The Social-Emotional Turing Challenge

William Jarrold, Peter Z. Yeh

■ *Social-emotional intelligence is an essential part of being a competent human and is thus required for human-level AI. When considering alternatives to the Turing test it is therefore a capacity that is important to test. We characterize this capacity as affective theory of mind and describe some unique challenges associated with its interpretive or generative nature. Mindful of these challenges we describe a five-step method along with preliminary investigations into its application. We also describe certain characteristics of the approach such as its incremental nature, and countermeasures that make it difficult to game or cheat.*

The ability to make reasonably good predictions about the emotions of others is an essential part of being a socially functioning human. Without it we would not know what actions will most likely make others around us happy versus mad or sad. Our abilities to please friends, placate enemies, inspire our children, and secure cooperation from our colleagues would suffer.

For these reasons a truly intelligent human-level AI will need the ability to reason about other agents' emotions in addition to intellectual capabilities embodied in other tasks such as the Winograd schema challenge, textbook reading and question answering (Gunning et al. 2010, Clark 2015), image understanding, or task planning. Thinking at the

human level also requires the ability to have reasonable hunches about other agents' emotions.

Social-Emotional Intelligence as Affective Theory of Mind

The ability to predict and understand another agent's emotional reactions is subsumed by a cognitive capacity that goes by various names including folk psychology, naïve psychology, mindreading, empathy, and theory of mind. We prefer the latter term, considering it is more precise and is more frequently used by psychologists nowadays. Theory of mind encompasses the capacity to attribute and explain the mental states of others such as beliefs, desires, intentions, and emotions. In this article, we focus on affective theory of mind because it restricts itself to emotions. We further restrict ourselves to consensual affective theory of mind (AToM) to rule out idiosyncratic beliefs of particular individuals.

Is There a Logic to Emotion?

Each of us humans has our own oftentimes unique affective reaction to a given situation. Although we live in the same world, our emotional interpretations of it are multitudinous. Does this mean that emotion is an "anything goes" free-for-all? In spite of the extreme variability in our affective evaluations, there nonetheless seems to be a rationality, a logic, of what constitutes a viable, believable, or sensible emotional response to a given situation.

When we hear of someone's emotional reaction to a situation sometimes, we think to ourselves, "I would have responded the same way." For other reactions, we might say, "That would not be my reaction, but I can certainly understand why he or she would feel that way." At still other times, another's actual emotional reaction may vary far afield of our prediction and we say, "I cannot make any sense out of his or her reaction."

For these reasons there does appear to be some sort of "logic" to emotion. Yet, how do we resolve the tension between the extreme possible richness and variability in emotional response and the sense that only certain reactions are sensible, legitimate, or understandable?

In the next two sections, we show how the concepts of falsifiability — the possibility of proving an axiom or prediction incorrect (for example, all swans are white is disproven by finding a black swan [Popper 2005]) — and generativity — the capacity of a system to be highly productive and original — play an important role in the resolution of this tension. Later, in the Proposed Framework section, we shall see how these two concepts influence the methods we propose for assessing machine social-emotional intelligence.

Falsifiability and AToM

In our approach to assessing affective theory of mind, we take the term *theory* seriously. Prominent philosophers of science claim that scientific theories are, by definition, falsifiable (Popper 2005). Although an optimistic agent may view a situation with a glass half full bias and pessimistic agents may tend to view the very same situations with a glass half empty bias, they can still both be correct. How then do we demonstrate the falsifiability of affective theory of mind? The answer comes when one considers a predicted emotion paired with the explanation of this prediction. If we consider both together then we have a theory that is falsifiable.

Consider the following situation and the following predictions:

Situation: Sue and Mary notice it is raining.
Appraisal U1: Sue feels happy because she expects the sun will come out tomorrow.
Appraisal U2: Mary feels sad because she hates rain and it will probably keep on raining.

Although some of us may tend to agree more with one or the other's reaction, virtually all of us will judge both of these replies as potentially valid (modulo some relatively minor assumptions about normal personality differences). By contrast, consider what happens if we invert the emotions felt by each character:

Appraisal R1: Mary feels sad because she expects the sun will come out tomorrow.
Appraisal R2: Sue feels happy because she hates rain and it will probably keep on raining.

We take it as a given that the vast majority of typical humans representative of a given cultural group will judge the immediately above appraisals as invalid or extremely puzzling.

