playing Watson system. The question is whether in all situations, when no proposed solution is above a confidence threshold, there is a realistic way of avoiding local failure. After all, not in every application there will be an option, as in *Jeopardy!*, to forgo offering a solution.

As orthotic systems improve, they may reach a stage when their confidence in their own results is above the desired quality threshold at all times. For that matter, prosthetic systems may improve to such a degree that in practice no human postprocessing will be necessary. Is either of these scenarios realistic? Is one of them more realistic than the other? First of all, let us not forget that Google Translate is fully operational while an orthotic MT system (or any AI system) with the capabilities sketched above does not exist today. Why? Because to build an orthotic system one must first meet a large number of hard prerequisites. Just for starters, in order to seamlessly communicate with people on a topic of mutual interest, such a system must understand people: it must understand and speak their language, understand their needs and motives, be aware of at least some of their beliefs, remember past experiences and be able to use them in decision making. Of course, many more capabilities will be needed to address additional modes of perception and reasoning and to absorb and integrate input from a combination of several kinds of perception.

Irrespective of whether AI applications are orthotic or prosthetic, mechanisms of life-long learning will be required for extending and improving the inventories of parameters to be used in a variety of decision functions and algorithms they will use. In orthotic systems, it is preferable that such parameters make sense to people because the symbiotic nature of such systems' functioning requires that people understand not only the results of the artificial intelligent agent's behavior but also how and on the basis of what beliefs these results were obtained. By contrast, current prosthetic AI applications derive parameter inventories using one of the many forms of machine learning (ML), and explaining parameter choice in a way that people can understand is not among the first-order tasks of AI researchers working in this paradigm. Both orthotic and prosthetic systems will also equally benefit from a quest for ever better methods for deriving time-sensitive values for the selected inventory of parameters.

As the quality of AI technologies asymptotically grows, so will the quality of all resulting systems, no matter their genre. One can envisage that, ultimately, the quality of component technologies will be so high that the choice to configure an AI system as orthotic or prosthetic will be based exclusively on the needs of the application, rather than the degree to which the applications will tolerate less-than-optimum results.

Recent spectacular advances in AI have been made possible by enormous increases in the speed and storage capacity of computers, as well as by the ubiquity of big data — most notably, the Internet. Lightning speeds enabled the broad application of machine-learning algorithms to practical tasks on ever larger data sets. Indeed, most current AI systems exploit regularities and analogies detected in large data sets without relying on preconceived models of the world, lan-

guage, or agency. Such systems tilt toward being prosthetic by necessity. They are the best that the field can offer in the short term, as they outperform handcrafted AI systems while requiring much less human labor to create. The immediate utility of such systems has blunted the impetus to develop new methods of overcoming the difficulties of achieving human-level AI, notably, approaches to overcoming the notorious *knowledge acquisition bottleneck:* the difficulty and cost of acquiring, representing, managing, and using the massive amounts of knowledge that humans bring to bear in their functioning. A stored knowledge substrate that models that available to people is generally considered necessary to configure comprehensive, multifunctional, human-level artificial intelligent agents. That the recent successful ML-based AI applications prefer, whenever possible, to bypass a reliance on stored knowledge (and thus not get bogged down in the difficulties of knowledge acquisition) offers a reason for why these applications — such as playing *Jeopardy!* — are by and large narrowly specialized. By comparison, the work toward building more humanlike capabilities and more comprehensive systems, which can be applied to a large scope of application areas, has not attracted a sufficient amount of recent attention among AI researchers, funders, or the general public alike.

It is instructive in this connection to review the responses from a large group of scientists and public intellectuals to the Edge.org 2015 annual question: "What Do You Think About Machines That Think?" (Brockman 2015). Some opinions from this survey will be quoted later, but we begin with one from Gary Marcus: "We still have no machine that can, say, read all that the Web has to say about war and plot a decent campaign, nor do we even have an open-ended AI system that can figure out how to write an essay to pass a freshman composition class, or an eighth-grade science exam. Why so little progress, despite the spectacular increases in memory and CPU power?"

