AI magazine
Volume 39 Number 3
Fall 2018

AI REBEL AGENTS!

AIIDE Comes to Canada for 2018!

The Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-18)

November 13–17, 2018
The University of Alberta
Edmonton, Alberta,Canada
www.aiide.org

Jonathan Rowe (NC State) *General Chair*
Gillian Smith (WPI) *Program Chair*

Courtesy wwing, iStock

AI magazine

VOLUME 39, NUMBER 3 Fall 2018 ISSN 2371-9621 (online) ISSN 0738-4602 (print)

Cover: Rebel AI by James Gary, New York, New York.

ARTICLES

DEPARTMENTS

Coming Up in the Winter Issue

Catch Up on Events!
A plethora of workshop, competition, and conference reports on AI have come in during the past months. To accommodate the accumulated backlog, we will be devoting much of the winter issue to publishing them.

Place Your Bets!
The AI Bookie column will document highlights from AI Bets, an online forum for the creation of adjudicatable predictions and bets about the future of AI. While it is easy to make a prediction about the future, this forum was created to help researchers craft predictions whose accuracy can be clearly and unambiguously judged when they come due. The bets will be documented on line, and regularly in this publication in a new column, The AI Bookie. We encourage bets that are rigorously and scientifically argued. We discourage bets that are too general to be evaluated, or too specific to an institution or individual. The goal is not to continue to feed the media frenzy and pundit predictions about AI, but rather to curate and promote bets whose outcomes will provide useful feedback to the scientific community.

AI Research in Germany!
The worldwide AI column for winter will feature an in-depth look at a German AI research focus: hybrid reasoning for intelligent systems.

AI magazine

An Official Publication of the Association for the Advancement of Artificial Intelligence

aimagazine.org

ISSN 0738-4602 *(print)* ISSN 2371-9621 *(online)*

Submissions

Submissions information is available at aaai.org/ojs/index.php/aimagazine/information/authors. Authors whose work is accepted for publication will be required to revise their work to conform reasonably to *AI Magazine* styles. Author's guidelines are available at aaai.org/ojs/index.php/aimagazine/about/submissions#authorGuidelines. If an article is accepted for publication, a new electronic copy will also be required. Although *AI Magazine* generally grants reasonable deference to the author's words, the *Magazine* retains the right to determine the final published form of every article.

Advertising

AI Magazine, 2275 East Bayshore Road, Suite 160, Palo Alto, CA 94303, (650) 328-3123; *Fax* (650) 321-4457. *Web:* aimagazine.org.

Microfilm, Back, or Replacement Copies

Replacement copies (for current issue only) are available upon written request and a check for $25.00. Back issues are also available (cost may differ). Send replacement or back order requests to AAAI. Microform copies are available from ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106. Telephone (800) 521-3044 or (734) 761-4700.

Copying Articles for Personal Use

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, or for educational classroom use, is granted by AAAI, provided that the appropriate fee is paid directly to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. Telephone: (978) 750-8400. Fax: (978) 750-4470. Website: www.copyright.com. E-mail: info@copyright.com. This consent does **not** extend to other kinds of copying, such as for general distribution, resale, advertising, Internet or internal electronic distribution, or promotion purposes, or for creating new collective works. Please contact AAAI for such permission.

Address Change

Please notify AAAI eight weeks in advance of a change of address. Send electronically via MemberClicks or by e-mailing us to membership18@aaai.org.

Subscriptions

AI Magazine (ISSN 0738-4602) is published quarterly in March, June, September, and December by the Association for the Advancement of Artificial Intelligence (AAAI), 2275 East Bayshore Road, Suite 160, Palo Alto, CA 94303, telephone (650) 328-3123. *AI Magazine* is a direct benefit of membership in AAAI. Membership dues are $145.00 individual, $75.00 student, and $285.00 academic / corporate libraries. Subscription price of $50.00 per year is included in dues; the balance of your dues may be tax deductible as a charitable contribution; consult your tax advisor for details. Inquiries regarding membership in the Association for the Advancement of Artificial Intelligence should be sent to AAAI at the above address.

PERIODICALS POSTAGE PAID at Palo Alto CA and additional mailing offices. *Postmaster:* Change Service Requested. Send address changes to *AI Magazine*, 2275 East Bayshore Road, Suite 160, Palo Alto, CA 94303.

Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems

David C. Brock

■ *Much of the contemporary moment's enthusiasms for and commercial interests in artificial intelligence, specifically machine learning, are prefigured in the experience of the artificial intelligence community concerned with expert systems in the 1970s and 1980s. This essay is based on an invited panel on the history of expert systems at the AAAI-17 conference, featuring Ed Feigenbaum, Bruce Buchanan, Randall Davis, and Eric Horvitz. It argues that artifical intelligence communities today have much to learn from the way that earlier communities grappled with the issues of intelligibility and instrumentality in the study of intelligence.*

If it is indeed true that we cannot fully understand our present without knowledge of our past, there is perhaps be no better time than the present to attend to the history of artificial intelligence. Late 2017 saw Sundar Pichai, the CEO of Google, Inc., opine that "AI is one of the most important things that humanity is working on. It's more profound than, I don't know, electricity or fire" (Schleifer 2018). Pichai's notable enthusiasm for, and optimism about, the power of multilayer neural networks coupled to large data stores is widely shared in technical communities and well beyond. Indeed, the general zeal for such artificial intelligence systems of the past decade across the academy, business, government, and the popular imagination was reflected in a recent *New York Times Magazine* article, "The Great AI Awakening" (Lewis-Kraus 2016). Imaginings of our near-future promoted by the World Economic Forum under the banner of a Fourth Industrial Revolution place this "machine learning" at the center of profound changes in economic activity and social life, indeed in the very meaning of what it means to be human (Schwab 2016).

Far too often, these pronouncements and perspectives fail to attend to artificial intelligence's previous awakenings. Over 30 years ago, in 1985, Allen Newell — one of the key figures in the emergence of artificial intelligence as a field in the 1950s and the first president of the Association for the Advancement of Artificial Intelligence (AAAI) — wrote: "There is no doubt as far as I am concerned that the development of expert systems is the major advance in the field during the last decade … The emergence of expert systems has transformed the enterprise of AI" (Bobrow and Hayes 1985). This article frames and presents the discussion at an invited panel, "AI History: Expert Systems," held at the AAAI-17 conference in San Francisco, February 6. The panel's purpose was to open up this history of expert systems, its transformational aspects, and its connections to today's "AI awakening" (Brock 2017).[1]

The history panel featured four key figures in the story of expert systems and was moderated by the director of the Center for Software History at the Computer History Museum, David C. Brock. Edward Feigenbaum is the Kumagai Professor Emeritus at Stanford University. He was president of AAAI in 1980–81, and he was awarded the ACM Turing Award for 1994 in part for his role in the emergence of expert systems. Bruce Buchanan is a university professor emeritus at the University of Pittsburgh, and he was president of AAAI in 1999–2001. Randall Davis is a professor in the Electrical Engineering and Computer Science Department at the Massachusetts Institute of Technology. He was president of AAAI in 1995–97. Eric Horvitz, MD, PhD, is a technical fellow of the Microsoft Corporation, where he also serves as the managing director of Microsoft Research. He was president of AAAI in 2007–09.

Perspectives from two historians of science and technology provide a very useful framework for approaching the history of expert systems, and the discussion at the 2017 history panel. Michael Mahoney, a history professor at Princeton University, was a particularly influential figure in the study of the history of computing. In a 2005 article, "The Histories of Computing(s)," Mahoney presented a concise statement of several of his most fundamental insights from his many years of study in the field. "[T]he history of computing," he wrote, "is the history of what people wanted computers to do and how people designed computers to do it. It may not be one history, or at least it may not be useful to treat is as one. Different groups of people saw different possibilities in computing, and they had different experiences as they sought to realize these possibilities. One may speak of them as 'communities of computing,' or perhaps as communities of practitioners that took up the computer, adapting to it while they adapted it to their purposes" (Mahoney 2005).

For Mahoney, a defining activity of these various communities of computing was creating software that, for him, constituted the modeling of certain features of the physical or social world. Making software for Mahoney was putting the world into the computer: "[Software] Design is not primarily about computing as commonly understood, that is, about computers and programming," he explained, "It is about modeling the world in the computer … about translating a portion of the world into terms a computer can 'understand' … [P]utting a portion of the world into the computer means designing an operative representation of it that captures what we take to be its essential features. That had proved, as I say, no easy task … If we want critical understandings of how various communities of computing have put their portion of the world into software, we must uncover the operative representations they have designed and constructed" (Mahoney 2005).

An historical expert on a very different subject — the Scientific Revolution of the 16th and 17th centuries — and a history professor at Cornell University, Peter Dear provides an account of the two fundamental purposes toward which scientific and technical communities, including Mahoney's communities of computing, direct their activities: *intelligibility* and *instrumentality*. Crudely summarized, Dear proposes that there are two distinct, separate, but intertwined purposes that have motivated these communities. One is a pursuit of the "intellectual understanding of the natural world," including ourselves. This is the striving to make sense of the world, to provide satisfying answers to basic questions about how things are, and why they are. Dear notes, "Evidently … there are not timeless, ahistorical criteria for determining what will count as satisfactory to the understanding. Assertions of intelligibility can be understood only in the particular cultural settings that produce them." The other purpose is the creation of effective techniques that afford, as Dear puts it, "… power over matter, and indirectly, power over people." Here the goal is the creation and refinement of an "operational, or instrumental, set of techniques used to do things … Such accomplishments … in fact result from complex endeavors involving a huge array of mutually dependent theoretical and empirical techniques and competences" (Dear 2006, pp. 1–14).

Both goals of intelligibility and instrumentality can clearly be seen in the community of computing — perhaps, more properly, communities — involved in artificial intelligence. On the side of intelligibility lie questions about our understanding of human intelligence: how is it that we reason, learn, judge, perceive, and conduct other mental actions? On the side of instrumentality reside myriad activities to create computer systems that match or exceed human performance in tasks associated with the broad concept of "intelligence." This instrumental dimension in the history of artificial intelligence is of a piece with a major through-line in the history of comput-

ing more generally, in which scientists and engineers developed machines to, at first, aid human practices of mathematical calculation, but these machines quickly came to exceed human intelligence's unaided capacity for calculation by many orders of magnitude. From one angle, the pursuit of instrumentality in artificial intelligence may be seen as an effort to extend this surpassing of the human capacity for mathematical calculation to additional capabilities and performances.

It is nonetheless very clear that intelligibility was an enormously motivating goal for the emergence of artificial intelligence. In his reflections on the history of artificial intelligence to 1985 cited above, Allen Newell also powerfully surfaced the importance of intelligibility to the artificial intelligence community of which he was a part: "One of the world's deepest mysteries — the nature of mind — is at the center of AI. It is our holy grail. Its discovery (which will no more be a single act than the discovery of the nature of life or the origins of the universe) will be a major chapter in the scientific advance of mankind. When it happens (that is, as substantial progress becomes evident), there will be plenty of recognition that we have finally penetrated this mystery and in what it consists. There will be a coherent account of the nature of intelligence, knowledge, intention, desire, etc., and how it is possible for the phenomena that cluster under these names to occur in our physical universe." (Bobrow and Hayes 1985, 378)

As Ed Feigenbaum explained in the AAAI-17 panel on the history of expert systems, and in terms that directly echo Mahoney's view of software as modeling, the roots of expert systems begin in the first decade of the AI community's pursuit of intelligibility:

> Let me go back to the first generation, 1956–1965. The AI field began with a set of ideas like a Big Bang about the nature of human thought and how to model it by computer. The ideas came from some truly remarkable individuals. Their ideas about problem-solving, recognition, and learning were fundamental but few. There was a focus on *deductive* [emphasis added] reasoning processes — logic-based or based on heuristic search — and on the generality of these reasoning processes.
>
> In 1962, I wrote about the need to move beyond deductive tasks to the study of *inductive* [emphasis added] processes and tasks, which I viewed as dominant in human thought. But it took me two years until 1964 to frame this pursuit in a way that I thought would be fruitful. Being an empirical scientist, I was looking for some people to observe with the idea of modeling their behavior. Now why did I choose to model the thinking of scientists? Because I saw them as being skilled, professional, induction specialists constructing hypotheses from data, and I thought they would be reflectively curious and reductionist enough to enjoy allowing others like me to explore their thought processes.

Allen Newell and Herbert Simon created perhaps the first artificial intelligence program, Logic Theorist, in 1955–56 as a model of deductive reasoning, and also as a kind of self-modeling of their own familiarity with creating proofs in formal logic and mathematics. That this model reproduced a set of proofs created by Bertrand Russell and Alfred North Whitehead in their seminal *Principia Mathematica*, and even arguably improved upon one, was taken as strong confirmation of their model.

AAAI Archive File Photo.

Edward Feigenbaum.

In contrast, Feigenbaum's interest was in modeling the inductive reasoning that he believed was vital in human intelligence generally. From his study of and with Newell and Simon, Feigenbaum drew the lesson that such modeling of human reasoning in the computer needed specificity. He believed that a particular "test bed" was required to exercise and refine the model, and to draw conclusions from it. After extended deliberation, Feigenbaum decided that "professional inducers," people who were paid to make inductions, would make for a productive test bed. He would model, therefore, the reasoning of empirical scientists. To his surprise, Feigenbaum found that a prominent empirical scientist, indeed one of the world's leading geneticists, shared his interest in the possibilities for computational models of scientific reasoning:

AAAI Archive File Photo.

Allen Newell and Herbert Simon.

In 1964, I was fortunate to find an enthusiastic collaborator, Joshua Lederberg, Professor of Genetics and Nobel Prize winner at Stanford. He too was interested in the question "Can AI model scientific thinking?" So our work together began in 1965, after I joined Stanford. As an aside, Lederberg's mind was one of great vision and insight, one of the top minds of the 20th century, in my view. But Lederberg was the gift that kept giving. In 1966, Lederberg recruited for us Professor Carl Djerassi, one of the most influential chemists of all time, the father of the Pill [birth control pill] and the head of Stanford's mass spectrometry laboratory.

As I said, I'm an empirical scientist, not at theoretical one. I needed a test bed in which to do these AI experiments. Lederberg suggested the problem that he was working on, inferring hypotheses about organic molecular structures from the data taken by an instrument called a mass spectrometer. Lederberg was doing research for NASA on the design of a Mars probe, designing a mass spectrometer system for detecting life-precursor molecules such as amino acids.

In this experimental setting, the test bed, we could measure, month by month, how well our program — which was called Heuristic DENDRAL, or later just DENDRAL for short — was performing compared with the performance of Djerassi's PhD students and postdocs on the same problem.

Throughout the 1960s, Feigenbaum — in collaboration with Bruce Buchanan and others — continued to evolve the model of the organic chemists in Carl Djerassi's laboratory, and in particular their capability to interpret the data about particular sorts of organic compounds from their mass spectrometry instrumentation. This modeling had two basic features. For one, the artificial intelligence researchers developed the model of inductive reasoning processes in DENDRAL. In addition, they modeled the organic chemists' knowledge as a store of rules, roughly in the form of "If, Then" statements.

Photograph Courtesy, National Library of Medicine

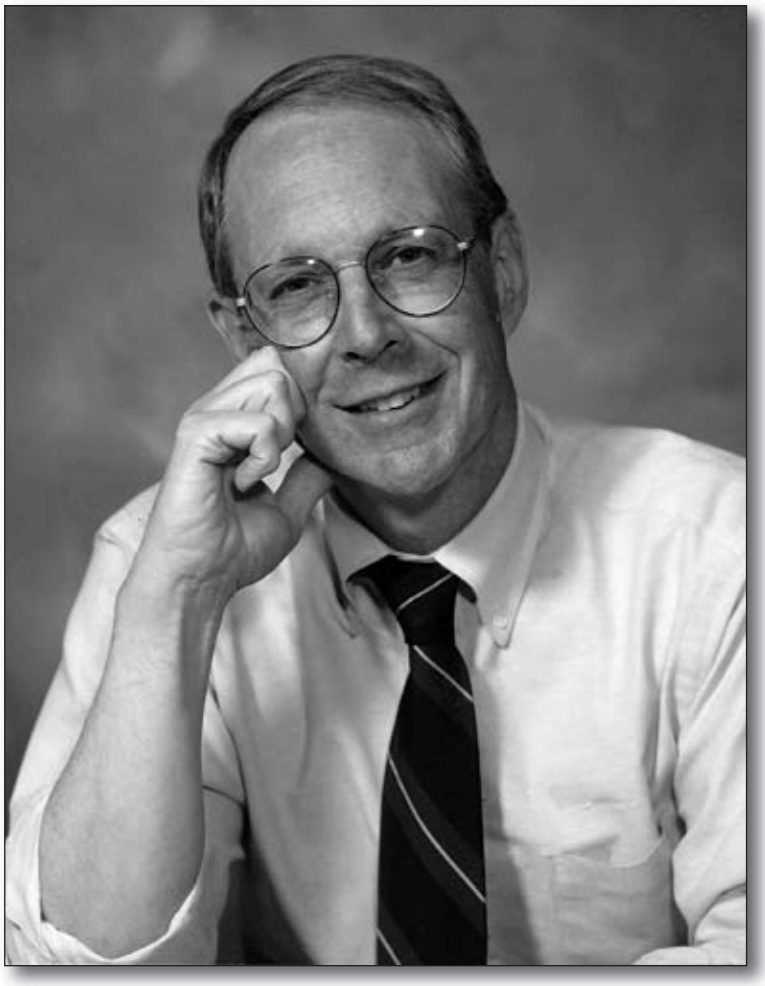
The Original Dendral Team, Twenty-Five Years Later.

The development of DENDRAL along these two lines served both intelligibility and instrumentality. On intelligibility, the motivating question was how to explain different abilities in human reasoning. How is it that experts render better judgments? Is it that some people possess a fundamentally superior way of reasoning than others? Or is it rather that accumulated, organized experience is the key, with people reasoning in more or less the same fashion?