In sum, emotion is not an anything goes phenomenon — we have demonstrated that some appraisals violate our intuitions about what makes sense. Although there are a multitude of different emotions that could make sense, falsifiability is demonstrable when one considers the predicted emotion label along with its explanation (Jarrold 2004). As will be described next, falsifiability of AToM is important in the context of Turing test alternatives.

A Generative AToM

Leaving falsifiability aside, there remains the need to provide an account for the multitude of potential emotional appraisals of a situation. The need is addressed by viewing appraisal not as an inference but rather as a generative process.

Generative processes are highly productive, able to produce novel patterns of outputs such as cellular automata, generative grammars, and fractals such as the Mandelbroit or Julia Set. Ortony (2001) posited that generative capacity is critical to computational accounts of emotion.

As a demonstration of this generativity, consider the range of appraisals obtained from "college sophomore" participants in Jarrold (2004) (see table 1).

Scenario	Tracy wants a banana. Mommy gives Tracy an apple.
Question	How will Tracy feel? (Choose from happy, sad, or indifferent)
	Appraisals
Valence	Explanation
Happy	She'll feel happy even though she didn't get exactly what she wanted; it is still something.
Indifferent	Because nonetheless she still has something to eat just not exactly what she wanted.
Indifferent	She will feel indifferent as long as she likes apples too. It isn't exactly what she wanted, but she was probably just hungry and if she likes apples then she would be satisfied because it would do the same thing as a banana.
Sad	Because she was probably excited about eating the banana that day and when mom gave her an apple instead she probably felt disappointed and wondered why her mom wouldn't give her what she wanted.
Sad	She did not get what she wanted.

Table 1. Five Human Appraisals of a Simple Scenario

Alhough research subjects were presented with a very simple scenario, answers ranged from happy to indifferent to sad. The explanations for a given emotion also varied in terms of assumptions, focus, and complexity.

Note that the inferences in explanations are often not deductions derived strictly from scenario premises. They can contain abductions or assumptions (for example, in table 1, row 3 "she is probably just hungry") and a series of subappraisals (for example, row 4 excitement yielding to disappointment).

Furthermore, note that the above data were generated in response to very simple scenarios derived from an autism therapy workbook (Howlin, Baron-Cohen, and Hadwin 1999). Imagine the generative diversity attainable in real-world appraisals where the scenarios can include *N* preceding chapters in a novel or a person's life history.

Typical humans predict and explain another's emotions and find it easy to generate, understand, and evaluate the full range of appraisal phenomena described above. For this reason it is important that human-level AI models of emotion be able to emulate this generative capacity.

Outline

In the remainder of this article, we will first describe how test items are involved in a five-stage framework or methodology for conducting an evaluation of computational social-emotional intelligence. Challenges to the integrity of the test are anticipated and countermeasures are described. Finally, issues with the specifics of implementing this framework are addressed.

Proposed Framework

Each of the framework's five stages (see figure 1) is described: first, developing the test items; second, obtaining ground truth; third, computational modeling; and, finally, two stages of evaluation. In these last two evaluation stages models are judged on the basis of two corresponding tasks: (1) generating appraisals (stage 4) and (2) their ability to evaluate other's appraisals — some of which have been manipulated (stage 5)

Test Items

The framework revolves around the ability of a system to predict the emotions of agents in particular situations in a human-like way across a sufficiently large number of test items. As will be explained in detail, test items are questions posed to examinees (both humans and machines). They require the examinee to generate appraisals (answers to the questions). Machine-generated appraisals are evaluated in terms of how well they compare to the human-generated ones.

Items have the following structural elements: (1) a scenario that is posed to the human or machine examinee and that consists of (1a) a target character whose emotion is to be predicted; a scenario involving the target (and possibly other characters). (2) a two-part emotion question that prompts the examinee to (2a) select through multiple choice an emotion descriptor that best matches the emotion he, she, or it predicts will likely be felt by the target character, and (2b) explain why the character might feel that way.