The reason is that, as mentioned earlier, ML-oriented systems have a narrow sphere of applicability. In order to burst through the quality ceiling and move toward comprehensive applications that are more like intelligent agents than mechanistic automata, the field must re-address, with newly available theories and methods, the development of systems featuring human-inspired computational models. Such models are essential for orthotic applications but, once developed, will offer an attractive option to prosthetic ones as well — after all, according to a familiar consensus, the old knowledge-based AI paradigm was abandoned not because it was proven wrong but because it was shown to be too expensive. The real significance of the big data and deep learning revolution of recent years may very well be not in demonstrating that it can yield a Go-playing system that beats the best human players. Its main promise may be in its potential to help cut the costs of overcoming the AI knowledge bottleneck by facilitating ever more sophisticated orthotic systems for knowledge acquisition. An example of what I have in mind would be an orthotic system that is a member of a human-computer team charged with acquiring knowledge about the world and/or about language using a combination of ML methods and smart environments for eliciting knowledge from people. An example of such an orthotic system is the Boas linguistic knowledge elicitation system (McShane 2003; Oflazer, Nirenburg, and McShane 2001), which supported the recording of machine-tractable knowledge about the grammar and lexicon of low-resource languages in order to speed up the configuration of MT systems.

Knowledge-lean AI systems have a built-in ceiling of quality. To mention just one example, true human-level understanding of language (text and dialogue) can only be achieved by a combination of understanding what is overtly mentioned and inferring what is implied. Interpreting implied meanings requires extensive knowledge and context-based reasoning, as detailed in the discussion of understanding that follows. So, whereas knowledge-lean language processing systems might be able to generate rough translations of texts without attempting to derive implied meanings. But systems that aspire to achieve full understanding in even the more challenging communication genres, such as dialogue, must be able to recover many elements of hidden meaning. It is time for scientific (as opposed to technological) AI to take a long view and increase the efforts on tackling the hard residue that will remain after the rate of improvement in data-driven AI levels out (as is noted already by some forward-looking practitioners — see the discussion of Ken Church's opinions in McShane's contribution in this issue [McShane 2017]).

The cognitive systems community takes a long view and concentrates on modeling the cognitive abilities of people in comprehensive artificial intelligent agent systems operating in human-agent teams. As a rule, work on these systems prioritizes deep understanding of phenomena over the immediate utility of narrowly specified applications.

Cognitive Systems

The modifier *cognitive* has become very fashionable in the recent past. It has been used to qualify a wide variety of concepts defining research communities, not only the more traditional *psychology* and *science* but also *economics, computing, neuroscience, linguistics,* and others. So, *cognitive systems* was added to an already long list of fields whose names include this popular modifier. What is behind this label? Here is what Daniel Dennett says in response to the above-mentioned Edge.org question:

In the earliest days of AI, an attempt was made to

> enforce a sharp distinction between artificial intelligence and cognitive simulation. The former was to be a branch of engineering, getting the job done by hook or by crook, with no attempt to mimic human thought processes—except when that proved to be an effective way of proceeding. Cognitive simulation, in contrast, was to be psychology and neuroscience conducted by computer modeling … [Still], to lay people AI means passing the Turing Test, being humanoid. The recent breakthroughs in AI have been largely the result of turning away from (what we thought we understood about) human thought processes and using the awesome data-mining powers of super-computers to grind out valuable connections and patterns without trying to make them understand what they are doing.

Substitute *systems* for *simulation* and we are coming a step closer to delineating the purview of our community. The AI historian Pamela McCorduck adds in her response to the Edge.org survey:

> The original motive for AI at Carnegie Mellon in the mid-1950s was to model some part of human cognition, in this case, logical reasoning—a "small but fairly important subset of what's going on in mind," Herb Simon put it to me. Marvin Minsky and John McCarthy were also fascinated by the human mind … Minsky later told me that whatever detours he later took, he returned finally to focusing on the human mind.