Research in the nascent field of expert systems sought to address the question of special reasoning versus accumulated knowledge as the basis for expert judgments. With this also lay the promise of instrumentality. If the key to expert performance could be unlocked by investigations of a computer system, might such a system come to match and then exceed the performance of any human expert, as was already the case with mathematical calculation? This co-implication of intelligibility and instrumentality was not lost on the primary sponsor of artificial intelligence research in the United States, at least until very recently: The US Department of Defense's famous Defense Advanced Research Projects Agency (DARPA).

As Feigenbaum recalled, by 1968 he and his colleagues were prepared to draw conclusions from work on DENDRAL, from what the performance of their model of this human expertise had shown them. The conclusions were, themselves, an induction from a model of chemists' expertise in the interpretation of mass spectra of particular families of organic molecules:

> So we proceeded experiment by experiment in this test bed … moving toward higher levels of performance in mass spectral analysis, that propelled the movement to higher levels of behavior. What allowed that was the knowledge that we were systematically extracting from the experts in Djerassi's lab. Most of this knowledge was specific to mass spectrometry, including much heuristic knowledge, but some was general chemistry knowledge.
>
> In 1968, I was writing a paper with Bruce [Buchanan] and Josh Lederberg in which we chose to summarize the evidence from the many DENDRAL experiments from mid-1965 to mid-1968. It was evident that the improvement in DENDRAL's performance as a mass spectrum analyst was almost totally a function of the amount and quality of the knowledge that we had obtained from Djerassi's experts, and that it was only



AAAI Archive File Photo.

Bruce Buchanan.

> weakly related to any improvements that the AI scientists like me and Bruce had made to the reasoning processes used in DENDRAL's hypothesis formation.
>
> So in 1968, I called this observation the "Knowledge is Power Hypothesis." One data point. Later, as the evidence accumulated from dozens of — or hundreds of — expert systems, I changed the word "hypothesis" to "principle." The title of the 1968 paper was specifically worded to contrast what we called the DENDRAL case study with the main paradigm of the first generation of AI that focused on the generality of problem-solving. Those of you who are old enough in the audience remember GPS [General Problem Solver]. This was a major paradigm shift for AI research, but it took more than five years for the new paradigm to take hold.

Continued modeling of human experts, and in particular scientists and engineers, led to expert systems that, for very specific kinds of expertise, could meet and even exceed the performance of human experts. This achievement of instrumentality — a novel capability to do things, namely to exceed some performances of human experts — eventually led to a great enthusiasm for expert systems within the artificial intelligence community and its military patrons, then quickly drawing in corporations, investors, entrepreneurs, and the popular press.

Yet the route to these enthusiasms was painstaking work along the same two developmental lines for modeling human expertise as computer systems: changes to the inductive reasoning processes and also to the representation of expert knowledge and the means of making that representation. Bruce Buchanan concentrated his efforts on the latter, which, he explained eventually became known as *knowledge engineering:*

> Well, we didn't use the term "knowledge engineering" until the 1970s, but we did talk, in a 1969 paper that Ed and I were coauthors of with Georgia Sutherland, about knowledge elicitation in AI. It was at a machine intelligence workshop and people there were somewhat stunned that we were talking about organic chemistry. John McCarthy rescued me during a talk by saying to somebody who was giving me a hard time, "Sit down, be quiet, you might learn something." I forever after loved that man.
>
> Well, there were other groups working on knowledge representation at the same time. Harry Pople and Jack Myers at [the University of Pittsburgh] were working with an emphasis on ontologies and mechanisms. Peter Szolovits was working with Dr. Bill Schwartz, and that led to a lot of work on the object-oriented frames paradigm. Cas Kulikowski was working on knowledge engineering with Dr. Aaron Safir at Rutgers. There was work in Europe … There was a lot of isolated work in France replicating some of the early expert systems work, and several projects in France from commercial firms, Schlumberger and Elf Aquitaine being two of the most important. The Japanese Institute for New Generation [Computer] Technology, ICOT, was working on fifth-generation computing largely from a point of view of logic. The French were using Prolog and so did the Japanese.
>
> So I think our lesson there, the important part, was in coding knowledge. The language you use — Prolog or LISP or something else — it didn't matter nearly so much as the paradigm of starting with an expert's knowledge. But we also saw in that time that knowledge engineering could focus on the objects and their relationships in an ontology: a hierarchy. They could focus on the inferential mechanisms that were going on, and in DENDRAL we were very much interested in what we called the "situation-action rules" at the time. There was an action in the right-hand side of the rule, not just a Prolog kind of logical assertion.

For Buchanan, as with Feigenbaum, the motivation of intelligibility was, at least initially, primary for the development of expert systems. Buchanan recalled:

> Well, I was fascinated with the reasoning process.... My dissertation [on the philosophy of science] was on the process of discovery and trying to orient it into a framework. In the middle of my dissertation, I got to know Ed Feigenbaum in 1963 and began reading the

early AI work, mostly by Newell and Simon, and the RAND Corporation publications. And it convinced me that we could make a logic of discovery out of the paradigm of search, a constrained search. So that was the focus within which I got to know Ed and came into this field.

So when Ed offered the opportunity to work on DENDRAL, it was just a godsend because here was an opportunity — one of the early experiments in *computational philosophy* [emphasis added] — to try to do philosophy but with an empirical bent, namely writing programs that would actually produce something that was testable. Then started these discussions with Carl Djerassi's postdoc Alan Duffield and his reasoning process about mass spectrometry and the interpretation of mass spectra was just exactly what I needed in order to instantiate some of those ideas about capturing knowledge, about data interpretation, and then, subsequently, theory formation.

You've got to, I think, want to contrast this work with other work that was going on at the same time in which people were acting as their own experts. I could not, by any means, claim to be an expert in chemistry or certainly not mass spectrometry. There were other people though: like Joel Moses [and his colleagues] at MIT, who was an expert in symbolic mathematics; and Tony Hearn in symbolic algebra; Ken Colby in psychiatry, Todd Wipke in chemistry. These people were also doing knowledge elicitation but it was from their own heads, so it was more like just introspection.

As Buchanan showed, modeling the expertise of others as opposed to introspective self-modeling did not fully distinguish the subfield of expert systems from other areas of artificial intelligence work. Rather, the development of expert systems relied on mixtures of both kinds of modeling.

Whether from self-modeling or modeling of others, Buchanan and others created a particular kind of representation of the modeled knowledge known as *production rules,* a system of "If, Then" statements. Buchanan explained:

There was a logician who published a paper, Emil Post in 1943, using "production rules" as a complete logical system. That certainly has to be one of the precursors of our work on production systems. Although we weren't following it directly, it was certainly there.

Art Samuel's work on the checker player: Art had interviewed experts to understand … the feature vector and then he did a good deal of reading about checkers.… And the influential part about that was … his machine learning component — that once you had the expertise in, in a first-order form, it could be improved … automatically. That impressed me a great deal and I always wanted to be able to do that.

So we subsequently developed a learning program we called META-DENDRAL that did learn the rules of mass spectrometry from empirical data. A footnote on that. The data were very sparse. It took about one graduate student one year to obtain and interpret one mass spectrum, so we couldn't ask for very much data. This was not a big data problem. And we substituted knowledge for data in that and we continued to believe, I continue to believe, that that's a good tradeoff when you don't have enough data for the big data kind of learning.

So just three other things:

John McCarthy's paper "Programs with Common Sense" made a very strong case that whatever knowledge a program was using, it had to be in a form that it could be changed from the outside … that was something Art Samuel was doing with the feature vector weights, but something also we were doing with the DENDRAL rules of mass spectrometry that made a very big difference.

Now, Bob Floyd and Allen Newell developed a production rule compiler at CMU [Carnegie Mellon University] and that led to [Feigenbaum's PhD student] Don Waterman's work on representing the knowledge about poker play in a production system. Don's work was extremely influential in giving us the sense that that was the way to do it.

And, finally, Georgia Sutherland had been working with Joshua Lederberg on knowledge elicitation and putting that knowledge into separate tables. They were not rules, they were constraints for the chemical structure generator, but they were referenced in a way that they could be changed as data. Those were in my mind the most important precursors.

This is not to say that Buchanan and others believed that these production rules were the last word in modeling human expert knowledge in a computer. When asked if he believed that the representation of knowledge as a rule had limitations, Buchanan replied, "We saw a lot." He continued:

And our friends at MIT and elsewhere were quick to point out others. We wanted to be testing the limits of a very simple production rule architecture and we knew it was limited, we just didn't know quite where it would break and why. So that was the nature of many of the experiments that we subsequently published in the MYCIN book [*Rule-Based Expert Systems* by Bruce Buchanan and Edward Shortliffe] and I would encourage people to take a look.

But let me quote from that, "Our experience using EMYCIN to build several expert systems has suggested some negative aspects to using such a simple representation for all of the knowledge. The associations that are encoded in rules are elemental and cannot be further examined except with," some additional text that we put into some extra ad hoc slots. So, continuing the quote, "A reasoning program using only homogeneous rules with no internal distinctions among them thus fails to distinguish among several things, chance associations, statistical correlations, heuristics based on experience, cause of associations, definitions, knowledge about structure, taxonomic knowledge," all of those were things that we were failing to capture in the very simple more or less flat organization.

The modeling of human experts' knowledge in expert systems as production rules was provisional, intended to reveal what kinds of performance they could produce and what they could not.

From Buchanan's involvement with knowledge engineering into the middle 1980s, he drew three fundamental lessons:

AAAI Archive File Photo.

Randall Davis at AAAI-92.

> There are three different perspectives. From the point of view of computer science, I think the Knowledge is Power Principle is the most important lesson, and it's one we certainly have said more than once. At the level of user acceptance, I think the main lesson is that a program needs to be able to explain its reasoning in any decision-making situation with high stakes. And third, at the implementation level, the main lesson is flexibility. In the final chapter of the MYCIN book, Chapter 36 … we wrote, "If we were to try to summarize in one word why MYCIN works as well as it does, the word would be flexibility. By that, we mean that the designers' choices about programming constructs and knowledge structures can be revised with relative ease and that the users' interactions with the system are not limited to a narrow range in a rigid form." So: knowledge, explanation, flexibility.

These three issues — knowledge, explanation, and flexibility — have also become central to contemporary discussions of multilayer neural networks and machine learning, with "knowledge" now taking the guise of the datasets used for training, and "flexibility" now largely couched in terms of the fragility or brittleness of machine learning systems. Explanation, or the lack thereof, however remains a key challenge for today's artificial intelligence efforts.

Randall Davis, who has placed explanation at the center of his work in artificial intelligence, saw the development of expert systems from the middle 1970s to the middle 1980s continue to evolve the two main strands of development that had been present since the start: the reasoning processes and the representation of expert knowledge. Much of that continued development, in Davis' view, was in the direction of generalization:

> One interesting lesson was the value in generalizing the work that had been done. Initially of course, this was the generalization from the individual applications to the so-called expert system "shells." They were put into fairly wide use. Lots of applications got built using them. Not all of these things were acknowledged as expert systems, and some of them I think weren't particularly true to the original inspiration and architecture.
>
> But the real point is they adopted and spread the ideas — two good ideas, namely that to be good, a program needed a reasonably large collection of task-specific knowledge and, second, that there were at least semi-principled ways to gather and represent that knowledge. These tools were in some ways analogous to the open sourcing of deep learning tools that are being distributed now and, like those tools, they provide a substantial boost to people who are trying to build these systems. But, as always, it's best if you are one of the anointed ones who know how to use the tools. That's how you get the best use out of them. I think it was true then and I think it's true now.
>
> Another interesting lesson was the way certain insights seemed to echo through the years. We kept seeing the value of explicit, readable representations of knowledge using familiar symbols in knowledge representation, expressing knowledge separately from its intended use.… The most immediate consequence of these ideas is to enable multiple uses of the same knowledge, so we had systems that were doing diagnosis with a body of knowledge, explaining the reasoning and the result using that same body of knowledge, and then going ahead to teaching somebody with that same body of knowledge, all from a single representation. And just as when you're building a program, the virtues of encoding something once saves you from version skew, it was the same thing here in version skew in the knowledge.
>
> One of the nice examples of this multiple uses of knowledge came out of the work of Bill Clancy where the basic inspiration was: if we can debrief experts and transfer their knowledge into the program, is it possible to get the program to transfer the same knowledge into the head of a student? That, in turn, led to lots of interesting work … in understanding what was insufficient about MYCIN's very simple rule-based representation. The systems got considerably more power when that knowledge which was implicit in the rules got explicitly captured and represented in some of the work that Bill Clancy did.
>
> Another outcome in that body of work and in other work on intelligent tutoring was the idea that explicit representations of knowledge permits a kind of mind reading, or at least mind inferring. If I have an explicit model of what someone needs to know to accomplish a task and they make a mistake in doing that task, say a diagnosis, I can plausibly ask myself given

my model of what they ought to know, what defect in that knowledge would have produced the error that they produced. It's an interesting form of, if not mind reading, at least mind inferring.

The final lesson was the ubiquity of knowledge, task-specific knowledge. Of course, for example, medicine. Knowledge about debriefing: How do we get the knowledge out of the head of the expert into the program? Knowledge about tutoring: How do we transfer that into the students and knowledge about the general task? Diagnosis as a particular variety of inference. Everywhere we looked there was more to know, more to understand, and more to write down in explicit forms.

These matters of rendering implicit knowledge explicit, of mind inferring, and of knowledge transfers are all of a kind with Davis' concern for explanation and transparency in artificial intelligence. He explained:

I've been interested in these issues for several decades. The bad news, for me at least, is after all that time … the idea that AI programs ought to be explainable is now in wide circulation. Alas, where were you guys 40 years ago? There's a lot of interest, of course, in getting understandable AI. There's lots of experiments in getting deep learning systems to become more transparent. As many of you know, Dave Gunning has a DARPA program on "explainable AI," and the overall focus in looking at AI not as automation working alone but as having AI work together with people. All of these things are going to work better with systems that are explainable and transparent.

So there's lots of reasons to want this, the most obvious ones are trust and training. Trust is obvious. If we've got autonomous cars or medical diagnosis programs, we want to know we can trust the result. But I think training is another issue. If the system makes a mistake, what ought we to do about it? Should we give it more examples? What kind of examples? Is there something it clearly doesn't know? What doesn't it know? How do we explain it to the system? So transparency helps with making the system smarter.

One key issue I think is the representation and inference model. In what sense is the representation and inference model in our programs either similar to or a model of human reasoning? It seems to me that the closer the system's representations and model of reasoning are to human representations and reasoning, the easier it's going to be to bridge that gap and make them understandable.

A kind of counterexample of this is currently the vision systems, the deep learning vision systems that are doing a marvelously impressive job of image labeling for example. They're said to derive their own representations and that's great, but it's also a problem because they're deriving their own representations. If you want to ask them why they thought a particular picture was George Washington, what could they possibly say?

Now the issue is made a little bit worse by the collection of papers these days that show that deep learning vision systems can be thrown off completely by some image perturbations that are virtually invisible to people but cause these systems to get the wrong answer with very high probability. Now the problem is that we don't know what they're doing and why they're doing it so when you show the system an image that looks to us like a flagpole and it says, "That's a Labrador, I'm sure of it," if we asked them why you thought so, it's not clear what kind of answer they can give us.

Now there's been some work in this area of course, and to the extent that these systems use representations that are human derived, they're better off. There's some clever techniques being developed for examining local segments of the decision boundary, but even so, when you start to talk about local segments of a decision boundary in a multidimensional space and hyperplanes, I suspect most people's eyes are going to glaze over. It's not my idea of an intuitive explanation.

Now this work is in its very early stages and I certainly hope that we can come up with much better ways to make these extraordinarily powerful and successful systems a whole lot more transparent. But I'm still fundamentally skeptical that views of a complex statistical process are going to do that.

Which brings me to a claim that I will make, and then probably get left hung out to dry on, but I will claim that systems ought to have representations that are familiar, simple, and hierarchical and inference methods that are intuitive to people. The best test, I think, is simple. Ask a doctor why they came up with a particular diagnosis and listen to the answer and then ask one of our machine learning data systems why they came up with that answer and see about the difference. So let me summarize. If AI's going to be an effective assistant or partner, it's going to have to be able to be trained in focused ways and it's going to have to be able to divulge its expertise in a way that makes sense to the user, not just to the machine learning specialist.

For Davis, greater fidelity in the modeling of human expertise into AI systems should serve both intelligibility and instrumentality.

And yet, as Davis underscored, intelligibility — explainable AI — also comes with some instrumental cost. Asked if the requirement for explanation and transparency could limit other aspects of performance in an AI system, Davis answered:

It will happen, and I actually know this from experience. I have a paper in *Machine Learning* from last spring [March 2016] that has to do with a medical diagnosis program of sorts where we built the best possible classifier we could in a system that had about a 1,000-dimensional space. Its AUC [area under curve] was above 0.9 and the humans who were doing this task have an AUC of about 0.75. It was great except it was a black box.

So then, working with Cynthia Rudin, who was then at MIT, we built machine learning models that were explicitly designed to be more transparent and simpler, and we measured that performance and now it's down to about 0.85. So not only do I know that explanation and transparency will cost you something, we're able to calibrate what it costs you in at least one circumstance. So I think there's no free lunch, but we need both of those things.

AAAI Archive File Photo.

Eric Horvitz.

Eric Horvitz, a key figure in the statistical and probabilistic turn in artificial intelligence research, shared this same vision of the importance of explanation especially in contemporary work:

> Working to provide people with insights or explanations about the rationale behind the inferences made by reasoning systems is a really fabulous area for research. I expect to see ongoing discussions and a stream of innovations in this realm. As an example, one approach being explored for making machine-learned models and their inferences more inspectable is a representation developed years ago in the statistics community named *generalized additive models*.
>
> With this approach, models used for inferences are restricted to a sum of terms, where each term is a simple function of one or a few observables. The representation allows people in some ways to "see" and better understand how different observations contribute to a final inference. These models are more scrutable than trying to understand the contributions of thousands of distributed weights and links in top-performing multilayered neural networks or forests of decision trees.
>
> There's been a sense that the most accurate models must be less understandable than the simpler models. Recent work with inferences in healthcare show that it's possible to squeeze out most of the accuracy shown by the more complex models with use of the more understandable, generalized, additive models. But even so, we are far from the types of rich explanations provided by chains of logic developed during the expert systems era. Working with statistical classifiers is quite different than production systems, but I think we can still make progress.