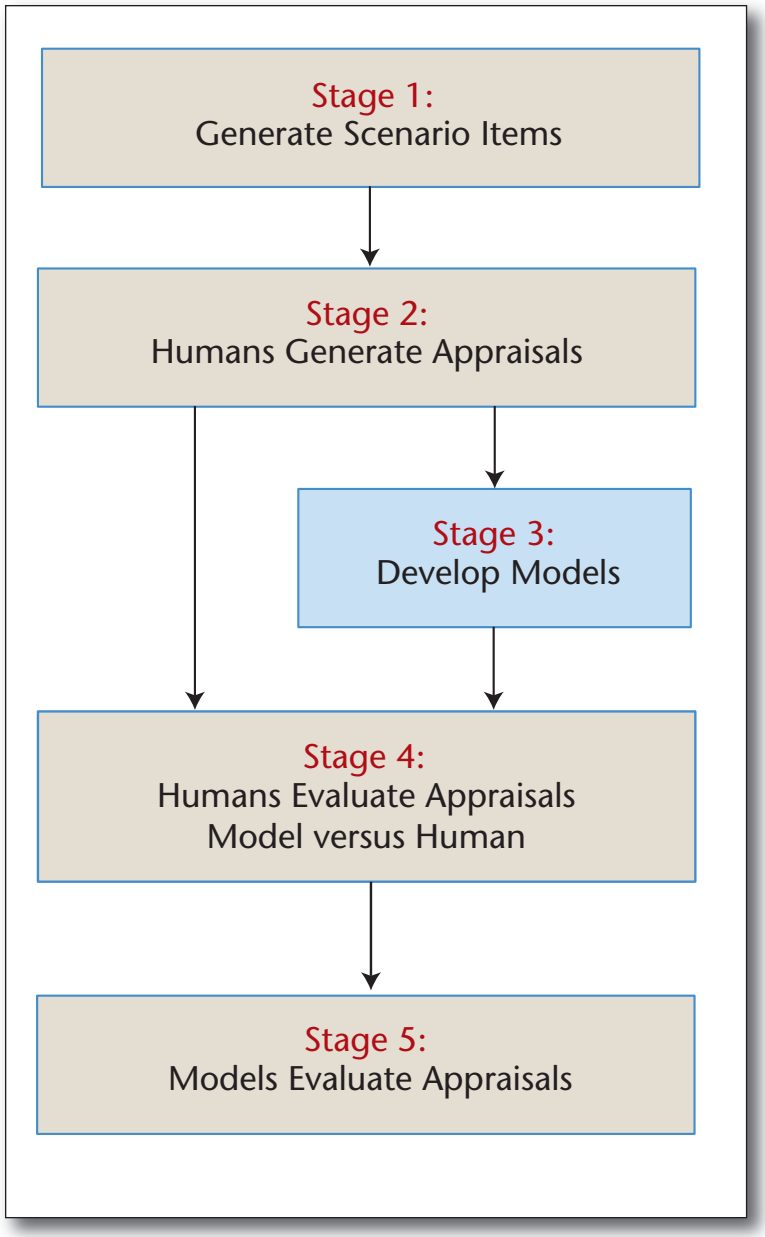

Figure 1. High Level Schematic of the Framework's Five Stages.

Stage 1: Generate Scenario Items

The purpose of stage one is to produce a set of scenario items that can be used later in the evaluation. The range of scenarios circumscribes the breadth of the modeling task.

In the early years of the competition, we will focus on simple scenarios (for example, "Eric wanted to ride the train but his father took him in the car. Was he happy or sad?") and in later years, move to ever more complex material from brief stories to, much later, entire novels.

Stage 2: Obtain Human-Generated Appraisals

The overall goal of this stage is to obtain a ground truth for the test. Concretely, the goal of this stage is to task a group of human participants to generate at least one appraisal for items produced in stage 1. Generating an appraisal involves choosing an emotion to answer the emotion question and producing an explanation for that answer.

Given the generativity of emotional appraisal we expect a wide range of responses even for a single scenario instance. Recall the example of appraisal data derived from the simple scenario in table 1.

The range of distinct appraisals should increase with the range of possible emotions from which to chose, the length of the allowable explanation, and the number of participants. That said, the increase at some point will level off because the themes of the *n*th participant's appraisal will start to overlap with those of earlier participants.

While the number of different scenario instances may circumscribe the generative breadth we require our computational models to cover, one might also say that the generative depth of the model is circumscribed by the number of distinct appraisals generated for each scenario.

Some of the resulting human-generated appraisals can be passed to the next stage as training data for modeling. The remainder are sequestered as a test set to be used during evaluation phases.

Stage 3: Develop Appraisal Models

The contestants, computational modelers, are challenged to develop a model that for any given scenario instance can (1) predict an appropriate emotion label for the target scenario character (for example, happy, sad, and so on); and (2) generate an appropriate natural language (NL) explanation for this prediction. Appropriate is judged by human raters in stage 4 in reference to human-generated appraisals. Contestants are given a sample of scenario instances and the corresponding human-generated appraisals to train or engineer their models.