The perceived need to define the concept of cognitive systems brings to mind Minsky's "suitcase words," which have a vague meaning but refer to important though not universally clearly defined (or definable) concepts, such as consciousness, emotion, perception, awareness, intelligence, attention, happiness or, for that matter, knowledge and heuristics in the manner they are used in this article. (Without going into too much detail, when I use the term *knowledge,* I do not mean to imply that it is necessarily true or correct.) Minsky (2006) points out that it might be counterproductive to try to make the corresponding notions more precise right from the outset: "We need to use those suitcase words in our everyday lives to keep from being distracted by thinking about how our thinking works" (p. 128). I think this is true even for people whose everyday lives involve working in AI. So, let us treat *cognitive systems* as a suitcase phrase. In lieu of a definition, let us at this time stress the field's emphasis on modeling human abilities, such as understanding, learning, and explaining — each of which we will consider in turn.

Understanding in cognitive systems involves interpreting, within the agent's world model, the meaning of sensory inputs from the outside world. Crucially, and uniquely for humans, these include the overt and implied meanings of language utterances. As an illustration, consider how an agent capable of human-level understanding — which, at this time, has to actually be a human but, over time, could be a machine — will behave in the following situation. If in a storm the agent observes a tree teetering and hears a characteristic cracking sound — or, alternatively, receives a verbal report about this happening — it should be able to infer that the tree may be about to fall, so that anything that should be protected from the resulting impact should be, if at all possible, removed from the expected crash path. In fact, hearing the utterance "That birch tree in the front yard is swaying suspiciously" should be understood as an indirect speech act requesting this kind of action. If the agent cannot estimate the potential trajectory of the falling tree, it should apply the clearing-out goal to a circle around the tree with a radius roughly corresponding to the tree's height. If the agent can't physically remove everything under potential threat, it should choose the most important things to be rescued or alternatively recruit help for this task (this typically involves generating and interpreting a sequence of speech acts in a dialogue). Or, if humans or animals are threatened, it could choose to remove them by signaling to them using language or some other means of inducing them to activate a context-sensitive instance of the goal of self-preservation).

Understanding involves not only taking stock of the current situation in the world of the agent (for example, a chess position in which it must make a move or a single question just asked of it). It also involves understanding the goals and plans of all agents active in the environment. An agent's reasoning process starts with understanding the direct meaning of sensory inputs — be they utterances or other stimuli from other senses, it continues with understanding the intended meaning of these inputs along with the intentions of relevant other agents, and then it smoothly extends into reasoning about the following:

1. How to integrate this newly understood meaning into the its short-term and long-term memories
2. How this meaning impacts its owns goals and plans
3. How this meaning impacts its emotional state
4. Whether this input requires it to carry out an action, and
5. If it does, then what action should be selected

These five tasks are relevant not only after perceptual input is fully processed and interpreted but also during the interpretation process itself. That is, expectations stored in the agent's long-term memory (its world model, theory of the minds of other agents and self, knowledge of language) help in the interpretation process and also support general reasoning and decision making. That's because more often than not context-free processing of perceptual input does not yield complete or satisfactory results. This applies, by the way, not only to language but also to vision, as described in Summerfield and Egner (2009). To summarize: the paucity of the input signal is ubiquitous and, in the case of human dialogue, typically deliberate (Piantadosi, Tily, and Gibson 2012), which is explained by the application of the principle of least effort (as applied to the management of the agent's cognitive load). Therefore, understanding is a func-

tion of both the input and the reasoning on the basis of the knowledge (beliefs, norms, conjectures, narratives, desires, biases, emotions, and so on) stored in the agent's memory.

The ability to understand is tightly coupled with the ability to *learn*. As emphasized earlier, in order to understand, people must possess a lot of knowledge, which must be learned. In artificial intelligent agents, learning can be delegated to human knowledge acquirers (of the type well-known in expert systems of the past), or it can be modeled as a dynamic capability of agents. Cognitive systems pursue the computational modeling of human abilities with a methodological preference for emulating not just the human-quality outcomes but — within the limitations of modeling tools — the processes used by humans in arriving at those outcomes. Although modeling early childhood learning is a fascinating topic, it is probably not the most realistic objective at this time. However, it is plausible to simulate the kind of language-centered learning that most people are actively involved in starting at around age 6. In her response to the Edge.org question, Alison Gopnik puts it this way:

> Starting at a rather tender age and throughout their lifetimes [...] people learn predominantly by being taught using natural language — at the very least, to accompany visual or other sensory input. Language abilities are thus a prerequisite to human-like learning. But these language abilities are integrated with reasoning capabilities and beliefs about how the world is organized. It is important to stress that the world includes people and AI agents. This means that in order to function at the human level AI agents must have a theory of mind (being able to ascribe beliefs and attitudes to other agents) and self-awareness (doing the same with respect to oneself). Note that these beliefs and attitudes could be wrong. But they must be there — otherwise the agent will not be able to provide explanations of events and states in the world and thus will fall short of being considered a model of a human.