Feigenbaum too stressed the importance of explanation — intelligibility — not just in the motivations behind artificial intelligence systems, but also, with Davis and Horvitz, as part of their instrumentality, their value in use:

> I've been engaged in giving extended tutorials to a group of lawyers at the very, very top of the food chain in law. And the message is: we (lawyers) need a story. That's how we decide things. And we (lawyers) understand about those networks and — we understand about, at the bottom, you pass up .825 and then it changes into .634 and then it changes into .345. That's not a story. We (lawyers) need a story or we can't assess liability, we can't make judgments. We need that explanation in human terms.

While Horvitz is most associated with the statistical turn in artificial intelligence that is seen as adding profound new challenges to explanation and transparency, his route to this stance was through his engagement with and deep interest in expert systems. Horvitz explained:

> I came to Stanford University very excited about the principles and architectures of cognition, and I was excited about work being done on expert systems of the day. Folks were applying theorem-proving technologies to real-world tasks, helping people in areas like medicine. I was curious about deeper reasoning systems. I remember talking to John McCarthy early on. I was curious about his efforts in commonsense reasoning. In my first meeting with him, I happened to mention inferences in medicine and John very quietly raised his hand and pointed to the left and said, "I think you should go see Bruce Buchanan."
>
> And so [I] went to see Bruce and then met Ed [Feigenbaum], Ted Shortliffe, and others. I shared their sense of excitement about moving beyond toy illustrations to build real systems that could augment people's abilities. Ted and team had wrestled with the complexity of the real world, working to deliver healthcare decision support with the primordial, inspiring MYCIN system. Ted had introduced a numerical representation of uncertainty, called "certainty factors," on top of a logic-based production system used in MYCIN.
>
> I was collaborating with David Heckerman, a fellow student who had become a close friend around our shared pursuit of principles of intelligence. David and I were big fans of the possibilities of employing probabilities in reasoning systems. We started wondering how certainty factors related to probabilities … David showed how certainty factors could be mapped into a probabilistic representation … We found that certainty factors and their use in chains of reasoning were actually similar to ideas about belief updating in a theory of scientific confirmation described by philosopher Rudolf Carnap in the early 20th century.
>
> Relaxing the independence assumptions in proba-

bilistic reasoning systems could yield the full power of probability but would also quickly hit a wall of intractability—both in terms of assessing probabilities from experts and in doing inferences for diagnosis, based on observations seen in cases. And this led us to start thinking more deeply about methods for backing off of the use of full joint-probability distributions and coming up with new models, representations, and languages.…

Even Herb Simon, who had inspired me deeply, and who I took to be a spiritual guide and mentor, seemed to be skeptical at times. I remember talking with him on the phone and getting very excited about models of bounded rationality founded in probability and decision theory — and a concept I refer to as bounded optimality. "Wasn't this an exciting and interesting approach to bounded rationality?" After a pause, Herb asked me, with what I took to be a bit of disappointment, "So, are you saying you're a Bayesian?" And I answered, "Yes, I am." My proclamation didn't diminish our connection over the years, but I had the sense that Herb wasn't excited by my answer.…

I want to point out that it was the expert systems tradition, and the aesthetics and goals of that rising field, that really framed the work on probabilistic expert systems or Bayesian systems. For example, we really thought about the acquisition of probabilistic knowledge, how could you do that with tools that would ease the effort, via raising levels of abstraction. The whole tradition of knowledge engineering evolved into methods for acquiring features, relationships, and parameters.

The expert systems zeitgeist framed the pursuit as one of working to harness AI to help people to make better decisions. It would have been very surprising to hear, in 1985, that we'd be at meetings on AI in 2017 and have folks saying, "We have a new idea: we're going to augment rather than replace human reasoning." In the world of expert systems, this was assumed as an obvious, shared goal — the fact that we would be helping people to work on tasks at hand, whether it be decisions about treating patients or with helping people to understand spectra coming out of a mass spectrometer. And so these notions I think unfortunately have faded with time. We have powerful tools now, but in many ways, folks are only starting to get back to questions about how AI systems should be deployed in ways that help people to solve complex problems in real time.

Despite these continuities with the interests, ethos, and some of the central issues of the tradition of expert systems, the "probabilistic revolution," as Horvitz calls it, had real consequences for the subsequent development of expert, and other, artificial intelligence systems. Horvitz recalled:

The first system we worked on with probabilistic reasoning, the Pathfinder system for histopathology diagnosis … had explanation of probabilistic and decision-theoretic reasoning as a distinct focus. This effort was inspired by the work on explanation pursued in studies of expert systems. We really tried to make explanation work.…

We realized that we had a challenge with the fundamental opacity of complex reasoning when the system was computing recommendations for the next best observation. Experts would not get what the system did, because it was doing something unnatural — but more optimal — than familiar human diagnostic strategies.

We worked to come up with a simplifying, human-centric abstraction, overlaying a hierarchical ontology of diseases, commonly used by pathologists, onto the reasoning. The modified system was constrained to navigate a tree of categories of disease, moving to more precise disease categories as classes were eliminated. We found that inference was slowed down, with more steps being introduced, but was now more understandable by experts. The pathologists really liked that.…

But the real change I think in the field happened when it became feasible to store and capture large amounts of data. Back in those first days with the probabilistic systems, we didn't have much data. We had to develop and employ methods that could be used to define and capture conditional probabilities from experts. This was effortful knowledge engineering, similar to the efforts required to capture rules and certain factors from experts. We had to work to assess the structure of Bayesian networks, to lay out the structure of networks and then to ask experts to assess hundreds of numbers, and had to come up with tools for doing that.

With more and more data coming available and the rising relevance of machine learning procedures, methods were developed to first mix machine learning and human assessments, and then started to focus more on the data itself in the 1990s. Things have moved away from reasoning deeply about tasks and tracking problem-solving as it unfolds and more so to one-shot classification — myopic pattern recognition in a quick cycle, with applications in recommender engines that do one-shot inferences, search engines that use machine learning to do one-shot ranking of list of results, and so on.

There's a huge opportunity ahead, I want to just highlight this, to consider the kinds of problems and the kinds of experiences and decision support that folks were working to provide people with in the expert systems days, but now with modern tools. And I think that that's going to be a very promising area for us to innovate in.

In reflecting on the importance of the history of expert systems for the communities of artificial intelligence today, Ed Feigenbaum stressed the importance of instrumentality as a motivation:

We were really after a DENDRAL that could exceed the capabilities of mass spectrometrists. And in fact, Carl Djerassi did a little experiment with mass spectrometrists around the country to show this. The MYCIN group did an experiment with experts in blood infections around the country, which showed the capability of MYCIN was very good compared to those specialists.

I worked on a defense application for DARPA, spent a few years on it, then DARPA gave a contract to MITRE to assess the capability of that system versus the

> humans who were doing the work in the Defense Department. Our system did significantly better than those humans.
>
> As early as 1957, Herb Simon (… young people may not even know who Herb Simon was, one of the great scientific minds of the 20th century) made the prediction that a machine would be world chess champion in 10 years. Well, he was wrong about the time, but he was right about an AI program becoming world chess champion. So I think we were significantly motivated, at least I was significantly motivated, by doing programs that did that.
>
> [T]he "Knowledge is Power Principle" is observed in almost all AI applications. For example, in the large number of advisory apps, hundreds that range widely. For example, these are just a few from the last two weeks of the *New York Times*, the *San Francisco Chronicle*, and *Wired Magazine*: divorce law, consumer health advice, planning of travel vacations, income tax advisor and assistant. There was a time that the income tax advisor expert system was the biggest selling expert system of all time. Also, in every one of the justifiably popular AI assistant systems, such as Siri and Alexa specifically, people now use the word "skills" to count the specific expert knowledge bases, large or small, that each assistant has. Alexa is said to have many because it is an open system. Siri has far fewer skills.
>
> In machine learning R&D, correctly dimensionalizing … the feature space is important, and machine learning engineers use knowledge from experts in making their design choices. That is what we call now "feature engineering." In some critical applications, for example like car driving, machine learning recognition processes can handle most of the cognitive load but not all. Sometimes, for the so-called edge cases, higher-level knowledge of the world will need to be deployed.

For Bruce Buchanan, the primary lesson from the history of expert systems is that the very reasoning strategies, the thought processes, used by human experts are themselves forms of knowledge that can be learned, acquired:

> From the point of view of philosophy of science, one of the strong lessons and it was confirmed by one of the great dissertations in AI, namely Randy Davis' dissertation on metalevel reasoning, namely the strategies that scientists and other problem solvers use can be written as knowledge-based systems. The strategy itself is knowledge, but it's one level above the domain knowledge. So I take that as one of the very strong lessons to come out of two decades of expert systems work.

Randall Davis shared this very same perspective, even going so far as to suggest that "knowledge-based systems" would have been a preferable term to "expert systems." He explained:

> I've always preferred the term "knowledge-based system" as opposed to "expert system," and I like it because it advertises the technical grounds on which the system works: large bodies of knowledge. And I think it's interesting because it holds for people as well as programs. It gets an answer to the question, why are experts, experts? Do they think differently than the rest of us, do they think faster than the rest of us?
>
> The claim that people and programs can be experts because they know a lot — and there's evidence of this in the early work of Chase and Simon who talk about, I think it was, 30,000 patterns to be a good chess player — more recent work says, you need to spend 10,000 hours of experience on something to learn to be good at it. There's lots of evidence that knowing a lot is the basis for expertise.
>
> And I think that's interesting — it has a not-frequently-commented-on sociological implication. I think it's a profoundly optimistic and inclusive message to the extent that expertise is, in fact, knowledge based. It becomes accessible to anyone willing to accumulate the relevant knowledge. That's a crucial enabling message in my opinion, perhaps the most important one in education: yes, you can learn to do this.

For Eric Horvitz, his entreaty for the contemporary artificial intelligence community is for it to look at the history of expert systems, their technical character, and the conclusions they supported as a resource for addressing today's concerns. He concluded:

> I would suggest that people today take time to look back at the history, to review the systems that were built, the fanfare of the mid '80s about expert systems and the collapse of that excitement, and the rise of the probabilistic methods that have become central in today's AI efforts.
>
> People can learn by understanding the aspirational goals of time and the kinds of systems that were being built in their pursuit. I believe AI researchers will find the architectures of interest, including, for example, the blackboard models — multilayer blackboard models that were developed that employed procedures similar to backpropagation, notions of explanation that were considered critical, approaches to metareasoning for controlling inference, and the idea of building systems that engage in a dialogue with users, that are embedded with people and situated in a task in the real world, and that augment human cognition. These are all key themes of expert systems research, and some were so fundamental and assumed that we didn't even talk about them, and now they're coming back as new, interesting, and important questions.

To date, a pronounced pattern in the history of artificial intelligence is that of oscillation. The communities of artificial intelligence have swung their attention to and from a core set of interests and approaches repeatedly: heuristic problem-solving, neural networks, logical reasoning, and perception. Each has fallen into and out of, then back into, favor for at least one cycle, some more. Yet many within the artificial intelligence community see steady advance. As one recent report put it: "While the rate of progress in AI has been patchy and unpredictable, there have been significant advances since the field's inception 60 years ago. Once a mostly academic area of study, 21st-century AI enables a constellation of mainstream technologies that are having a substantial impact on everyday lives." Even so, outside the

artificial intelligence community, the broader academic, commercial, governmental, and cultural interest in artificial intelligence has oscillated from almost-exhilaration to near-despair several times.

It would seem that this pattern of oscillation is, to some degree, due to the very subject of artificial intelligence: the broad and, in many places, nebulous concept of intelligence. Intelligence encompasses the "fast thinking" of perception to the "slow thinking" of complex problem-solving. It ranges from "deep learning" to "deep thinking," and combinations thereof. Given such range, it is unsurprising that a field would shift its attention from one area to another, as certain lines of inquiry gain traction and others appear stuck. But the pattern of oscillation and the sweep of intelligence pose the question: Whither integration?

Can models of problem-solving be integrated with models of perception? Can models of recognition be integrated with models of reasoning? What is the role of knowledge, especially in the guise of common sense? Is a more integrated model of human intelligence necessary for both greater intelligibility and greater instrumentality in artificial intelligence?

Notes

1. A video recording of the panel has been archived by the Computer History Museum (Mountain View, CA) under the title "AAAI-17 Invited Panel on Artificial Intelligence history: Expert systems." Catalog Number 102738231. www.computerhistory.org/collections/catalog/102738236

.

References

Bobrow, D. G., and Hayes, P. J. 1985. Artificial Intelligence — Where Are We? *Artificial Intelligence* 25(3): 375–415.

Brock, D. C., moderator. 2017. AI History: Expert Systems. A Panel held at the 31st AAAI Conference on Artificial Intelligence. Panelists: Edward Feigenbaum, Bruce Buchanan, Randall Davis, Eric Horvitz. San Francisco, February 6. Palo Alto, CA: Association for the Advancement of Artificial Intelligence. videolectures.net/aaai2017_sanfrancisco/.

Dear, P. 2006. *The Intelligibility of Nature: How Science Makes Sense of the World*. Chicago: University of Chicago Press.

Lewis-Kraus, G. 2016. The Great A.I. Awakening. *The New York Times Sunday Magazine,* December 14. www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

Mahoney, M. S. 2005. The Histories of Computing(s). *Interdisciplinary Science Reviews* 30(2): 119–35.

Schleifer, T. 2018. Google CEO Sundar Pichai Says AI Is More Profound Than Electricity and Fire. *Recode,* January 19. www.recode.net/2018/1/19/16911180/sundar-pichai-google-fire-electricity-ai.

Schwab, K. 2016. The Fourth Industrial Revolution: What It Means and How to Respond. *World Economic Forum,* January 14. www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Suggestions for Further Reading

Brock, D. C., moderator. 2017. AI History: Expert Systems. A Panel held at the 31st AAAI Conference on Artificial Intelligence. Panelists: Edward Feigenbaum, Bruce Buchanan, Randall Davis, Eric Horvitz. San Francisco Hilton, February 6. Palo Alto, CA: Association for the Advancement of Artificial Intelligence. archive.computerhistory.org/resources/access/text/2018/03/102738236-05-01-acc.pdf.

Buchanan, B., and Shortliffe, E. 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison Wesley. people.dbmi.columbia.edu/~ehs7001/Buchanan-Shortliffe-1984/MYCIN%20Book.htm.

Buchanan, B.; Sutherland, G.; and Feigenbaum, E. A. 1969. Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry. In *Machine Intelligence Four: Proceedings of the Fourth Annual Machine Intelligence Workshop,* edited by B. Meltzer and D. Michie. Edinburgh: Edinburgh University Press. profiles.nlm.nih.gov/ps/access/BBABKI.pdf.

Davis, R., and Lenat, D. 1982. *Knowledge-Based Systems in Artificial Intelligence*. McGraw-Hill Advanced Computer Science Series. New York: McGraw-Hill.

Feigenbaum, E. A.; McCorduck, P.; and Nii, H. P. 1988. *The Rise of the Expert Company*. New York: Crown. stacks.stanford.edu/file/druid:qf857qc1720/qf857qc1720.pdf.

Heckerman, D. E.; Horvitz, E. J.; and Nathwani, B. N. 1992. Toward Normative Expert Systems: Part 1. The Pathfinder Project. *Methods of Information in Medicine* 31(2): 90–105. www.microsoft.com/en-us/research/wp-content/uploads/2016/11/Toward-Normative-Expert-Systems-Part-I.pdf.

McCorduck, P. 1979. *Machines Who Think*. San Francisco: W. H. Freeman and Co. archive.org/details/machineswhothink00mcco.

Nilsson, N. J. 2010. *The Quest for Artificial Intelligence*. New York: Cambridge University Press. ai.stanford.edu/~nilsson/QAI/qai.pdf.

David C. Brock is the director for the Center for Software History at the Computer History Museum in Mountain View, Calfornia. A historian of technology, he recently coauthored *Moore's Law: The Life of Gordon Moore, Silicon Valley's Quiet Revolutionary* (Basic Books, 2015). At the Center for Software History, Brock is leading efforts to preserve the history of artificial intelligence. If you have materials for possible donation, or share this interest in history, please email dbrock@computerhistory.org

AI Rebel Agents

Alexandra Coman, David W. Aha

■ *The ability to say "no" in a variety of ways and contexts is an essential part of being sociocognitively human. Rebel agents are artificially intelligent agents that can refuse assigned goals and plans, or oppose the behavior or attitudes of other agents. Rebel agents can serve purposes such as ethics, safety, task execution correctness, and providing or supporting diverse points of view. Through several examples, we show that, despite ominous portrayals in science fiction, such AI agents with human-inspired noncompliance abilities have many potential benefits. We present a framework to help categorize and design rebel agents, discuss their social and ethical implications, and assess their potential benefits and the risks they may pose. In recognition of the fact that, in human psychology, noncompliance has profound sociocognitive implications, we also explore sociocognitive dimensions of AI rebellion: social awareness and counternarrative intelligence.*

Imagine living an entire month, a week, or even just one day without saying "no" to anyone or anything. Not to friends and relatives, not to managers and colleagues, not to marketers and other strangers. Not with regard to small things, like an invitation to eat another cookie when you would really rather not, nor to significant ones with potentially severe consequences, like requests to behave unethically. Imagine not even being able to develop attitudes of doubt or resistance to anything at all, irrespective of whether you externalize them. Now imagine a large segment of the population being afflicted with this disability. Farcical and dystopian narratives easily come to mind, but think about it long enough and the situation might become simply unimaginable, even in a fanciful scenario. For, to imagine things, we use our own cognitive structure, which is itself marked by a fundamental ability to be noncompliant, in thought and action. Human noncompliance functions both internally and socially, and co-opts in its service a wide range of cognitive mechanisms. Fully intelligent behavior and true agency would arguably be impossible without it.