Stage 4: Evaluate Appraisals: Model Versus Human

The purpose of this stage is to obtain an evaluation of how well a given model performs appraisal in comparison to humans. This is achieved by a new group of human participants serving as raters. The input to this process is a set of appraisals including human-generated ones from stage 2 and model-generated ones from stage 3.

Valence Reversal

Before being submitted to a human judge, each appraisal has a 50 percent chance of being subject to an experimental manipulation known as valence reversal. Operationally, this means replacing the emotion label of a given appraisal with a different

label of preferably "opposite" emotional valence. Under such a manipulation, happy would be replaced with sad, and sad with happy. For example:

Situation: Eric wants a train ride and his father gives him one.
Unreversed Appraisal: Eric feels happy because he got what he wanted.
Reversed Appraisal: Eric feels sad because he got what he wanted.

Reversal provides a contrast variable. We expect the statistical effect of reversal on appraisal quality to be strong. In contrast, if the model's appraisals are adequate, then among unreversed appraisals there should be no significant difference between human- versus model-generated appraisals. This methodology was successfully used in Jarrold (2004) and this article is essentially a scaling up of that approach.

Submission to Human Evaluators
Either the reversed or unreversed version of each appraisal is administered to at least one judge. The judges are to rate appraisals independently according to some particular subjective measure(s) of quality such as commonsensicality, believability, novelty, and so on. The measure is specified by the contest organizers. Judges are blinded to the reversal status — reversed or unreversed — and source — human or machine — of each item.

Stage 5: Model and Evaluate Human Meta-Appraisal

The purpose of this stage is to evaluate a model's ability not to generate but rather to validate appraisals. This capacity is important because human-level AToM involves not just the capacity to make one decent prediction and explanation of another agent's emotions in a given situation. It also involves breadth, the ability to assess the validity of any of the multitude of the generatable appraisals of that situation. If a model's pattern of quality ratings for all the stage 4 appraisals — be they model or human generated, reversed or unreversed — matches the pattern of ratings given by stage 4 human judges, then it demonstrates the full generative breadth of understanding.

The capacity for validating appraisals is important for another reason — detecting the authenticity of an emotional reaction. Consider the following:

Bob: How are you today?
Fred: Deeply depressed — no espresso.

People know that Fred is kidding. A deep depression is not a believable or commonsensical appraisal of a situation in which one is missing one's espresso.

The input to stage 5 is the output of stage 4, that is, human evaluations of appraisals. The appraisals evaluated include all manner of appraisals generated in prior stages: that is, both human and machine generated, both unreversed and reversed. These rated appraisals are segregated by the organizers into two groups, a training set and a test set.

Modelers are given the training set and tasked with enhancing their preexisting models by giving them the ability to evaluate the validity of others' appraisals. Once modeling is completed, the organizers evaluate the enhanced self-reflective models against the test data. Model appraisal ratings should be similar to human ratings — unreversed appraisals should receive high-quality ratings, and reversed ones, poorer ratings.

This phase may add new layers of model complexity and may be too difficult for the early years. Thus, for reasons of incrementality we consider it a stage that is phased in gradually over successive years.

Issues in Implementation

In this section we discuss specific issues associated with actually running the experiments and competitions.

Incrementality

Hector Levesque (2011) described the benefits of an incremental staged approach. Any challenge should be matched to existing capabilities. If too easy, the challenge will not be discriminative nor exciting enough to attract developers. If too hard, solutions will fail to generalize and developers will be discouraged. In addition, systems advance every year. In view of all of these needs, it is best to have a test for which it is easy to raise or lower the bar.

How can incrementality be implemented within the framework? As will be explained in the next section, parameterization of scenarios provides one relatively low-effort means of adapting the difficulty of the test.

Parameterization of Test Scenarios

It is important to be able to have a lot of test scenarios. More scenarios means more training data, a more fine-grained evaluation, a greater guarantee of comprehensive coverage. Cohen et al. (1998) used parameterization to create numerous natural language test questions that deviate from sample questions in specific controlled ways. The space of variation within given parameterization can be combinatorially large thus ensuring the ability to cover a broad range of materials. Parameterization was successfully used by Sosnovsky, Shcherbinina, and Brusilovsky (2003) to produce large numbers of training and test items for human education with relatively low effort.