We can model language-mediated learning — by reading, by verbal instruction, through dialogue, and so on — once we endow the agent with a basic ability to *understand*. That ability, as I argued earlier, is predicated on the availability of several kinds of stored knowledge. When at the start of the learning process we endow the agent with the necessary minimum inventory of concepts — as well as language-processing, reasoning, and decision-making algorithms — it will be able to apply these resources to expand its knowledge of grammar, lexicon, world model elements, and other components of its stored knowledge by understanding the meaning of explanations that its teachers provide in natural language. The ability for deep understanding of the language and the world (including discourse situations) is an immutable prerequisite for modeling human-style learning. Early attempts to overcome the knowledge acquisition bottleneck by using the information in machine-readable dictionaries (MRDs) did not succeed precisely because those dictionaries were intended for people: the definitions used ambiguous terminology and were often circular, and the information provided was variable in depth and coverage. Interpretation of the meaning of dictionary entries was not a part of these MRD-harvesting efforts. Predictably, these efforts proved useful only for systems that could succeed by manipulating uninterpreted text strings.

The ability to learn through language is, in turn, tightly coupled with the ability to explain (DeJong 2004). This ability is not a forte of the current generation of AI systems. From Rodney Brooks's Edge.org contribution:

> Today's chess programs have no way of saying why a particular move is "better" than another move, save that it moves the game to a part of a tree where the opponent has less good options. A human player can make generalizations and describe why certain types of moves are good, and use that to teach a human player. Brute force programs cannot teach a human player, except by being a sparring partner. It is up to the human to make the inferences, the analogies, and to do any learning on their own. The chess program doesn't know that it is outsmarting the person, doesn't know that it is a teaching aid, doesn't know that it is playing something called chess nor even what "playing" is. Making brute force chess playing perform better than any human gets us no closer to competence in chess.

The explanatory capabilities of social agents, including cognitive systems, must include explaining to other agents what they are doing (or intend to do, or have done); explaining to themselves what other agents are doing; and understanding other agents' explanations. So, explanation can be viewed as the other side of learning in social settings. To be effective, it must be geared toward a specific audience: deciding what to say to explain something depends on the explainer's model of the explainee's theory of mind, the explainer's memory of past interactions, and the explainee's perceived goals or needs.

The need for explanation has been widely recognized in the AI, as evidenced, for example, by the recently started DARPA Explainable AI program. A workshop on the topic was held at IJCAI-2017. This is a positive development. Constructing explanations is not an easy task. Constructing relevant explanations is an even more difficult one. It seems that very few things can demonstrate that an artificial intelligent agent possesses at least a vestige of human-level intelligence as well as its ability to generate explanations specifically for a particular audience and state of affairs in the world. Without these constraints, many explanations, while being technically accurate, might prove unedifying or inappropriate. Plato's reported definition of humans as "featherless bipeds" may have engendered Diogenes' witty repartee but

will not be treated by most people in most situations as an enlightening characterization.

In my view, cognitive systems are *explanatory* systems committed to computational and representational theories of mind. They are theory based, geared toward responding to *why* questions, not only the much simpler *how* and *what* questions. The nature of the explanation can be causal or empirical. Causal explanations can appeal either to laws of physics or biology or to folk psychology: "because people tend to like people whom they helped." Empirical explanations can range from "have always done it this way and succeeded" to appeals to authority, as in "this is what my teacher told me to do."