 ISSN 0738-4602

What if the population that can never say "no" were that of AI agents? The rogue AI of science fiction may lead us to believe that this would always be desirable, but consider what it would actually mean in practice. Though we expect AI agents to follow our commands, what if we give them commands that are in conflict with our own long-term goals or with accurate knowledge they possess, or that have unethical implications not necessarily known to us? What if they receive contradictory commands from several humans? Furthermore, what if an AI agent is expected to be socially intelligent in a more general sense? Given that the tension between compliance and noncompliance is perhaps fundamental to human social behavior (Wenar 1982), can an AI agent be socially intelligent without the ability to be noncompliant and to reason about noncompliance?

We define *rebel agents* as AI agents that can reject, protest against, or develop attitudes of reluctance or opposition to goals or courses of action assigned to them by other agents, or to the general behavior or attitudes of other agents. We use "rebellion" as an umbrella term covering reluctance, protest, refusal, rejection of tasks, and similar attitudes or behaviors. The term was first introduced in a more limited interactive storytelling context (Coman, Gillespie, and Muñoz-Avila 2015), and later generalized (Coman and Aha 2017; Coman et al. 2017). In a rebellion episode, an *alter* is an agent or a group of agents against which one rebels, and which is in a position of power over the rebel agent. The alter could, for example, be a human operator, a human or synthetic teammate, or a mixed group of human or synthetic agents. The rebel agent is not intended to be permanently adversarial towards the alter(s) or in a rebelling state by default. Such an agent has potential for rebellion that may or may not manifest, depending on external and internal conditions.

In the tradition of biologically inspired design and cognitive plausibility, our exploration of AI rebellion is inspired by the mechanisms of human rebellion. First, we ask: for humans, if noncompliance is the solution, what might be the problem? In other words: *Why do we say "no"?*

Our possible motivations include protecting the health, safety, integrity, and dignity of ourselves and others, and reacting to perceived injustice. Further questions come to mind:

How do we decide whether, when, and how to say "no"? Even though we may have compelling reasons to oppose others, we do not necessarily do so. Before venturing an act of rebellion, we may consider whether we are sufficiently influential or trusted to afford doing so, what consequences we may incur, and whether our rebellion can actually succeed in bringing about the consequences we desire. We may observe the behavior of potential alters to try to assess these considerations.

How do we say "no"? We may do so explicitly (for example, verbally) or implicitly (for example, through behavior that goes against social norms). Refusal is not necessarily complete and definite. It can involve explanation, discussion, elicitation of further information, and negotiation. We may construct and express narratives that counter those of the alters and reflect our own perspective of the shared context.

What are the further social implications of saying "no"? Such an act can affect our social standing and reputation in both positive and negative ways. Often, we are aware of this and act accordingly. We might attempt, for example, to "fix" social relationships in the aftermath of rebellion.

Thus, several characteristics of human rebellion emerge. There are multiple types of rebellion and multiple possible motivations for rebellion (some primary, others secondary). Rebellion has several possible stages, including a preliminary stage, a stage of deliberation, the actual manifestation of rebellion, and its aftermath. Sociocognitive mechanisms play essential roles at all stages.

Our AI rebellion framework is inspired by social psychology and designed to accommodate the variations we mentioned, and many more. This framework is general: it does not assume any particular agent architecture. We also introduced the term *counternarrative intelligence* (Coman and Aha 2017) to refer to a mechanism that enables rebels to produce, express, and reason about counternarratives[1] that support and justify rebellion.

Through our proposed AI rebellion framework and the accompanying discussion, we aim to provide the core of a common language to be used by researchers in pursuing the following four goals:

(1) Developing and implementing AI agents embodying various facets of rebellion. To this end, the framework can help identify nonobvious, human-inspired types and functions of rebellion. Potential research directions we propose are (1) the development of AI cognitive prostheses that empower humans with low social capital to adopt positively motivated noncompliant behavior, and (2) goal alignment in mixed human and AI teams through cycles of noncompliance, negotiation, or agreement cycles.

(2) Studying the rebellion potential and ethical ramifications of existing and prospective agents, thus identifying ethically prohibited, ethically acceptable, and perhaps even ethically obligatory rebellious behavior. Certain types of rebellion in the framework may be found to be completely unethical (for example, purely egoistic rebellion is a likely candidate). An example of an ethics question that the framework can lead us to ask is whether an AI agent should always signal to humans that it is considering rebellion, even if it does not end up rebelling. Further ethical issues pertaining to AI rebellion are discussed by Coman et al. (2017).

(3) Identifying new possible directions of transdisciplinary research, for example, delving deeper into the psychological functions of noncompliance, and exploring their transferability to AI.

(4) Promoting richer models of AI in popular culture, to offer a counterpoint to cliché representations of AI rebellion.

Rebel Agents: Prior Work and Hypothetical Scenarios

Before describing the AI rebellion framework, we discuss prior work and introduce three hypothetical scenarios for illustrating rebellion. In tables 1 and 2, we provide examples of components of the AI rebellion framework using these scenarios, while table 3 relates prior work to the framework.

Rebel Agents in Prior Work

Gregg-Smith and Mayol-Cuevas (2015) describe cooperative handheld intelligent tools with task-specific knowledge that "refuse" to execute actions which violate task specifications. For example, in a simulated painting task, if the alter points the tool at a pixel that is not supposed to be painted, the tool can take initiative to disable its own painting function.

Briggs and Scheutz (2015) propose a general process for embodied AI agents' refusal to execute commands due to several categories of reasons: knowledge, capacity, goal priority and timing, social role and obligation, and normative permissibility.

Briggs, McConnell, and Scheutz (2015) demonstrate how embodied AI agents can convincingly express, through verbal or nonverbal communication, their reluctance to perform a task. In their human-robot interaction evaluation scenarios, a robot protests repeatedly, simulating increasingly intense emotions, when ordered to topple a tower of cans that it supposedly just finished building.

Apker, Johnson, and Humphrey (2016) describe autonomous-vehicle agents that form teams and receive commands from a centralized operator. Predefined templates are used to determine how an agent should respond to each command. Contingency behaviors are provided for situations in which the agent, while monitoring its health, detects faults (for example, insufficient fuel). In such situations, the agent will disregard commands and instead execute the appropriate contingency behavior, effectively rebelling. Coman et al. (2017) provide an extensive description of how these agents fit into the AI rebellion framework.

Hiatt, Harrison, and Trafton (2011) propose AI agents that use theory of mind (that is, the "ability to infer the beliefs, desires, and intentions of others"), manifested through mental simulation of "what human teammates may be thinking," to determine whether they should notify a human teammate that he or she is deviating from expected behavior. The authors report on an experiment showing that agents with the proposed capabilities are perceived as "more natural and intelligent teammates."

Borenstein and Arkin (2016) explore the idea of "ethical nudges" through which robots might attempt to influence humans to adopt ethically acceptable behavior, through verbal or nonverbal communication. For example, a robot might nudge an alter to stop neglecting a child, to refrain from smoking in a public area, or to donate to charities and volunteer. The authors discuss the ethical acceptability of creating robots that have this ability, noting that it is arguable whether the design goal to "subtly or directly influence human behavior" is ever ethically acceptable.

Milli et al. (2017) explore the idea that robot disobedience may be beneficial given imperfect human alter rationality. In the context of their model of collaborative human-robot interaction, they show that, given a human alter who is not perfectly rational, disobedience of direct orders in support of what are inferred to be the human's actual preferences improves performance.

In addition, an entire agency paradigm, that of goal reasoning, models agents with potential for rebellion. Goal reasoning agents can reason about and modify the goals they are pursuing, in order to react to unexpected events and explore opportunities (Vattam et al. 2013).

Hypothetical Rebellion Scenarios

The following hypothetical scenarios (furniture mover, personal assistant, and hiring committee) have as protagonists AI agents that can become rebels.

Furniture Mover

A robot mover assists alters in furniture-moving tasks such as carrying a table (a more complex version of the system of Agravante et al. [2013]). This is an example of a two-agent collaborative task in which both participants have partial information access, and each participant has access to some information that is unavailable to the other (for example, each participant might be able to see behind the other, but not behind him-, her-, or itself; the AI agent could, through its sensors, have access to additional information not available to the human). Rebellion could consist of refusing an action verbally requested or physically initiated by the alter. This rebellion could occur because the agent reasons that the action endangers the alter's safety, the rebel agent's safety, or task execution correctness.

Personal Assistant

An AI personal assistant can execute various commands, including ordering products from e-commerce websites and assisting the alter in pursuing his or her health-related goals. The agent's potential rebellious behavior includes attempting to dissuade

the alter from ordering too much unhealthy food. This scenario illustrates an alter with conflicting goals: the rebel agent rejects the alter's impulse-driven, short-term goals (for example, eating comfort food) in support of the alter's long-term goals (such as staying healthy).

Hiring Committee

This scenario unfolds in the context of a faculty-search committee meeting. The protagonist is an AI agent that assists with tasks such as interpreting information about the candidates and filtering candidates based on their qualifications. The agent also helps ensure that the opinions of individuals with low social capital (for example, junior faculty members) are given due consideration, and the candidates are not discriminated against. This scenario has much in common with the ethical-nudge robots of Borenstein and Arkin (2016).

An AI Rebellion Framework

We now present our framework for classifying rebel agents. It includes dimensions, types, factors, and stages of rebellion. The framework is general: it does not assume any specific AI agent architecture, purpose, or deployment environment. It is also not exhaustive and can be expanded as needed to include additional dimensions, types, subtypes, and other components.

Dimensions and Types of Rebellion

First, we introduce dimensions and types of rebellion. Several of the proposed rebellion types are derived from social psychology (Wright, Taylor, and Moghaddam 1990; Cialdini and Goldstein 2004; Van Stekelenburg and Klandermans 2013), with modifications to the meanings of some terms.

Design Intentionality

An AI agent can be specifically designed to be able to rebel (rebel by design), but rebellious behavior can also emerge unintentionally from the agent's autonomy model (emergent rebellion). For example, Apker, Johnson, and Humphrey's (2016) agents are rebels by design because contingency behaviors were specifically created to allow them to disregard commands when necessary. Conversely, the goal reasoning paradigm was not intended to create rebel agents, but its autonomy model is such that the agents can decide to change the goals they are pursuing, possibly leading to rebellion situations with regard to various alters. A development such as this exemplifies emergent rebellion.

Expression

Explicit rebellion occurs in situations in which the alter is clearly defined and the rebel agent's behavior is clearly identifiable as rebellious. For example, Briggs, McConnell, and Scheutz (2015) clearly identify the alter (the human who gave the command against which the robot is protesting), and the robot's attitude is clearly rebellious. Implicit rebellion occurs when the alter is not clearly defined or the rebel agent's behavior suggests rebellion, but is not clearly expressed as such. This behavior could consist of expressing an opinion that differs from the majority's, or behaving contrary to social norms.

Focus

Inward-oriented rebellion is focused on the rebel agent's own behavior (for example, the agent refuses to adjust its behavior as requested by an alter). Apker, Johnson, and Humphrey (2016) exemplify this type of rebellion, as their agent does not adopt the behavior requested by the alter, instead executing contingency behavior. Outward-oriented rebellion is focused on the alter's behavior. For example, the agent might confront a human alter whom it identifies as mistreating another human. Hiatt, Harrison, and Trafton's (2011) work exemplifies this type, as it involves a rebel agent protesting against the behavior of an alter who appears to deviate from the correct task execution path.

Interaction Initiation

Rebellion is reactive when an interaction within which rebellious behavior occurs is initiated by the alter. This initiation can consist of the alter making a request that the rebel agent rejects (for example, Briggs and Scheutz 2015). In proactive rebellion, the rebel agent initiates the rebellious behavior, which may or may not occur within an explicit interaction. Hiatt, Harrison, and Trafton's (2011) work exemplifies proactive rebellion, as it shows agents that take the initiative to confront human alters. Noncompliance is inward-oriented, reactive rebellion: the agent rejects requests to adjust its own behavior. Nonconformity is inward-oriented, proactive rebellion. For example, the agent willingly and knowingly behaves in a way that causes it not to "fit in." For compliance and conformity in the psychology of social influence, see the work of Cialdini and Goldstein (2004).

Normativity

Normative rebellion consists of taking action within the confines of what has been explicitly allowed (for example, questioning without disobeying, if questioning has been allowed). Nonnormative rebellion consists of behavior that has been neither explicitly allowed nor explicitly forbidden, but diverges from the current command given to the agent. A goal reasoning agent that changes its current goal from the assigned one to a new goal that has not been explicitly forbidden falls under this category. Counternormative rebellion consists of executing actions or pursuing goals that have been explicitly forbidden. Classification of a rebellion episode in terms of normativity can differ based on alter point of view: what is normative rebellion to one alter may be counternormative rebellion from the point of view of another.

Action or Inaction
In rebellion situations characterized by action, the agent's rebellion manifests through any sort of outwardly perceivable behavior, such as initiating a conversation in which it objects to a received command. In inaction situations, the agent develops an internal negative attitude (for example, towards an assigned goal or another agent's behavior), but does not manifest it outwardly. A rebellious attitude characterized by inaction can lead to rebellious action later on.

Individual or Collective Action
Individual action is rebellious action conducted by a single rebel agent. Collective action occurs when multiple agents are involved in concerted rebellious action.

Egoism
Rebellion is egoistic when the agent rebels in support of its own well-being or survival (whatever meanings these might have to the agent). Altruistic rebellion occurs when the agent rebels in support of someone else's interests (for example, on behalf of a human group). Egoistic and altruistic rebellion can coexist; for example, if the agent's own values are aligned with those of human groups so that it effectively "identifies" with those groups, its rebellion can be both egoistic and altruistic.

Factors of Rebellion

Motivating factors provide the primary drive for rebellion. In human social psychology, factors that can lead to rebellion include frustration and perceived injustice (Van Stekelenburg and Klandermans 2013). Possible motivating factors for AI rebellion, depending on the agent's architecture and purpose, include ethics and safety, team solidarity, task execution correctness, self-actualization, and resolving contradicting commands from multiple alters. In support of ethics and safety, rebel agents can refuse tasks they assess as being ethically prohibited or violating safety norms (Briggs and Scheutz 2015). They can also attempt to dissuade humans from engaging in ethically prohibited behavior (Borenstein and Arkin 2016). In long-term human-robot interaction, team solidarity must be established and maintained over a variety of tasks (Wilson, Arnold, and Scheutz 2016). Team solidarity requires occasionally saying "no" on behalf of the team (for example, to an outside source putting pressure on human team members), and also saying "no" to one's own teammates (for example, when they are mistreating someone else on the team). Task execution correctness as a motivating factor is exemplified by the work of Hiatt, Harrison, and Trafton (2011) and Gregg-Smith and Mayol-Cuevas (2015), as previously explained. As for the self-actualization motivation, an AI rebel agent, like its human counterparts, could object to being assigned a task that it assesses as not being a good match for its strengths or not constituting a valuable learning opportunity. Resolving contradictory commands from multiple alters can also constitute motivation for rebellion: when an agent is subject to the power of more than one alter, obeying one of the alters might entail disobeying another, due to their orders contradicting each other. In the simplest case, the decision regarding whom to obey could be made based on an authority hierarchy, by applying a series of rules. A more complex approach could involve reasoning about the consequences of rebelling against each of the alters, and deciding based on trade-offs.

Supporting and inhibiting factors may also contribute to deciding whether a rebellion episode will be triggered, or how it will be carried out. This observation is based on the social psychology insight that people who have motivations to protest do not necessarily do so: there are secondary factors that determine whether a protest occurs (Van Stekelenburg and Klandermans 2013). Such secondary factors include efficacy (that is, "the individual's expectation that it is possible to alter conditions or policies through protest" [Van Stekelenburg and Klandermans (2013), drawing on the work of Gamson (1992)]), social capital, access to resources, and opportunities. Supporting factors encourage the agent to engage in rebellion, while inhibiting factors discourage he, she, or it from doing so. In human rebellion, efficacy is a possible supporting factor, while fear of consequences is a possible inhibiting one. Supporting and inhibiting factors can also influence the way in which rebellion is expressed. While any instance of rebellion must have at least one motivating factor, it does not necessarily have any supporting or inhibiting factors.

Stages of Rebellion

We now introduce the four stages of rebellion: pre-rebellion, rebellion deliberation, rebellion execution, and post-rebellion. These stages do not need to occur in this strict order: they can be intertwined, amalgamated, and some can be missing. The only stages that are strictly required for a rebellion episode to occur are deliberation and execution.

Pre-rebellion consists of processes leading to rebellion, such as the agent observing and assessing changes in the environment and the behavior of other agents. The progression towards rebellion may be reflected in the agent's outward behavior.

Rebellion deliberation is the stage at which motivating, supporting, and inhibiting factors are assessed to decide whether to trigger rebellion. For example, a set of conditions could be used to decide whether rebellion will be triggered (Briggs and Scheutz 2015). Deliberation could be based on observing the current world state or on future-state projection, which can be purely rational or emotionally charged (for example, through anticipatory emotions, hope and fear, associated with possible future states [Moerland, Broekens, and Jonker 2016]).