A parameterized scenario is essentially a scenario template. Such templates can be created by taking an existing scenario and replacing particular objects in the scenario with variables of the appropriate type. Consider the following scenario instance:

Scenario: Tracy wants a banana. Mommy gives Tracy an apple for lunch.

Emotion Question: How will Tracy feel? (Choose from one of happy or sad.)
Explanation: Explain why she will feel that way (in less than 50 words).

This item can be parameterized by replacing Tracy, banana, Mommy, and others with variables as shown next.

Scenario Template
<target-character> wants <object1>. <alt-character> gives
<target-character> <object2> for <condition>

Answer Template
Emotion: How does <target-character> feel?
Choose from: <range of emotion terms / levels>
Explanation: <answer constraints — length, vocabulary, and others>

The range for each parameter is specified by the test administrator. For example the range for <object1> could include any object within the vocabulary of a four year old (for example, banana, lump of coal, chocolate, napkin). Additional item instances are instantiated by choosing values for the parameters of a given template. If parameters can take on a large set of values, a very large set of items can be generated.

To meet the needs of incrementality, one can increase (or decrease) the level of difficulty by increasing the range of values that scenario parameters may take on. Alternatively one can add more templates.

How the Framework Prevents Gaming Evaluation

Like any contest, it can be gamed by clever trickery that violates the spirit of the rules and evades constructive progress in the field. We describe a variety of gaming tactics and how the Framework prevents them.

Bag of Words to Predict Emotion

A bag of words (BOW) classifier assigns an input document to one of a predefined set of categories based on weighted word frequencies. Thus, one "cheat" is to use this simple technique to predict the correct emotion label.

One problem is that such classifiers ignore word order — thus "John loves Mary" and "Mary loves John" would assign the same emotion to Mary. Further, they are not generative and thus unable to produce novel explanations necessary in stage 4. In stage 5, it is hard to imagine how such a shallow approach would do well in evaluating the match between a scenario plus the appraisal emotion and explanation.

Chatbots

In stage 5, a chatbot will not do well because the task involves no NL generation — it just involves producing scores rating the quality of an appraisal.

In stage 4, the case against the chatbot is more involved. A chatbot hack for this stage would be to chose an arbitrary emotion and generate explanation through a chatbot. Chatty or snarky explanations might sound human but contain no specific content. Such explanations would intentionally be a form of empty speech hand-crafted by the modeler to go with any chosen emotion. For example, a Eugene Goostman-like agent could chose happy or sad and provide the same explanation, "Tracy feels that way just because that's the way she is."

A related but slightly more sophisticated tactic is always to chose the same emotion but devise a hand-crafted appraisal that could go with virtually any scenario. For example, "Tracy feels happy because she has a very upbeat personality — no matter what happens she's always looking on the bright side."

There are several reasons a chatbot will likely fail. First, we expect chatbots may be detectable through the human ratings. Although humans may sometimes provide answers like the above, more often than not, we expect their answers to exhibit greater specificity to the scenario and emotion chosen. We suspect that direct answers will generally receive higher ratings than chatty ones. Unlike the Turing test, there is no chance to build conversational rapport because there is no conversation and thus little for the chat bot to hide behind.

If necessary, contest administrators can give specific instructions to human judges to penalize appraisals that are ironic, chatty, not specific to the scenario, and so on. These considerations could be woven into a single overall judgment score per appraisal or by allowing for additional rating scales (for example, one dimension might be believability, another could be specificity, and so on). Elaborating the instructions in this way demands more training of judges and raises some issues associated with inter-rater reliability and multidimensional scoring.

The second countermeasure leverages falsifiability and the valence reversal manipulation done to all appraisals (machine as well as human generated) in stage 4. A chatbot lacks an (affective) theory of mind and thus does not know what kind of emotion goes with what kind of explanation in an appraisal. There should therefore be little to no dependency between its emotion labels and explanations. Put another way, being "theory free," chatbot "predictions" about other agents' appraisals are not falsifiable. Thus, valence-reversed appraisals from a chatbot will likely not be judged worse than their unreversed counterparts. Thus if a given appraisal and its reversed counterpart score about as well, this should factor negatively in that contestant's overall score.

Contest Evolution

An attractive design feature of this method is the number of contest configuration variables that can be readjusted each year in response to advancing technology, pitfalls, changing goals, or emphasis.