Philosophers and psychologists (notably, Lombrozo 2013) have devoted significant attention to the varieties and theories of explanation, often coming to unexpected conclusions, as when Nancy Cartwright (1983) persuasively argues that, despite their great explanatory power, fundamental scientific laws do not describe reality. Cognitive systems are artificial intelligent agents that are not primarily intended — at least not at this time — to model scientists. This licenses an emphasis on implementing models of explanation that are not necessarily scientific, nor necessarily (always) true, but are always contextually appropriate and take into account the goals, plans, biases, and beliefs of both the explainer and the explainee, not only immediate sensory inputs. Such models will also rely heavily on folk psychology as the basis for explanation. Systems must follow folk psychology because they collaborate with humans, and humans need explanations in terms they understand and are used to. When doing science, people do not usually behave in accordance with folk psychology but in everyday life those same people certainly do.

The Current Consensus and a Few Personal Desiderata ("Aspirational Issues")

This section presents a summary catalogue of issues that cognitive system developers currently address and methodological preferences that they, by and large, share. For some of the issues, the consensus is not entirely universal, which is to be expected for a group of active developers. Still, the general points of consensus should help to characterize the overall trends in the community. In addition to an assessment of current developments, I add several personal opinions about directions that the field needs to address if we aspire to make a qualitative leap in modeling human-level functionality using computers. The descriptions that follow are necessarily very brief and high level. The other cognitive systems contributions in this issue present a much more detailed view of many of the points that follow. Each of these points can clearly be discussed and debated in great detail. If this article facilitates such debates, it will have served its primary purpose.

The cognitive systems paradigm develops artificial intelligent agents that model higher-level human abilities and are implemented as programs on digital computers. Approaches to implementation can be widely diverse but the processes and phenomena that are modeled as components of a cognitive system are common for the field: decision making, preferences, models of the world including models of other agents, perception and action capabilities, and so on.

At the moment, the modeling of perception-related processes in cognitive systems usually concentrates on the tasks that take place after physical input signals are transformed in a symbolic representation appropriate for a particular kind of perception — for example, for language input this will be an encoding of text (possibly, obtained as a result of transcribing speech). Cognitive systems then translate symbolic representations of inputs into expressions in the language they use to record the contents of their memory and carry out reasoning and decision-making. It would be an understatement to say that this is a difficult task: it is a rule, not an exception that perceptual inputs are incomplete or ambiguous, or both; and that the system's internal resources will never cover all the possible inputs. For cognitive systems researchers the important point is that people habitually face precisely the same difficulties, and are known to cope with them adequately. Modeling this human ability is an important facet of work in our field (McShane's [2017] contribution describes such a model for language understanding).

Once the input interpretation task is completed, the system may turn to deciding what actions, if any, it should plan and execute as a result of this input. Cognitive systems may be equipped with a variety of components that can influence this decision. In addition to a model of the world and a reasoning mechanism for computing expected utility, these may include an inventory of goals and plans, a set of personality traits, a model of emotional states, a repository of personal and societal norms, preferences, or biases. This is an area where statistics- and ML-based systems can be symbiotic with cognitive systems: the former can provide advanced computation frameworks while the latter can provide content-related insights into the choice of the inventory of features to be used in making decisions. *Aspirational issue:* Most current models of decision making are normative. This is understandable — artificial intelligent agents must be endowed with normative theories, for two reasons: (1) to have a yardstick against which to judge individual behavior instances and their deviations from what is societally expected; and (2) to be able to teach others about societal norms and desiderata — although fully expecting deviations to keep occurring in the future. This last point leads to the desideratum of paying attention to developing descriptive deci-

sion theories that take into account ever-present human inconsistency. In a variety of application domains modeling human frailties is as important as modeling human achievements. In his Edge.com survey response, Dennett says: "A cognitive simulation model that nicely exhibited recognizably human errors or confusions would be a triumph, not a failure." I fully agree.

Actions in cognitive systems can be physical (several such systems are implemented using robots); verbal (both dialogue and narrative oriented); and mental (for example, interpreting intended, rather than direct meaning of language utterances, such as indirect speech acts ("Do you have a watch?") or deciding what to do next). *Aspirational issue:* people often ideate without this process being overtly triggered by a sensory input. The repertoire of cognitive systems will be enhanced if, when they have an opportunity, they could learn by introspection and reasoning, whereby an agent, in the absence of a more pressing goal, will, for example, use its reasoning capacity to seek and resolve contradictions that are guaranteed to be present in the world model of a learning agent or make overt the conclusions that had not been overtly recorded but whose premises are available in its world model. This seems to suggest a path toward a theory that might explain at least some facets of creativity.