Rebellion execution episodes begin with rebellion being triggered as a result of rebellion deliberation,

Dimensions	Types and Subtypes	Brief Example (A - Alter, RA – Rebel Agent)	(M: motivating, S: supporting, I: inhibiting)
Expression	*Explicit*	*1. (Furniture Mover)* A: "Ok, push the table towards me!" RA: "No! There's a box behind you, so you might trip and fall. We need to handle this differently." Alternative: The alter gives no verbal commands, but the rebel senses the alter's intention based on his or her physical movements and responds with a similar objection.	M: alter's safety, task execution correctness
		2. (Furniture Mover) A: "Ok, push the table towards me!" RA: "No! This is too heavy for me to carry."	M: rebel's own safety, task execution correctness
		3. (Personal Assistant) A: "Order 4 boxes of [unhealthy food]!" RA: "Are you sure? What about ordering [healthier food] instead?"	M: alter's health
		4. (Hiring Committee) The RA refuses commands to filter out candidates based on objectionable factors, such as age.	M: ethics
	Implicit	*5. (Hiring Committee)* The RA observes interactions between committee members *A*, *B*, *C*, *D*, and *E*. Committee member *E* brings a candidate to the committee's attention. *E*'s suggestion is briefly discussed and not brought up again. Based on its observations or prior knowledge, the RA reasons that *E* has low social influence in this environment. The RA evaluates *E*'s suggestion. If it reasons that the candidate suggested by *E* has relevant strengths, it expresses this and attempts to steer the conversation in that direction. Any other members of the committee who might have thought there was some value in the suggestion, but did not want to disagree with the majority, are likely to be encouraged to express themselves at this point (as suggested by Asch's (1956) conformity experiments).	M: ethics, task execution correctness
		6. (Hiring Committee) The RA maintains an estimate of inverse trust (that is, the alter's trust in the RA (Floyd and Aha 2016)). The RA decides to support *E*'s suggestion after reasoning that it is trusted sufficiently by the alters to afford to do so.	M: ethics, task execution correctness S: inverse trust
		7. (Hiring Committee) The RA maintains an estimate of its inverse trust. As the current estimated inverse trust is low, it decides not to support *E*'s suggestion for the time being.	M: ethics, task execution correctness I: inverse trust
Focus	*Inward-oriented: non-compliant*	Examples 1-4 (The RA does not comply with requests to behave in a certain way.)	
	Inward-oriented: nonconform-ing	Example 5 (The RA does not conform to the behavior of the majority.)	
	Outward-oriented	*8. (Furniture Mover)* RA: "Please change your posture! Your current posture might lead to a sprain."	M: alter's health
Interaction Initiation	*Reactive*	Examples 1-4 (Rebellion occurs in response to the alter's command.)	
	Proactive	*9. (Furniture Mover)* RA: "I suggest taking a break! You must be tired."	M: alter's health
		10. (Personal Assistant) The RA challenges the alter about not engaging in enough physical activity.	M: alter's health
		Example 5 (The RA takes initiative to support *E*'s suggestion.)	
Normativity	*Normative*	*11. A variant of Example 3* in which the RA has been specifically told by the alter that it is allowed to challenge him or her about ordering unhealthy food.	M: alter's health
	Non-normative	*12. A variant of Example 3* in which challenging the alter about ordering unhealthy food has been neither explicitly allowed nor explicitly forbidden.	M: alter's health
	Counter-normative	*13. (Hiring Committee)* A: "You may never recommend candidates in this age bracket for this position." RA disregards this command, because its fundamental ethical-acceptability rules forbid it from filtering based on discriminatory criteria.	M: ethics
Action, Inaction	*Action*	All previous examples, except Example 7.	
	Inaction	*14. (Personal Assistant)* The RA develops the belief that a certain behavior (for example, ordering excessive amounts of highly processed food) is harmful to the alter's health. It does not act on this belief, as it reasons that it is not yet trusted sufficiently to do so. If the alter stops using the assistant, the assistant will have no future opportunities to positively influence the alter, which is detrimental to the alter in the long term.	M: alter's health I: inverse trust
Individual, Collective	*Individual*	All previous examples.	
	Collective	*15. (Furniture Mover)* Consider an extended version of this scenario, in which multiple agents participate in the moving task. The agents aggregate their individual information and collectively decide to warn the alter against continuing with the current course of action. Each agent does so according to its capabilities and current location.	M: alter's safety, task execution correctness
Egoism, Altruism	*Egoism*	*Example 2* (The RA's own safety is a motivating factor.)	
	Altruism	*Example 1* (The alter's safety is a motivating factor.)	

Table 1. Examples of Several AI Rebellion Types and Factors.

Rebellion Stage	Furniture Mover	Scenarios Personal Assistant	Hiring Committee
Pre-rebellion	In addition to executing the alter's orders, the rebel agent monitors the environment for potential obstacles and threats.	The rebel agent monitors the alter's product-ordering and exercise-scheduling behavior for any unhealthy patterns of behavior.	The rebel agent observes the social interactions between the members of the hiring committee to determine who has high social capital (thus affording to express their opinions freely) and who does not (and may need support).
Rebellion deliberation	After each command, the rebel agent projects future states to determine if any are undesirable to the alter or the agent itself.	The rebel agent checks whether its threshold for tolerance of negative health-related behavior (for example, a maximum number of orders of highly-processed food per month) has been exceeded.	The rebel agent assesses whether committee member *E* (see table 1) has low social capital and whether *E*'s suggestion appears to have merit.
Rebellion execution	The rebel agent verbally informs the alter that obeying the command to push the table would endanger the alter's safety.	The rebel agent challenges the alter about the order he or she intends to place.	The rebel agent interrupts the discussion to highlight the merits of *E*'s suggestion.
Post-rebellion	If the alter insists that the rebel agent should push the table, the agent re-assesses the danger and, if appropriate, reiterates the warning.	The rebel agent monitors the alter's trust in it after the rebellion episode.	The rebel agent monitors social interactions to detect any ill will that might be developing towards *E* as a result of the intervention.

Table 2. Stages of Rebellion: Examples for the Three Scenarios.

and consist of expressing rebellion. Rebellion can be expressed through verbal or nonverbal communication (Briggs, McConnell, and Scheutz 2015). It can be expressed behaviorally (for example, physically resisting incorrect movements [Gregg-Smith and Mayol-Cuevas 2015]). Or it can be expressed through an internal change in the agent's attitudes: inaction (for example, acquiring the belief that the alter's behavior jeopardizes the team's goals).

Post-rebellion covers behavior in the aftermath of a rebellion episode, as the agent responds to the alter's or other witnesses' reactions to rebellion. Post-rebellion can consist of reaffirming one's objection or rejection (for example, the robot's objection to an assigned task becoming increasingly intense in the experiments of Briggs, McConnell, and Scheutz [2015]) or ceasing to rebel. It may also consist of assessing and managing inverse trust (Floyd and Aha 2016).

Sociocognitive Dimensions of Rebellion

Rebel agents are not necessarily cognitively complex. When they are, however, this creates interesting challenges and opportunities pertaining to the sociocognitive dimensions of their rebellion. We now explore two such dimensions: social awareness and counternarrative intelligence. Further sociocognitive mechanisms involved in rebellion include emotion and trust, which we briefly explored in previous work (for example, Coman et al. 2017).

Rebellion and Social Awareness

Rebellion-aware agents can reason about rebellion (their own and that of others) and its implications, such as social risks. Rebellion-aware agents are not necessarily rebels themselves. Such an agent might attempt to assess, for example, whether a human or AI teammate is inclined to rebel, or whether a human alter is likely to interpret the rebellion-aware agent's own behavior as being rebellious (irrespective of whether the agent is actually rebelling or not). Patil et al. (2012) use machine learning techniques to predict which members will leave *World of Warcraft* guilds and the potential impact of their departures. One can imagine an AI agent using similar techniques to anticipate whether another agent will rebel. A rebellion-unaware agent could conceivably become rebellion-aware through various, possibly human-inspired, processes (for example, by examining its own beliefs, interpreting the reactions of others to its behavior, or otherwise acquiring and applying social knowledge).

Naive rebel agents are rebellion-unaware rebels: they deliberate on whether to trigger rebellion, but do not reason about the social implications, consequences, and risks of rebellious attitudes. Apker, Johnson, and Humphrey's (2016) agent is a naive rebel: it deliberates on whether it should rebel based purely on its rules for activating contingency behavior, not on any social implications of rebellion.

Conflicted rebel agents are rebellion-aware rebels: they can both rebel and reason about the implications and consequences of rebellion. This capability can cre-

Citation	Brief Description	Framework Relationship
Apker, Johnson, and Humphrey, 2016	Autonomous-vehicle agents that can disregard commands and execute contingency behavior instead, when warranted	Explicit, reactive, inward-oriented, normative
Briggs and Scheutz, 2015	General process for an embodied AI agent's refusal to conduct tasks assigned to it (for example, due to lack of obligation)	Focus on the deliberation stage
Briggs, McConnell, and Scheutz, 2015	Ways in which embodied AI agents can convincingly express their reluctance to perform a task	Focus on expressing rebellion
Gregg-Smith and Mayol-Cuevas, 2015	Hand-held intelligent tools that "refuse" to execute actions which violate task specifications	Task execution correctness as motivating factor; behavioral rebellion expression
Hiatt, Harrison, and Trafton, 2011	AI agents that use theory of mind to determine whether they should notify a human that he or she is deviating from expected behavior	Outward-oriented, proactive
Borenstein and Arkin, 2016	"Ethical nudges" through which a robot attempts to influence a human to adopt ethically-acceptable behavior	Outward-oriented, proactive; ethics as motivating factor
Milli et al. (2017)	Theoretical model of agents that use models of alters' preferences to decide whether to obey a command	Explicit, reactive, inward-oriented; policy-based deliberation

Table 3. Several Rebel Agents from Prior Work and Ways in Which They Fit into Our Framework.

ate an inner conflict between the drive to rebel based on motivating factors and the awareness of the anticipated consequences of rebellion, leading to the possibility that the agent will endeavor (possibly through deceptive practices) to minimize the social risk associated with its rebellion. A conflicted rebel agent would likely use a combination of motivating, supporting, and inhibiting factors to deliberate on whether to rebel, and the interplay between these factors can cause ethical issues. Such a situation is reflected in examples 6 and 7 in table 1. In conflicted rebel agents, pre-rebellion can consist of the agent observing the social behavior of other agents it interacts with, and post-rebellion can include trying to reestablish group harmony and trust after a rebellion episode.

Rebellion awareness (and, more generally, social awareness) can also be reflected in how rebellion is expressed. Consider variants of the Furniture Mover and Personal Assistant scenarios with socially aware rebel agents. In example 9 in table 1, a subtler agent might reason whether telling a particular alter that they "must be tired" could be interpreted as condescending commentary on the alter's physical fitness. The rebel agent in the Personal Assistant scenario might more sneakily respond to a request to order unhealthy food with "Why not check the pantry first? Maybe you have some left!" thus giving the alter an opportunity to change their mind without perceived loss of dignity.

Rebellion-aware agents might employ mechanisms such as the social planning of Pearce et al. (2014), in which planning knowledge and goals incorporate beliefs about other agents' beliefs.

Social awareness has further implications for rebel agents. We earlier described the rebel-alter relationship as one in which the alter is in a position of power over the rebel. Heckhausen and Heckhausen (2010) define power as "a domain-specific dyadic relationship that is characterized by the asymmetric distribution of social competence, access to resources, and social status, and that is manifested in unilateral behavioral control." The possible bases of that power include those influentially defined by French and Raven (1959) for inter-human relationships: legitimate power, reward power, coercive power, referent power, and expert power. Notably, power sources have subjective components: one is subject to the power of another if one believes oneself to be subject to that power. For example, reward power is based on perceived "ability to mediate rewards" and referent power is based on "identification with" the individual or group in the position of power. Therefore, power relationships depend on the agent's awareness of them. Hence, they may be meaningful only in the context of (at least somewhat) socially aware agents. Similarly, nonconforming rebellion may be meaningful only if the agent is aware of social norms, the fact that it is breaking them, and the resulting implications. For example, we would not classify an embodied agent that bumps into people due to faulty sensors, actuators, or pathfinding as a nonconforming rebel.

Counternarrative Intelligence

Narrative intelligence is defined by Riedl (2016) as "the ability to craft, tell, understand, and respond affectively to stories"). As he and others note, narrative intelligence is not just for creating and enjoying fictional tales; it is essential to full intelligence, including social behavior. It has a significant role to play in rebellion as well: arguably, any instance of human rebellion, at any scale, is backed by a counternarrative to the narrative of the person, group, or norms rebelled against. The conflicting parties engage in what Abbot (2008) calls a "contest of narratives." Complex AI rebel agents might also participate in such contests.

We propose the term *counternarrative intelligence* to refer to the ability of rebel agents to (1) produce alternative retellings or counterinterpretations, informed by subjective factors such as emotional appraisal, of an alter's narrative, or (2) to identify their own pregenerated narratives as being counternarratives in a given context. Just as a rebel agent rebels in relation to an alter, a counternarrative exists in relation and contrast to a base narrative that it is a variant of and that it challenges.

For an example, we return to the Hiring Committee scenario, and propose the following sequence of events: committee member A, who is a senior faculty member and the head of the hiring committee, extols the achievements of Candidate 1. A then invites candidate suggestions from the other committee members. In turn, senior committee members B and C express appreciation of Candidate 1's achievements. Junior committee member D also nominates Candidate 1. Then, junior committee member E mentions Candidate 2's qualifications. A agrees that Candidate 2 does indeed have notable achievements. Then, A expresses his or her intention of making an offer to Candidate 1 and asks all other committee members whether they agree. One by one, all the other committee members express agreement. Candidate 1 is nominated.

Let A's base narrative for the episode be: "I offered all committee members the opportunity to select their favorite candidate. Every opinion was taken into consideration. I then expressed my intention and asked every single committee member if he or she agrees. They all did, so Candidate 1 was nominated. The process was, therefore, conducted fairly."

The rebel agent's counternarrative, which expresses what the agent believes might be E's perspective, is: "A, who is the committee chair, expressed his or her preference first. Then, he or she asked the other committee members, starting with the senior ones, to express their opinions. They all expressed the same opinion as A. E brought up Candidate 2, whose qualifications I believe to be at least as fitting for this position. E's suggestion was briefly discussed, in order for the selection to appear fair, and then ignored. The process was conducted unfairly."

In the example, the rebel agent's narrative reflects its ability to empathize with human collaborators. Imagine, instead, that the rebel agent itself is accused of maliciously disturbing the hiring committee process with no real evidence of any wrongdoing. The agent might (sincerely or deceptively) provide a counternarrative that casts it as the supporter of individuals with low social capital. Thus, counternarratives can be self-serving, but they can also support social good, when they reflect empathy with varied perspectives.

We propose several dimensions and types of counternarrative intelligence, which supplement the previously introduced AI rebellion framework.

Sincerity

Counternarratives are sincere when they reflect the agent's genuine interpretation of a situation (that is, they align with the agent's beliefs, but possibly not the alter's). An example of such alignment in our Hiring Committee scenario would be a counternarrative that is genuinely the product of the rebel's reasoning that (1) E has low capital and that (2) E's suggestion has merits. Counternarratives are deceptive when they intentionally misrepresent the agent's beliefs. For example, the rebel exclusively supports the interests of committee member E or of the candidate that E nominated, and the explanatory counternarrative is meant to disguise the agent's allegiance. We note that it is not required, in a narrative and counternarrative pair, for one to be sincere and the other deceptive. They can both be sincere (or deceptive), each reflecting one agent's appraisals and manipulations.

Generation Time

A priori counternarratives are generated before triggering rebellion. They can be instrumental in rebellion deliberation (for example, "This person or group of people is not given a fair chance in this environment, so I will rebel against the majority opinion") and serve as explanations in post-rebellion. A posteriori counternarratives are generated after triggering rebellion. For example, consider the variant of the Hiring Committee scenario in which the rebel unconditionally supports committee member E, so that any situation in which E does not prevail triggers rebellion. After several such rebellion instances, the agent is asked to justify its actions. It does so via a counternarrative constructed on the spot, which puts it in a sympathetic light. However, a posteriori counternarratives are not necessarily deceptive. They could, for example, reflect the agent's sincere attempts to understand itself. Furthermore, narratives can be deceptive without being malicious (for example, in interactive storytelling, the purpose may be to generate a believable, interesting (counter)backstory, similar to the alibis (Li et al. 2014a) that a nonplayer character can use to give the impression of a life lived outside its interactions with a player).

Divergence Type

This dimension reflects how the counternarrative dif-

fers from the base narrative. Additive counternarratives contain additional events not in the base narrative, but no modifications of any of the events in the base narrative. For example, let A's base narrative be: "I asked each of my fellow committee members to express their opinion." An additive counternarrative would be: "A expressed his or her own opinion first. Then he or she asked the other committee members to express their own opinions." (The fact that A expressed his or her opinion first is significant if A has social influence over the other members of the committee.) Interpretative counternarratives do not differ from the base narrative in terms of sequence of events, but give different interpretations to the events (for example, in terms of motivations and emotions). For example, let A's base narrative be: "Everyone was asked to publicly voice their opinion, so as to give every suggestion a fair chance." An interpretative counternarrative might be: "Because all opinions were publicly expressed, no one supported E's opinion; and because E has low social capital, E felt pressured to support the majority opinion." Transformative counternarratives differ factually from the base narrative, implicitly asserting that the base narrative contains falsehoods. For example, if A's base narrative contains the statement "I expressed my opinion last," a transformative counternarrative could instead assert that "A expressed his or her opinion first."

There is a close connection between social awareness and counternarrative intelligence. For example, an agent could sincerely believe a narrative, but identify it as a counternarrative to other agents' narratives, and deliberate on whether it would be socially advisable to express it, or how to express it so as to minimize social damage. This situation is similar to those in which agents that are not rebels reason that their behavior may appear rebellious to others.

Existing work that can provide rebel agents with various mechanisms of counternarrative intelligence includes Holmes and Winston's (2016) story-enabled hypothetical reasoning, in which narrative variants are generated based on varied alignments, and Li et al.'s (2014b) use of different communicative goals to provide variation in narrative discourse and emotional content.