If organizers want to maximize the generative pro-

ductivity of contestants' models they can use fewer scenario instances; involve more human participants to generate more appraisals at stage 2; allow longer appraisal explanations with a larger vocabulary; and / or reward models that generated multiple appraisals per scenario.

By contrast, to maximize the breadth of appraisal domains organizers can have more scenario templates, more parameters in a template, more parameter values for a given parameter; or adjust the size of vocabulary allowed for a scenario.

To increase an appraisal's algorithm sophistication one can increase the number of characters in each scenario, increase the number of emotions to chose between, or allow multiple or mixed emotions to be chosen

The first contests should involve a small handful of emotions because Jarrold (2004) demonstrated there is a tremendous amount of complexity yet to be modeled to simply distinguish between happy and sad.

Affective reasoning requires a substantial body of commonsense knowledge. To bound the amount of such background knowledge required and focus efforts on affective reasoning, organizers can decrease the diversity of scenario characters — for example, human children ages 3 to 5; narrow the range of scenario parameters to a focused knowledge domain; or restrict the vocabulary or length allowed in explanations.

In later contest years, there may be rater disagreement for some of the more nuanced or subtle scenario or appraisal pairs due to differing cultural or social-demographic representativeness factors. A variety of options present themselves — make rater "cultural group" a contextual variable; increase the cultural homogeneity of the human raters; or remove appraisals with low interrater reliability from the contest.

Crowdsourcing

It is possible that considerable numbers of participants will be required at certain stages. For example, modelers may desire a large number of appraisals to be generated in stage 2 as training data. Prior work in dialog systems (Yang et al. 2010) or the creation of ImageNet (Su, Deng, and Fei-Fei 2012) (to pick just two of many crowdsourced studies) has shown that large numbers of people can be recruited online (for example, through Amazon Mechanical Turk) as a form of crowdsourcing. It is hoped that over successive years a large library of scenarios each with a large number of appraisals and associated human ratings could be collected in this way over time to compose an emotion-oriented ImageNet analog.

Public Interest

Newsworthiness and public excitement are important because prior competitive challenges such as Robocup, IBM Watson, and Deep Blue have demonstrated how these factors drive talented individuals and other resources to attack a problem. One factor helping the social-emotional Turing challenge is that emotional content has mass appeal and may be less dry than other challenges such as chess.

Stage 5, where machine- and human-generated appraisals are judged side by side, may be the most accessible media-worthy part of the framework. Prior stages may be reserved for a qualifying round, which may be of more scientific interest. Akin to the Watson competition, both human and machine contestants may be placed side by side while scenarios are presented to them in real time. Judges will score each appraisal blind to whether it was human versus machine generated. Scores can be read off one by one akin to a gymnastics competition.

Conclusion

We argue for the importance of assessing social-emotional intelligence among Turing test alternatives. We focus on a specific aspect of this capacity, affective theory of mind, which enables prediction and explanation of others' emotional reactions to situations. We explain how a generative logic can account for the diversity yet specificity of predicted affective reactions. The falsifiability of these predictions is leveraged in a five-stage framework for assessing the degree to which computer models can emulate this behavior. Issues in implementation are discussed including the importance of incremental challenge, parameterization, and resisting hacks. It is hoped that over successive years a large set of scenarios, appraisals, and ratings would accrue and compose a kind of affective version of ImageNet.

Acknowledgement

We would like to thank Deepak Ramachandran for some helpful discussions.

References

Clark, P. 2015. Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge! In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 4019–4021. Palo Alto, CA: AAAI Press.

Cohen, P. R.; Schrag, R.; Jones, E.; Pease, A.; Lin, A.; Starr, B.; Gunning, D.; and Burke, M. 1998. The DARPA High-Performance Knowledge Bases Project. *AI Magazine* 19(4): 25.

Gunning, D.; Chaudhri, V. K.; Clark, P. E.; Barker, K.; Chaw, S.-Y.; Greaves, M.; Grosof, B.; Leung, A.; McDonald, D. D.; Mishra, S.; Pacheco, J.; Porter, B.; Spaulding, A.; Tecuci, D.; and Tien, J. 2010. Project Halo Update — Progress Toward Digital Aristotle. *AI Magazine* 31(3): 33–58.

Howlin, P.; Baron-Cohen, S.; and Hadwin, J. 1999. *Teaching Children with Autism to Mind-Read: A Practical Guide for Teachers and Parents*. Chichester, NY: J. Wiley & Sons.