An important component of any model of processing perceptual input and deciding what to do next is modeling attention (see Bello and Bridewell's [2017] contribution in this issue).

Study the sociological aspects of systems. Cognitive systems are largely intended to be deployed as members of human-agent teams. They are meant to enhance the quality of such current applications as companion agents, physician agents (aiming to combine the information abilities of an IBM Watson with features of a truly cognitive system), members of military teams, such as those under development in the Air Force Faithful Winger program, and many others. As a result, cognitive system developers study the social psychology of human-agent teams, with special attention to issues of trust and autonomy. Of special interest are the ethical parameters in the functioning of human-agent societies (see Scheutz's [2017] contribution in this issue). This line of research is in its infancy and offers a lot of promise for making agent models more humanlike — after all, an influential trend in the studies of consciousness (for example, Baumeister [1986]) stresses the pivotal role of society in forming an individual.

As already mentioned, artificial intelligent agents developed by cognitive systems researchers must be capable of human-style learning (not machine learning over big data). This kind of learning is not expected (at least in the foreseeable future) to start with an empty slate but rather will be facilitated by "seeding" the system with a modicum of knowledge and abilities and endowing the system with the ability to learn in a humanlike way as an apprentice in a social setting. While some work in this direction has been ongoing, this is still largely an aspirational issue. Language understanding ability is a key prerequisite.

Cognitive systems researchers typically target steady progress toward modeling as large a subset of human functionalities as possible rather than the pursuit of immediate (but limited) utility in narrowly defined applications. As a result of this preference, a lot of effort is being expended in the cognitive systems community on building support tools that facilitate implementations of models and theories of phenomena and processes. Most prominent among these tools are computational environments known as cognitive architectures (see Laird, Lebiere, and Rosenbloom's [2017] contribution in this issue). *Aspirational issue:* In my estimation, the work on knowledge representation schemata and reasoning and decision-making algorithms within cognitive architectures has historically occupied too central a place in work on cognitive systems. It is increasingly recognized that, while this work is still essential, the center of gravity at this point in the development of the field may very well need to shift toward developing theories of content (what should be encoded) and away from format (how to encode it).

Human functioning involves a wide variety of tasks. It is prima facie improbable that a single implementation method will serve all of them equally well. Indeed, any putative success in fitting the diverse needs of the components of a cognitive system into the Procrustean bed of one preferred algorithm and data structure may take too much time away from the primary task of modeling human cognitive functionality. A better strategy may be to select computational implementations for components of the overall system (for example, memory management, goal choice, or discourse analysis) and then to devise a blackboardlike noticeboard that will allow each of the system modules to use results from any other module as grist for the decision-making mill (however it is implemented in a specific system). Under this approach, many diverse methods can coexist within the same system architecture — symbolic and connectionist processing; abductive and analogy-based reasoning; ontological and distributional semantics, and so on. All considerations of efficiency of computation can be deferred till after an initial system of this kind is demonstrated.

A central component of an explanatory model of human cognitive functioning is the theory of mind of other agents. Stanislas Dehaene writes in his response to the Edge.org survey:

> Unless we suffer from a disease called autism, all of us constantly pay attention to others and adapt our behavior to their state of knowledge — or rather to what we think that they know … Future software should incorporate a model of its user … Unlike present-day computers, humans do not say utterly irrele-

vant things, because they pay attention to how their interlocutors will be affected by what they say. The navigator software that tells you "at the next roundabout, take the second exit" sounds stupid because it doesn't know that "go straight" would be a much more compact and relevant message.