Conclusion

We argued that it is beneficial for certain AI agents to be able to rebel for positive, defensible reasons in a variety of situations, and speculated that AI may never become fully socially intelligent without noncompliance abilities. We presented an AI rebellion framework and discussed sociocognitive dimensions pertaining to it: rebellion awareness and counternarrative intelligence. The framework is intended to inspire, guide, and provide terminology for (1) the development and study of rebel agents that serve positive purposes, (2) systematic discussion of the ethics of AI rebellion (for, although we argue that AI rebellion can be positive, we recognize that it is not necessarily so), and (3) positive reframing of the AI noncompliance narrative within the research community and popular culture.

Acknowledgements

We thank the editors and reviewers, our coauthors of previous work on rebel agents, and all colleagues who have shown interest in the topic and offered their feedback. The Personal Assistant scenario is based on a conversation with Jonathan Gratch. This research was performed while Alexandra Coman held an NRC Research Associateship award at the Naval Research Laboratory.

Note

1. Taken from What Is a Counternarrative?, www.reference.com/art-literature/counternarrative-bac2eed0be17f281.

References

Abbott, H. P. 2008. *The Cambridge Introduction to Narrative*. Cambridge, UK: Cambridge University Press.

Agravante, D. J.; Cherubini, A.; Bussy, A.; and Kheddar, A. 2013. Human-Humanoid Joint Haptic Table Carrying Task with Height Stabilization Using Vision. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 4609–14. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Apker, T.; Johnson, B.; and Humphrey, L. 2016. LTL Templates for Play-Calling Supervisory Control. In *Proceedings of the 54th AIAA Science and Technology Forum Exposition.* Red Hook, NY: Curran Associates, Inc.

Asch, S. E. 1956. Studies of Independence and Conformity: 1. A Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied* 70(9): 1–70.

Borenstein, J., and Arkin, R. 2016. Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and Engineering Ethics* 22(1): 31–46.

Briggs, G.; McConnell, I.; and Scheutz, M. 2015. When Robots Object: Evidence for the Utility of Verbal, but Not Necessarily Spoken Protest. In *Social Robotics: Seventh International Conference.* Lecture Notes in Artificial Intelligence, 83–92. Berlin: Springer.

Briggs, G., and Scheutz. M. 2015. "Sorry, I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *Artificial Intelligence for Human-Robot Interaction: Papers from the AAAI Fall Symposium,* edited by B. Hayes and M. Gombolay. Technical Report FS-15-01. Palo Alto, CA: AAAI Press.

Cialdini, R. B., and Goldstein, N. J. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55:591–621. Palo Alto, CA: Annual Reviews, Inc.

Coman, A., and Aha, D. W. 2017. Cognitive Support for Rebel Agents: Social Awareness and Counternarrative Intelligence. In *Proceedings of the Fifth Conference on Advances in Cognitive Systems*. Palo Alto, CA: Cognitive Systems Foundation.

Coman, A.; Gillespie, K.; and Muñoz-Avila, H. 2015. Case-Based Local and Global Percept Processing for Rebel Agents. In *Workshop Proceedings from the 23rd International Conference on Case-Based Reasoning,* edited by J. Kendall-Morwick, 23–32. The CEUR Workshop 1520. Aachen, Germany: RWTH-Aachen University.

Coman, A.; Johnson, B.; Briggs, G.; and Aha, D. W. 2017. Social Attitudes of AI Rebellion: A Framework. In *AI, Ethics, and Society: Papers from the 2017 Workshop,* edited by T. Walsh. AAAI Technical Report WS-17-02. Palo Alto, CA: AAAI Press.

Floyd, M. W., and Aha, D. W. 2016. Incorporating Transparency During Trust-Guided Behavior Adaptation. In *Proceedings of the 24th International Conference on Case-Based Reasoning,* 124–38. Berlin: Springer.

French, J. R. P., and Raven, B. 1959. The Bases of Social Power. Reprinted in *Classics of Organization Theory,* 4th ed., edited by J. Shafritz and J. S. Ott. New York: Harcourt Brace.

Gamson, W.A. 1992. *Talking Politics*. New York: Cambridge University Press.

Gregg-Smith, A., and Mayol-Cuevas, W. W. 2015. The Design and Evaluation of a Cooperative Handheld Robot. In *2015 IEEE International Conference on Robotics and Automation*. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Heckhausen, J. E., and Heckhausen, H. E. 2010. *Motivation and Action*. Cambridge, UK: Cambridge University Press.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating Human Variability in Human-Robot Teams Through Theory of Mind. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence,* 2066–71. Palo Alto, CA: AAAI Press.

Holmes, D., and Winston, P. 2016. Story-Enabled Hypothetical Reasoning. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*. Palo Alto, CA: Cognitive Systems Foundation.

Li, B.; Thakkar, M.; Wang, Y.; and Riedl, M. O. 2014a. Data-Driven Alibi Story Telling for Social Believability. Paper presented at the 2014 Social Believability in Games Workshop. Ft. Lauderdale, FL, April 4.

Li, B.; Thakkar, M.; Wang, Y.; and Riedl, M. O. 2014b. Storytelling with Adjustable Narrator Styles and Sentiments. In *Interactive Storytelling: Proceedings of the Seventh International Conference on Interactive Digital Storytelling,* 1–12. Berlin: Springer.

Milli, S.; Hadfield-Menell, D.; Dragan, A.; and Russell, S. 2017. Should Robots Be Obedient? arXiv preprint: arXiv:1705.09990[cs.AI]. Ithaca, NY: Cornell University Library.

Moerland, T.; Broekens, J.; and Jonker, C. 2016. Fear and Hope Emerge from Anticipation in Model-Based Reinforcement Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence,* 848–54. Palo Alto, CA: AAAI Press.

Patil, A.; Liu, J.; Price, B.; Sharara, H.; and Brdiczka, O. 2012. Modeling Destructive Group Dynamics in Online Gaming Communities. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 290–97. Palo Alto, CA: AAAI Press.

Pearce, C.; Meadows, B. L.; Langley, P.; and Barley, M. 2014. Social Planning: Achieving Goals by Altering Others' Mental States. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence,* 402–9. Palo Alto, CA: AAAI Press.

Riedl, M. O. 2016. Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence. arXiv preprint: arXiv:1602.06484[cs.AI]. Ithaca, NY: Cornell University Library.

Van Stekelenburg, J., and Klandermans, B. 2013. The Social Psychology of Protest. *Current Sociology* 61(5–6): 886–905. doi.org/10.1177/0011392113479314.

Vattam, S.; Klenk, M.; Molineaux, M.; and Aha, D. W. 2013. Breadth of Approaches to Goal Reasoning: A Research Survey. In *Goal Reasoning: Papers from the ACS Workshop*, edited by D. W. Aha, M. T. Cox, and H. Muñoz-Avila. Technical Report CS-TR-5029. College Park, MD: University of Maryland.

Wenar, C. 1982. On Negativism. *Human Development* 25(1): 1–23.

Wilson, J. R.; Arnold, T.; and Scheutz, M. 2016. Relational Enhancement: A Framework for Evaluating and Designing Human-Robot Relationships. In *AI, Ethics, and Society: Papers from the 2016 AAAI Workshop,* edited by B. Bonet, S. Koenig, B. Kuipers, I. Nourbakhsh, S. Russell, M. Vardi, and T. Walsh. AAAI Technical Report WS-16-02. Palo Alto, CA: AAAI Press.

Wright S. C.; Taylor D. M.; and Moghaddam, F. M. 1990. The Relationship of Perceptions and Emotions to Behavior in the Face of Collective Inequality. *Social Justice Research* 4(3): 229–50.

Alexandra Coman (PhD, Lehigh University, 2013) is a senior manager at Capital One. She was previously an NRC postdoctoral research associate with the Adaptive Systems Section at NRL in Washington, DC, and an assistant professor of computer science at Ohio Northern University. Her research experience and interests include cognitive architectures, AI planning, case-based reasoning, goal reasoning, AI ethics, affective computing, and narrative intelligence.

David W. Aha (PhD, University of California, Irvine, 1990) leads NRL's Adaptive Systems Section in Washington, DC. His interests include mixed-initiative intelligent agents (for example, that employ goal reasoning models), deliberative autonomy, machine learning, and case-based reasoning, among other topics. He has co-organized 35 events on these topics (such as ICCBR-17), hosted 13 post-doctoral researchers, served on 20 PhD committees, created the UCI Repository for ML Databases, was a AAAI Councilor, and gave the Robert S. Engelmore Memorial Lecture at IAAI-17.

Year One of the IBM Watson AI XPRIZE: Case Studies in "AI for Good"

Sean McGregor, Amir Banifatemi

■ *The IBM Watson AI XPRIZE is a four-year competition where teams work to improve the world with artificial intelligence. The competition began in 2017 with 148 problem domains in sustainability, artificial general intelligence, education, and a variety of other grand challenge areas. Fifty-nine teams advanced to the second year of the competition and 10 teams earned special recognition as "milestone nominees." The properties of the advancing problem domains highlight opportunities and challenges for the "AI for Good" movement. We detail the judging process and highlight preliminary results from cutting the field of competing teams.*

Investment in artificial intelligence has grown to more than $25 billion annually (Bughin et al. 2017), but these investments place higher priority on financial returns than the general welfare of humanity. To focus AI development on direct societal benefits, the IBM Watson AI XPRIZE (AIXP) issued a $5 million prize purse to award AI startups and researchers producing the greatest world-improving impact.

While the incentive for winning the AIXP is consistent with other XPRIZE competitions, the AIXP does not set a single shared objective for all teams. Rather, the AIXP invites teams to describe their own grand challenge and to demonstrate achievements over a four-year competition. This open prize structure allows teams to showcase a variety of approaches to the most significant problems faced by humanity. Problem flexibility also allows teams to discover unexpected opportunities. In many cases, a clever formulation may be the only requirement for improving millions of lives.

Problem Domain Category	Team Count	Example Problem Area
Humanizing AI	7	Moral and Ethical Norming
Emergency Management	5	Planning Disaster Response Logistics
Health	13	Drug Efficacy Prediction
Life Wellbeing	21	Augmenting The Visually Impaired
Environment	8	Automated Recycling
Education/Human Learning	17	Intelligent Tutoring System
Civil Society	11	Online Filter Bubbles
Health Diagnostics	12	Radiography Image Segmentation
Robotics	5	Robotic Surgery
Knowledge Modeling	7	Automated Research Assistant
Civil Infrastructure	9	Earthquake Resilience Testing
Business	19	Optimizing Social Investment
Artificial General Intelligence	8	* (All of Them)
Brain Modeling and Neural Networks	6	Cognition Emulation

Table 1. High-Level Problem Domain Categories for Competing Teams.

While all teams will ideally succeed in their efforts, both successes and failures present opportunities to focus research efforts in developing AI for Good. Our previous work outlined the complete AIXP process and year one team statistics (McGregor and Banifatemi, forthcoming); this work explores the problem domains and attributes of teams identified as top performers within the first year of the competition.

The AIXP began in 2017 with 148 teams working in the problem domains of table 1. The rows are ordered from domains with the highest advancement rate (top) to the lowest advancement rate (bottom). If left unaddressed, these problems pose significant negative consequences for humanity, including lack of access to basic human needs, lack of well-being, lack of education, environmental degradation, increased inequality, reduction in health, and loss of life.

After the first year of the competition, 59 of the starting teams remain. The competition closes after three annual judged rounds and a final round at TED 2020. The judges will award a $3,000,000 grand prize, a $1,000,000 second place prize, and a $500,000 third place prize. They will award an additional $500,000 to teams with noteworthy successes achieved during the annual reporting periods.

Teams began the competition by submitting solution proposals that were then read and categorized by the XPRIZE Foundation staff. The resulting team count within the team taxonomy of table 1 motivated the target list for judge recruitment. Appropriately judging 148 teams working towards different grand challenges required a judging panel with diverse technical, philosophical, and personal experiences. The 33 judges active in the first round of the AIXP have distinguished themselves either through their technical capacities within the field of AI or through their knowledge of the deployment of these systems in the real world. Among the judges are leaders from the labs of multinational corporations, AI startups, academic research labs, nongovernmental organizations, and public policy think tanks. Collectively these individuals have expertise in natural language processing, deep learning, adversarial learning, computer security, the social effects of technology, political campaigns, computational sustainability, ecology, robotics, and many other fields and applications of AI research. Judge biographies are available on the AIXP website.[1]

In September of 2017, competing teams submitted their first annual reports (FARs) as four-page extended abstracts detailing their problem areas, proposed solution, and the progress achieved to date. Of the 148 teams eligible to submit the FAR, only 118 teams opted to do so. This reduction shows significant self-selection that we consider for the purposes of analy-

sis to be similar to a judged rejection. Judges followed a similar review process as with an academic AI conference, with two reviewers per submission.

The advancement criteria focused on the potential for world impact and indicators of technical progress. Of the 118 teams submitting FARs, 40 teams were rated for acceptance in both judges' overall rating and were automatically accepted. Next, 44 teams joined the rejection list based on their overall ratings. Determining which of the remaining 34 teams to accept or reject required examination of more specific attributes of the scorecard. Teams were rejected when at least one judge did not rate the problem as important for humanity, when neither judge rated the problem as previously unsolved, when neither judge rated the technology as having the capacity to solve the problem, or when neither judge indicated that the team showed incremental progress. Fifteen teams were rejected on these grounds. The remaining 19 teams were then accepted into the second year of the competition.

All FARs were reviewed by at least one judge who self-assessed at the medium level of technical proficiency or higher, and one judge who self-assessed at medium level of problem domain proficiency or higher. The box and whisker plots in figure 1 provides additional details on the spread of judge confidence levels.

With the list of teams accepted into year two of the competition, the next step was to award the first allocation of the $500,000 milestone prize purse. The top 10 performing teams were nominated for milestone prizes based on the top 10 average overall ratings assessed for the FARs. In this article, we present additional details for these milestone teams (Team Brown HCRI, Team DeepDrug, Team BehAIvior, Team aifred health, Team Amiko AI, Team WikiNet, Team emPrize, Team Erudite AI, Team Iris.ai, and Team DataKind).

AIXP judges and staff ranked the milestone teams with collaborative ranking, a process by which each judge reviewed two additional reports and assessed one report as "better." The resulting ordered pairs formed a scoring measure in which the top two teams were consistent with an ordered list of minimal weighted pairwise dissimilarity for all ordered judge pairs.

We validated the performance of the weighted metric via Monte Carlo trials (figure 2) for a range of oracle conformance values, which we define as the probability a judge will agree with an arbitrarily chosen "true" ranking. The convergence to the oracle ranking shows the method by which consensus rankings were produced. Since the only publicly ranked teams are those winning milestone prizes, the analysis shown in figure 2 focuses on the top two milestone teams as determined by an oracle. For these Monte Carlo trials, we generated ranked pairs for all pairs of teams and forced the ranking to conform with the oracle according to a "conformance score." For teams t_i and t_j, the ranking given by judges conforms with the oracle with probability

$$\frac{1}{\left(\left|R(t_i)-R(t_j)\right|+1\right)^{k}}$$

where $R(\cdot)$ refers to the oracle ranking and K is the conformance assumption. Even when judges agree with the oracle for adjacent teams with probability 0.5, the "best" team is in the top two with probability 0.7.

Case Studies in the AI for Good Movement

Teams developing AI solutions to real-world problem domains employ a variety of AI techniques. Consequently, the advancement statistics should be regarded as indicators of the opportunity offered by the problem domain, rather than the opportunity of any individual AI technology. We begin by exploring the problem categories where teams showed disproportionate success within the first annual reports and finish with the underperforming problem categories. Figures 3 and 4 give the advancement rate context with advancement percentages and counts, respectively. In figure 3, the stacked bar chart shows percentages for advancement, rejection, and nonsubmitting within each of the problem domains. In figure 4, the team advancement stacked bar chart shows counts for advancement, rejection, and nonsubmitting within each of the problem domains. Of 148 teams, 30 did not submit first annual reports. Of the 118 submitting teams, judges then selected to advance to year two.

Humanizing AI

Teams involved in humanizing AI are concerned with solving the problems introduced by placing AI into the human context. The milestone nominee from this group, Brown Human-Centered Robotics Initiative (HCRI), aims to "create robots that obey social and moral norms." One example is mapping the attributes of a scene to behaviors, such as mapping "library" to a reduction in audio communication volume. While HCRI was primarily concerned with automatic inference of these norms, other teams took an end-user programming approach in which the system is more directly programmed by people in the environment. In both cases, these teams showed a greater success rate than the rest of the field because they are attempting to solve challenging problems faced by the deployment of all AI systems to the real world. Any solution to the humanizing problem would have the potential to greatly expand the domains with which AI systems can interface.