Jarrold, W. 2004. Towards a Theory of Affective Mind. Ph.D. Dissertation, Department of Educational Psychology, University of Texas at Austin, Austin, TX.

AI in Industry Columnists Wanted!

AI Magazine is soliciting contributions for a column on AI in industry. Contributions should inform *AI Magazine's* readers about the kind of AI technology that has been created or used in the company, what kinds of problems are addressed by the technology, and what lessons have been learned from its deployment (including successes and failures). Prospective columns should allow readers to understand what the current AI technology is and is not able to do for the commercial sector and what the industry cares about. We are looking for honest assessments (ideally tied carefully to the current state of the art in AI research) — not product ads. Articles simply describing commercially available products are not suitable for the column, although descriptions of interesting, innovative, or high impact uses of commercial products may be. Questions should be discussed with the column editors.

Columns should contain a title, names of authors, affiliations and email addresses (and a designation of one author as contact author), a 2–3 sentence abstract, and a brief bibliography (if appropriate). The main text should be brief (600–1,000 words) and provide the reader with high-level information about how AI is used in their companies (we understand the need to protect proprietary information), trends in AI use there, as well as an assessment of the contribution. Larger companies might want to focus on one or two suitable projects so that the description of their development or use of AI technology can be made sufficiently detailed. The column should be written for a knowledgeable audience of AI researchers and practitioners.

Reports go through an internal review process (acceptance is not guaranteed). The column editors and the *AI Magazine* editor-in-chief are the sole reviewers of summaries. All articles will be copyedited, and authors will be required to transfer copyright of their columns to AAAI.

If you are interested in submitting an article to the AI in Industry column, please contact column editors Sven Koenig (skoenig@usc.edu) and Sandip Sen (sandip-sen@utulsa.edu) before submission.

Levesque, H. J. 2011. The Winograd Schema Challenge. In *Logical Formalizations of Commonsense Reasoning: Papers from the 2011 AAAI Spring Symposium*, 63–68. Palo Alto, CA: AAAI Press.

Ortony, A. 2001. On Making Believable Emotional Agents Believable. In *Emotions in Humans and Artifacts,* ed. R. Trappl, P. Petta, and S. Payr, 189–213. Cambridge, MA: The MIT Press.

Popper, K. 2005. *The Logic of Scientific Discovery.* New York: Routledge / Taylor & Francis.

Sosnovsky, S.; Shcherbinina, O.; and Brusilovsky, P. 2003. Web-Based Parameterized Questions as a Tool for Learning. In *Proceedings of E-Learn 2003: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education,* 309–316. Waynesville, NC: Association for the Advancement of Computing in Education.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing Annotations for Visual Object Detection. In *Human Computation: Papers from the 2012 AAAI Workshop*. AAAI Technical Report WS-12-08, 40–46. Palo Alto, CA: AAAI Press.

Yang, Z.; Li, B.; Zhu, Y.; King, I.; Levow, G.; and Meng, H. 2010. Collection of User Judgments on Spoken Dialog System with Crowdsourcing. In *2010 IEEE Spoken Language Technology Workshop (SLT 2010),* 277–282. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

William Jarrold is a senior scientist at Nuance Communications. His research in intelligent conversational assistants draws upon expertise in ontology, knowledge representation and reasoning, natural language understanding, and statistical natural language processing (NLP). Throughout his career he has developed computational models to augment and understand human cognition. In prior work at the University of California, Davis and the SRI Artificial Intelligence Lab he has applied statistical NLP to the differential diagnosis of neuropsychiatric conditions. At SRI and the University of Texas he developed ontologies for intelligent tutoring (HALO) and cognitive assistants (CALO). Early in his career he worked at MCC and Cycorp developing ontologies to support commonsense reasoning in Cyc — a large general-purpose knowledge-based system. His Ph.D. is from the University of Texas at Austin and his BS is from the Massachusetts Institute of Technology.

Peter Z. Yeh is a senior principal research scientist at Nuance Communications. His research interests lie at the intersection of semantic technologies, data and web mining, and natural language understanding. Prior to joining Nuance, Yeh was a research lead at Accenture Technology Labs where he was responsible for investigating and applying AI technologies to various enterprise problems ranging from data management to advanced analytics. Yeh is currently working on enhancing interpretation intelligence within intelligent virtual assistants and automatically constructing large-scale knowledge repositories necessary to support such interpretations. He received his Ph.D. in computer science from The University of Texas at Austin.