Aspirational issue: In reality, a particular user may prefer the *robotic* message from the navigator. So, to function at the human level, a cognitive system will have to learn the theories of mind of each of the members of its team and act accordingly. The theory of mind must be extended to cover the mind of the cognitive system itself, that is, to help model the intelligent agent's self-knowledge; an interesting modeling angle here is that the agent's self-image may (and, as a rule, does) differ from how the agent is viewed by other agents or from the "objective" state of the world, as seen by an omniscient "demiurge" (that is, in reality, the system developer).

The ultimate criteria of success in building cognitive systems are (1) whether the resulting system behaves (understands, makes decisions) like a human and (2) whether its behavior can be explained in terms that make sense to humans that interact with them. Reliable formal evaluation procedures for establishing this are currently expensive, as they require experimentation with human subjects or settings such as Loebner competitions. Aspirational issue: Developing better measures of progress that are both specifically geared to cognitive systems and accepted by the broad AI community is an urgently required direction of research.

Conclusion

The purpose of this article is to present a bird's-eye view of a research community. It is clear that there will be omissions and lack of detail. The rest of the cognitive systems contributions in this issue will help to fill at least some of these lacunae.

This article is not a call to stop working on data-driven AI. In fact, there is a two-way symbiotic relationship between data-driven and knowledge-based AI. Thus, corpus annotation by people is typically a prerequisite for developing statistical NLP systems. Conversely, sophisticated analyses of large data sets offer immense help to knowledge acquirers — both human and, in the near future, automatic ones. Investigating the potential of using both approaches simultaneously in building AI systems is one of the most promising ways of overcoming the knowledge acquisition bottleneck of cognitive systems and the narrow applicability and quality bottleneck of ML-based ones. Building orthotic systems would be the first choice. But improvements may very well be as tangible in prosthetic ones.

Acknowledgments

This research was supported in part by Grant #N00014-16-1-2118 from the US Office of Naval Research. Any opinions or findings expressed in this material are those of the author and do not necessarily reflect the views of the Office of Naval Research. Many thanks to Paul Bello and Marge McShane for useful comments on an earlier draft. All remaining misunderstandings and obscurities are mine.

References

Baumeister, R. F. 1986. *Identity: Cultural Change and the Struggle for Self*. Oxford, UK: Oxford University Press.

Bello, P. F., Bridewell, W. 2017. There Is No Agency Without Attention. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2742

Brockman, J. 2015. *What to Think About Machines That Think: Today's Leading Thinkers on the Age of Machine Intelligence*. New York: Harper Perennial.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford, UK: Oxford University Press. doi.org/10.1093/0198247044.001.0001

Church, K. 2011. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology* 6(5): 1–27.

DeJong, G. 2004. Explanation-Based Learning. In *Computer Science Handbook,* ed. A. Tucker, 68–1–68–20. Boca Raton, FL: CRC Press.

Laird, J. E.; Lebiere, C.; Rosenbloom, P. S. A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2744

Lombrozo, T. 2013. Explanation and Abductive Inference. In *The Oxford Handbook of Thinking and Reasoning,* ed. K. J. Holyoak and R. G. Morrison. Oxford, UK: Oxford University Press.

McShane, M. 2017. Natural Language Understanding (NLU, not NLP) in Cognitive Systems. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2745

McShane, M. 2003. Applying Tools and Techniques of Natural Language Processing to the Creation of Resources for Less Commonly Taught Languages. *IALLT Journal of Language Learning Technologies* 35(1): 25–46.

Minsky, M. 2006. *The Emotion Machine*. New York: Pantheon, and Simon and Schuster.

Oflazer, K.; Nirenburg, S.; and McShane, M. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics* 27(1): 59–85. doi.org/10.1162/089120101300346804

Piantadosi, S. T.; Tily, H.; Gibson, E. 2012. The Communicative Function of Ambiguity in Language. *Cognition* 122(3): 280–291. doi.org/10.1016/j.cognition.2011.10.004

Scheutz, M. 2017. The Case for Explicit Ethical Agents. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2746

Summerfield, C., and Egner, T. 2009. Expectation (and Attention) in Visual Cognition. *Trends in Cognitive Sciences* 13(9): 403–409. doi.org/10.1016/j.tics.2009.06.003

Sergei Nirenburg is a professor in the Departments of Computer Science and Cognitive Science at Rensselaer Polytechnic Institute.