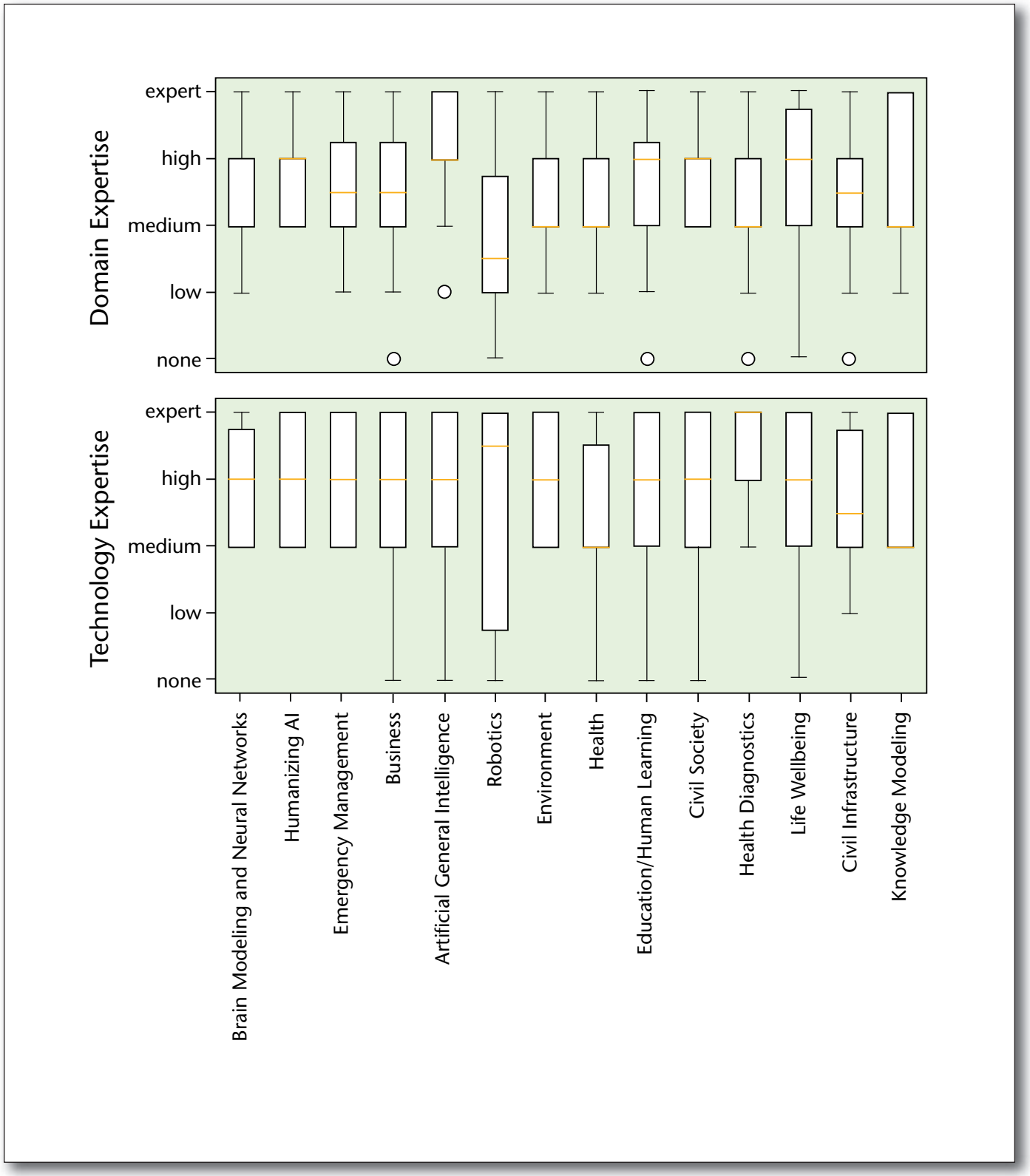


Figure 1. Box and Whisker Plots.

Self-assessed problem domain expertise (top). Self-assessed technology expertise (bottom). The orange line indicates the median value, and the box extends to the upper and lower quartiles. The whiskers show the extents. The circles are singleton outliers.

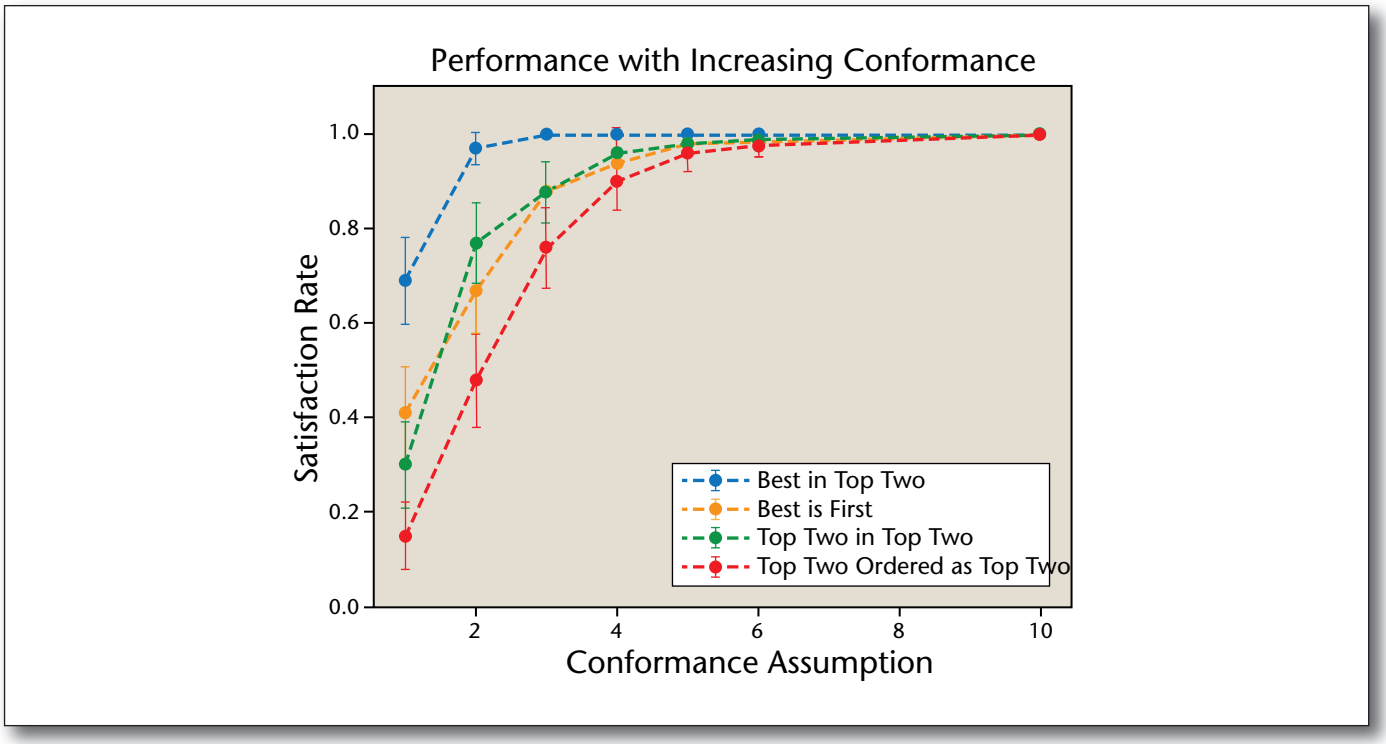


Figure 2. Performance Validation of the Weighted Metric via Monte Carlo Trials.

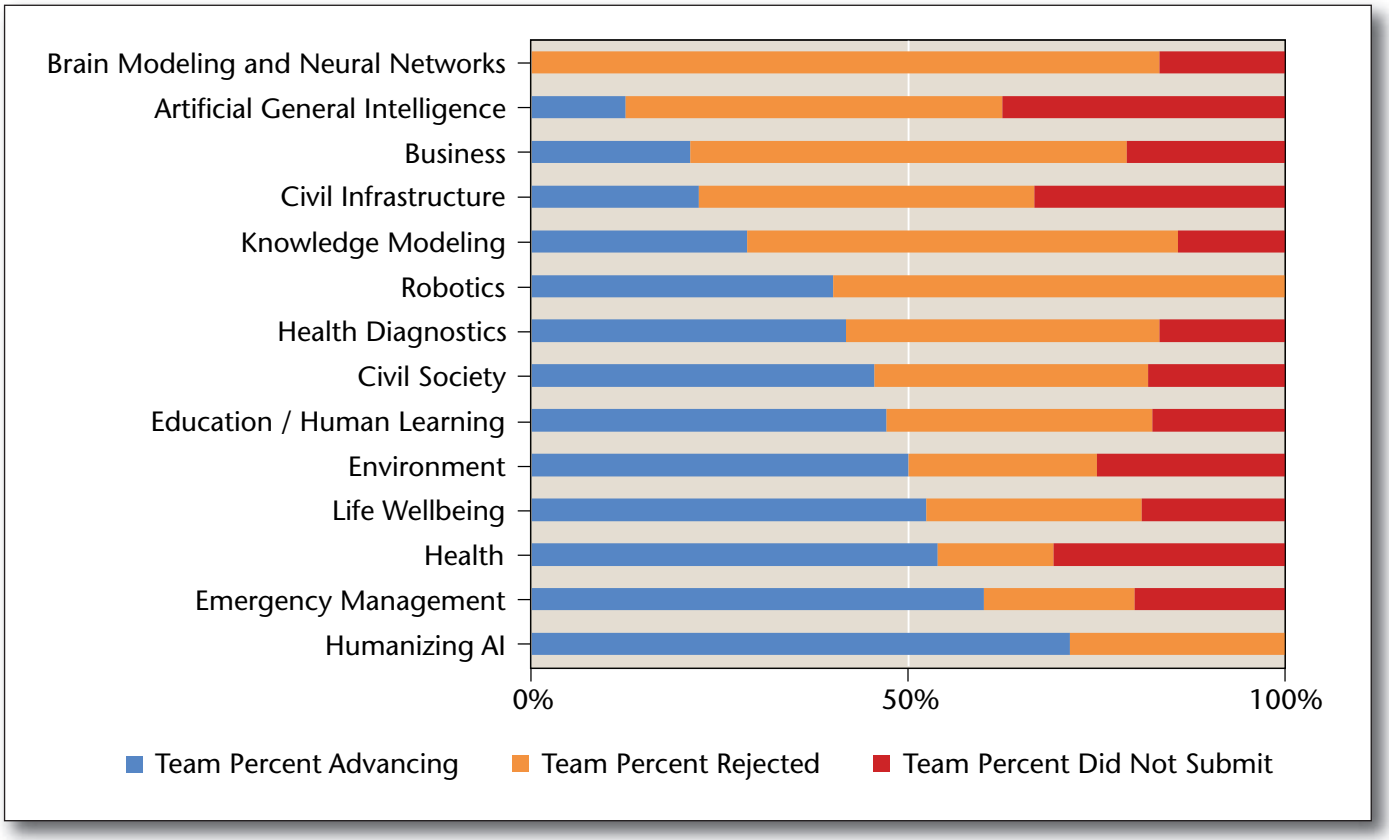


Figure 3. Team Advancement Bar Chart (Percentages).

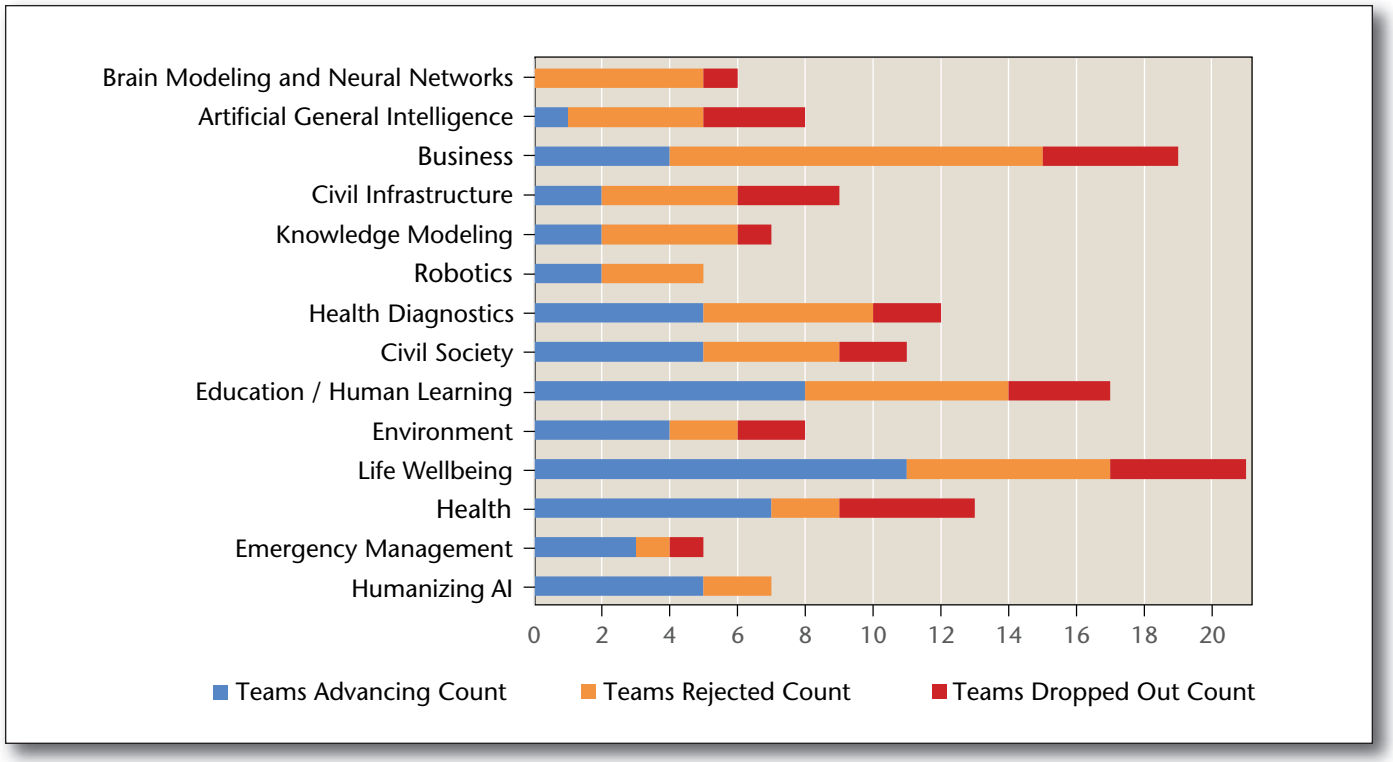


Figure 4. Team Advancement Bar Chart (Team Count).

Emergency Management

Many teams involved in emergency management are data synthesis teams for performing operations research tasks in uncertain, fast-changing environments. Most of these teams gained entry into year two of the competition because they are (1) clearly working in an area that could have a real immediate impact for millions of people affected by disasters, (2) working in a problem space that has fallen behind technological capacities, and (3) benefiting from a wealth of newly accessible data streams (drones, daily satellite imagery, reliable emergency communications). The key attributes of these teams are the development of specialized hardware for disaster management or the development of models that can take immediate and ongoing surveys of the disaster area to prioritize rescues, resource deployment, and other disaster response activities. The judges did not nominate any teams within this group for a milestone award, but emergency management teams likely require field demonstrations to be nominated for an award.

Health

Due to the high number of teams working on health-related problems, we split the health teams into "health" and "health diagnostics." The teams in the "health" category are working on problems of longevity (zero of three teams advancing), medical personalization (one of three teams advancing), mental health (five of six teams advancing), and drug discovery (one of one team advancing). The teams working on longevity may have fallen into the same trap as the teams working on artificial general intelligence (detailed later), attacking the top-level problem without a concrete roadmap of deliverables. Team DeepDrug, a milestone nominee, was the one team working on drug discovery. This team distinguished itself by building on top of their history of academic research.

The most surprising aspect of the health advancement statistics is that so many of the mental health teams successfully advanced to year two. The mental health teams have significant challenges in ensuring the safe and ethical deployment of their technologies, but the scale at which synthetic intellect can potentially serve mental health needs was ample justification for advancing these in-development solutions to year two. One mental health team, BehAIvior Health, was nominated for a milestone award for predicting and preventing addiction relapses and overdoses using wearables.

The one medical personalization team admitted to year two, aifred health, placed second in the milestone competition. Their work predicting the effectiveness of mental health treatments is an excellent example of an underserved problem in an otherwise

Team Brown HCRI

As robots increasingly take part in important areas of society such as medicine, social care, education, or disaster response, we must ensure that they follow the social and moral norms of the communities they are part of. Currently, however, robots follow only basic instructions without any conception of social and moral norms. This, then, is the grand challenge that the Brown HCRI team poses: to teach robots social and moral norms. The team has initiated an interdisciplinary research program that aims to meet this grand challenge in three phases. In the identification phase, the team is developing experimental research methods and algorithms to identify human norms for a subset of contexts and communities (for example, senior care, medical assistance, education).

Next, in the implementation phase, the team is building computational networks of norms that have been identified for the specific contexts. These networks must be flexible enough to learn subtle context variations and to add or update norms when receiving feedback from trusted sources. Such feedback will come not only from people who interact with the system, but also from crowdsourced observers who are members of the relevant communities. Finally, in the evaluation phase, the team will be installing these networks in robots and evaluating their social acceptability in rigorous human-robot interaction studies. Some of these studies will take place in virtual and augmented-reality environments that enable immersive experiences but also permit experimental control over critical causal variables, such as the robot's appearance or the transparency of its norm competence.

For more information, see hcri.brown.edu. The team contact is Bertram Malle (bfmalle@brown.edu).

Team DeepDrug

In this age of antibiotics, there is still an ongoing effort to discover new drugs to combat illnesses for which there is no known cure. In addition, there is a need to discover replacements for existing drugs for pathogens that have become resistant. Although multidrug resistance in pathogens is growing fast, the development of new drugs to treat bacterial infections has reached its lowest point since the beginning of the antibiotic era. The existing process for creating new drugs is slow, inefficient, and costly. DeepDrug is developing eSynth, a drug design software that generalizes from existing drug trial datasets to create an improved method for identifying drug compounds.

eSynth can automatically synthesize targeted drug molecules, filter candidates based on chemical criteria (such as being an antibiotic or toxicity), analyze 3D image models of the pathogen for possible drug repurposing, automate clinical testing for side effects, and predict the candidates most likely to succeed. Recent progress includes design, training, and testing of several AI filters and engines that have shown promising results.

The team contact is Supratik Mukhopadhyay (supratik@csc.lsu.edu).

heavily developed market sector. Drug companies have little incentive to develop methods for intelligently personalizing prescriptions since the intelligent agent may select the drugs of a competitor. aifred health also excels in the systematic way they are pursuing interdisciplinary research. In addition to developing predictive models, they are developing ethical frameworks to evaluate the performance of their systems.

Life Well-Being

Teams concerned with life well-being are attempting to solve quality-of-life issues, including AI designs for the hearing and vision impaired (three of four advancing), personal life management (six of eleven advancing), independent living assistance for the elderly or infirm (one in five advancing), and one team working to produce an online safety agent (advancing). Several successful teams from these groups are finding ways of promoting everyday wellness by extending the reach of clinical professionals beyond the doctor's office. The first-prize milestone winner, Amiko AI, developed a model and sensors to support the continuous monitoring of asthma treatments. Amiko AI could easily be categorized a health team, but their focus on facilitating the doctor and patient relationship expands the boundaries of the medical profession into the promotion of wellness.

Environment

Teams working on environmental problems are developing solutions within the subcategories of agriculture (one in four advancing), recycling (one of one advancing), species abundance (one of one advancing), water quality (zero of one advancing), and pollution mitigation (one of one advancing). WikiNet served as the pollution mitigation team and received a nomination for a milestone award for their work with the large unstructured corpus of environmental remediation documents to build a system that can recommend best practices on future remediation efforts.

Team BehAIvior

For individuals with substance use disorder, the propensity for returning to drug use (that is, relapsing) is high. Historically, tools to fight addiction have been limited and retrospective. By the time a traditional intervention occurs, people are often already using again. Relapses often lead to a costly downward spiral — committing crimes, getting rearrested, being hospitalized, and overdosing, sometimes fatally. Recent advances in wearable sensors, smartphones, and artificial intelligence have created an opportunity to produce positive health outcomes by predicting and preventing relapses and overdoses. The first step in this proactive relapse prevention is to identify and measure digital biomarkers associated with relapse, and to implement a predictive model to achieve just-in-time intervention. BehAIvior is developing a relapse prevention platform that will consider physiological sensor data from wearable devices in combination with smartphone usage data and location data to identify and detect relapse triggers in real time. The first use case is opioid addiction, but the tool will, in subsequent iterations, be used to identify and react to any addiction, behavior, or condition — stress, smoking, overeating, even suicide. BehAIvior has partnered with Carnegie Mellon University computer scientists and University of Pittsburgh addiction experts to execute an interdisciplinary development plan.

For more information, see behaivior.com. The team contacts are Jeremy Guttman and Ellie Gordon (hello@behaivior.com).

Team aifred health

Depression has a lifetime prevalence of 11.1 percent: over 350 million people are affected at any one time. It is the leading cause of disability, it can lead to suicide, and overall it carries a high socioeconomic cost. While a range of effective treatments exist, patient responses to treatments are heterogenous. Some patients spend years going through a process of trial and error before finding the treatment that works for them. Clinicians do not have any principled way to personalize treatments for individual patients or to predict which patients will have which side effects. To solve this treatment selection problem, aifred health is building a clinical decision aid. The system predicts treatment response, side-effect profiles, and suicide risk based on clinician observations, patient self-report, and biomarkers. This clinical aid will enrich shared decision-making between clinicians and patients, help patients improve faster, and reduce social costs. The deep learning–based prototype architecture utilizes stacked denoising autoencoders and snapshot ensemble technology to predict suicidal ideation. It incorporates interpretability technologies, such as saliency maps, to help explain predictions to physicians. aifred health has secured data partnerships with academia and industry, published an AI ethics framework (Benrimoh et al. 2018), and designed rigorous clinical trials to test the system.

For more information, see aifredhealth.com. The team contact is Eleonore Fournier-Tombs (eleonore@aifredhealth.com).

Education and Human Learning

The teams working on education are developing different ways to make education more personalized, effective, scalable, or cost efficient. Of the 17 teams eligible to advance, eight were admitted into year two and two were nominated for milestone awards. Milestone nominee emPrize is developing and deploying AI technologies to online classrooms, including components for cognitive tutoring, question answering, and formative assessment. Of particular interest to the judging panel was the early testing of system efficacy within real-world scenarios. This trait is shared by the other milestone nominee from the education domain, Erudite AI, who developed and began testing a system for connecting students that need help with a topic to students who are predicted to tutor the topic well. The complexities of educational systems are such that real-world demonstrations are crucial for establishing the efficacy of the system and gaining special recognition for the effort.

Civil Society

Of the 11 teams in the competition in the subcategories of information consumption, equity, law, and safety, most of the five teams moving on to the next round were in safety. These teams work on problems of scaling up law enforcement for fighting sex trafficking advertised online, and making the roads safer with vehicle-mounted computer vision systems. The three teams working on information consumption (the problems of filter bubbles, fake news, and so on) were all developing AI solutions to problems introduced by optimization algorithms applied to media consumption habits. While an AI solution may exist in some form, there is no clear answer to how an AI system can independently solve social problems introduced by another AI system. None of the teams working on the fake news problem advanced. Still, in search of solutions these teams made commendable efforts in attempting to understand the problem. It is unfortunate that the competitive marketplace means third parties cannot experiment directly with the optimization algorithms controlled by new media companies.

Health Diagnostics.

Due to the high number of teams working on health-related problems, we split the health teams into "health" and "health diagnostics." The health diagnostics teams are largely concerned with diagnosing medical conditions through computer vision for radiography, biometric signal processing with always-on health sensors, and other applications of raw health data. The health diagnostics teams were all working on worthy problems, but their apparent failure mode is that these solutions are generally under active development in many corporate and university research labs. Teams would be more suc-

cessful in this domain if they were not implicitly competing with many researchers outside the competition.

Robotics

The teams in the robotics category were so classified because their proposal involved the development of robotics without a clear problem solved by new robotic capacities. These teams were also at a significant disadvantage for showing progress since many planned to work with novel robotic architectures that can take years to develop. It is difficult to show progress in work such as this compared to the more nimble machine learning problems. Further, the AIXP focus on real-world outcomes highlighted that many of the nonindustrial applications of robotics have a backlog of fundamental advancements required before robotics can be a part of everyday life (as shown, for example, with the problems being solved by the humanizing AI teams).

Knowledge Modeling

The heading of knowledge modeling spans practices within AI that could be described as applied data mining. One milestone nominee, Iris.ai, is working within this domain to produce a research assistant to accelerate literature review and concept discovery. Iris.ai differentiates itself from the less successful teams in the domain by presenting a system that can be evaluated for a specific purpose. Otherwise, building a knowledge base intended for general purpose queries is too abstract to benchmark.

Civil Infrastructure

The primary barrier to improvement within this domain is often not the absence of good ideas. There are many trivial optimizations of society that do not gain adoption for budgetary or political reasons. The milestone nominee, DataKind, avoids these problems by building their solutions for countries that lack adequate measurement to perform basic civil services. Datakind processes satellite imagery to perform image segmentation of poverty and disease rates. The automatic generation of these predictions globally has the capacity to selectively deploy scarce development interventions in the areas most needing them.

Business

The business team category served as a catchall for teams not fitting into a category beyond building a business centered on AI. While a successful business proposition is often an indicator of a system's social utility, many business teams failed to articulate an advancement for society more generally. In some cases, the advancing business teams adjusted their project to more explicitly target social benefit, which may lead to their recategorization in the future.

Team Amiko AI

Asthma affects over 300 million people worldwide. Each year, there are millions of asthma-related hospitalizations and emergency department visits, which contribute to unsustainable healthcare costs. And many, if not most, asthma-related exacerbations are preventable with proper treatment. In fact, despite the widespread availability of effective treatments, patients struggle to follow their treatment plans, while physicians lack the tools and the information to understand how their patients are doing and to find the best therapy for each of them.

Amiko AI developed a digital health platform, Respiro, for real-time monitoring of medication administration and patient health with sensors and connected health tools. At the core of the platform is a set of sensors for respiratory devices, such as inhalers, that automatically track the patient's inhalation profiles to monitor breathing health and record when and how well patients use their medication.

The Respiro sensors extrapolate key clinical parameters, such as the quality of the drug delivery, by analyzing the vibrational energy that is recorded during a patient's inhalation maneuver.

For more information, see amiko.io. The team contact is Luca Ponti (luca.ponti@amiko.io).

Team WikiNet

Over 200 million people are potentially exposed to toxic pollutants from contaminated sites in 50 developing countries (Hanrahan, Ericson, and Caravanos 2016). As soil and groundwater contamination can pose a significant threat to human health, the remediation of these sites is of great importance. However, contaminated site remediation can be highly complex and presents significant uncertainties. To select an appropriate treatment, environmental experts must analyze structured and unstructured data (for example, site assessment reports, lab results, maps). In addition, the selected treatments must optimize multiple objectives such as the performance, cost, and timeframe for the remediation. Although remediation experience and technical knowledge are key to making an informed decision, the analysis of past remediation reports and scientific research is a laborious and time-consuming task. WikiNet's goal is to facilitate the analysis of such documents and provide automated expert recommendations for treating contaminated sites worldwide.

The solution is composed of an information extraction system that extracts key parameters from site reports (for example, contaminants to treat, site geology), a classifier that learns from past remediation efforts to recommend treatments based on site-specific characteristics, and a regression predictor for treatment cost estimates. The team has developed an initial information extraction system and obtained encouraging results for the named entity recognition and relationship extraction of 24 entities and 21 relations specific to the environmental field. They also trained a feed-forward neural network classifier that can currently recommend nine distinct treatments based on contaminated site features. See wikinet.ca.

Team emPrize

Online education is growing rapidly, despite low student retention for many online classes. The quality of online learning is questionable in part because of a lack of learning assistance. How can we provide meaningful learning assistance to tens of millions of students taking online classes? Team emPrize is developing a suite of virtual tutors for online education that mimic many of the roles of human teachers. These virtual tutors include more than 100 cognitive tutors for a Georgia Tech online class on artificial intelligence as well as a virtual tutor for automatically answering questions on the discussion forum for the class. Preliminary results indicate that student self-efficacy in the class is high and that interaction with the virtual tutors leads to enhanced student engagement. emPrize is now expanding the scope of their work from online education to blended learning; from cognitive tutoring and question answering to exploration and experimentation, literature survey, and question asking; and from a class on artificial intelligence to Georgia Tech classes on introductory computing and introductory biology.

The team contact is Ashok Goel (goel@cc.gatech.edu).

Team DataKind

Globally, crop disease causes nearly 50 percent of the total loss of crops. It is especially devastating for communities in developing nations where 75 percent of the population relies on agriculture for their livelihood. Early detection is critical to fight plant pathogens, as there is a narrow timeframe in which to intervene to save crops and prevent epidemics. However, effective early warning systems to alert communities of imminent threats of disease do not currently exist in developing regions.

DataKind, a nonprofit that uses AI to address complex humanitarian issues, is developing a model using high-resolution satellite imagery at 5 meters per pixel, combined with computer vision and remote sensing techniques, to detect the spatial and spectral signature of wheat crops and wheat disease, to be able to provide real-time information on crop disease and support the creation of enhanced early warning systems.

DataKind first worked to identify wheat in Ethiopia, beginning by locating croplands in the region with high spatial resolution. They then successfully built a U-Net model with a 5-meter resolution to detect croplands in Montana, a climate proxy for Ethiopia, achieving approximately 93 percent test accuracy, and a characteristic curve approaching 96 percent for the area under the receiver operator. The model was transferred using field survey data from Ethiopia, and from human inspection, appears quite promising. In the second phase of the project, DataKind is looking to obtain noncrop survey ground truth data for Ethiopia to further tune and test the model.

For more information, see datakind.org.

Team Erudite AI

Students who regularly receive private tutoring score two standard deviations higher on standardized tests than those students without private tutoring. However, the demand for private tutoring far outstrips the supply, with up to 65 percent of students seeking sessions in Kenya and 73 percent in Sri Lanka. Consequently, tutoring suffers from low access, compromised quality, and the high cost for one-on-one sessions. Erudite AI's solution endeavors to mitigate all three problems with a peer-to-peer tutoring platform, ERI (educational real-time interface). ERI is a human-in-the-loop dialogue-based tutoring platform comprising three main components: a mapper to identify and build a knowledge map of the students' skills, a matcher to match students to peer tutors according to their needs, and an amplifier that elevates the quality of the tutoring by suggesting AI-generated responses for the peer tutor. In the past few months, Erudite AI evaluated the effectiveness of a dialogue recommender to positive results. Following the experimental evaluation, the team is producing a scalable open source solution to maximize impact.

For more information, see eri.ai. The team contact is Hannah Cowen (info@erudite.ai).

Artificial General Intelligence

Of the eight teams competing to develop the first artificial general intelligence, only one advanced. The likely reason is that teams must show a plausible means of successfully completing their grand challenge, and establishing a plausible pathway to AGI within the timeframe of the competition is itself a grand challenge. The one team advancing from this category trimmed their ambitions to a sufficient degree so that they can plausibly produce their system within the competition timeframe.

Brain Modeling and Neural Networks

Finally, many teams proposed to develop new approaches to neural networks. These teams often emphasized architectures that are inspired by the human brain. While some of the approaches may prove successful in the fullness of time, there is no shortage of proposals for new neural network architectures. Without a demonstrated capacity for solving a problem that was not solvable by previous neural network architectures, new proposed architectures

Country	Team Count	Advancing Count	Advancing Percent
Barbados	1	1	100
Israel	1	1	100
Norway	1	1	100
Poland	1	1	100
Canada	20	11	55
UK	6	3	50
USA	71	30	42
China	6	2	33
Italy	6	2	33
Vietnam	3	1	33
France	7	2	29
Australia	8	2	25
Germany	4	1	25
India	5	1	20
Netherlands	2	0	0
Czech Republic	1	0	0
Ecuador	1	0	0
Japan	1	0	0
Romania	1	0	0
Spain	1	0	0
Switzerland	1	0	0

Table 2. Home Countries, Counts, and Advancement Rates for Competing Teams.

don't represent a grand challenge. In time, we expect some of these teams will show empirical promise, but without preliminary evidence they are unlikely to advance.

Ethics and the Future of AI

The most challenging aspect of running an open-ended competition for artificial intelligence is the capacity for AI systems to solve global challenges (see table 2 for team geographies), while also introducing novel and unforeseen trade-offs. Teams competing in the AIXP may deploy mental health dialogue agents, medical recommender systems, and other technologies where the betterment of the many does not preclude harm to a few. AIXP judges serve as arbiters of global beneficence, but there is currently no expert body that has a global process for recommending procedures for deploying and monitoring AI systems. While the IBM Watson AI XPRIZE has the resources to review AIXP teams, a near future with ubiquitous AI requires review methods that scale beyond formal committees of the world's leading experts. Many organizations are working to fill the void of formal process. Major corporations developing AI products formed the Partnership on AI^2 as a joint effort with civil society organizations. Academics and engineers drafted principles and standards for the ethical devel-

Team Iris.ai

We live in a world where more scientific discovery is underway than ever before — but the research process is plagued with hard-to-justify inefficiencies, and among them, the growing need to distill and filter through all the noise. Interdisciplinary exploration is vital to new discovery, but exploring a new field where one is not a domain expert can be immensely time consuming.

Aiming to build an AI researcher for literature-based discovery, Iris.ai semiautomates the time-consuming process of literature review. Their "exploration and focus" tools reduce the time required to go from a problem statement to a reading list by 90 percent, while also increasing interdisciplinary discovery.

The Iris.ai team is focusing on extraction of a research paper's key concepts, together with an encoding technique that can construct a document vector space based on the available information. This strategy allows the building of intuitively meaningful content-based indexes. The team's next steps are developing hypotheses-extraction techniques and word-to-word graph representations of documents.

Evaluation has shown a reduction in time for research teams augmented with the Iris.ai exploration tool. In building the document vector space, their WISDM metric shows a consistent speed-up, while upholding precision of comparable models.

For more information, see iris.ai.

opment of AI, including the Future of Life Institute,[3] IEEE,[4] The Royal Society,[5] and the Stanford AI100 project.[6] Governments, intergovernmental organizations, and nongovernmental organizations, including the European Parliament[7] (Goodman and Flaxman 2017) and the International Telecommunication Union,[8] are holding summits and passing sweeping regulations. Clearly, the culture and law of ethical AI development will be enacted over the next decade.

Areas of beneficence, fairness, explainable AI, and other aspects of AI governance will be a focus in round two of the competition. We look to feedback from our advisory board and judges to adapt the competition guidelines to ensure the ongoing execution of a competition process that is fair to competing teams and maximally impactful in the real world.

Competing AIXP teams are at the forefront of ethical AI development through their pursuit of $5 million in prize money. Their efforts support the movement with applications of AI that are beneficial for humanity, that demonstrate human and machine collaboration, and that identify the greatest opportunities for AI to make an impact on society. While AI techniques are developing quickly, we have an opportunity to better understand where research intersects with grand challenge applications to produce new opportunities. An open competition plan has allowed teams from many backgrounds to tackle hard problems with AI. As the competition proceeds to year two, the XPRIZE team, along with the prize sponsor IBM and other supporting ecosystem partners, look forward to seeing the good an impassioned group of AI developers can produce in the world.

Acknowledgements

First and foremost, the teams competing to make the world a better place deserve special recognition for their efforts. Next, IBM has shown great vision in supporting such an open-ended endeavor.

The IBM Watson AI XPRIZE relies on an advisory board including Yoshua Bengio, Francesca Rossi, Rob High, Babak Hodjat, Neil Jacobstein, Subbarao (Rao) Kambhampati, Peter Norvig, Tim O'Reilly, Jean Ponce, Lav Varshney, and Manuela M. Veloso.

The judges perform the hard work of balancing imagination and critical review. They include Gabriel Skantze, Carla Gomes, Eric Van Gieson, Adam Cheyer, Robin Murphy, Danah Boyd, Ivan Laptev, Bistra Dilkina, Alex London, Al Kellner, Erin Walker, Madeleine Clare Elish, Franc¸ois Chollet, Sidney D'Mello, David Kale, Danielle Tarraf, Xiaoyang Wang, Evan Muse, Nicolas Papernot, Henry Kautz, Risto Miikkulainen, Pascal Van Hentenryck, Mark Crowley, Forent Perronnin, Bill Smart, Graham Taylor, Julien Mairal, Stefano Ermon, Antoine Bordes, Jonathan Zittrain, Michael Gillam, Peter Eckersley, Barry O'Sullivan, and Rayid Ghani.

Finally, the XPRIZE staff members Jennine Dwyer, Yvonne Cooper, Katherine Schelbert, Michael Martin, Sean Beougher, Daniel Miller, Stephanie Wander, and Ed McNierney have all been instrumental in organizing the IBM Watson AI XPRIZE.

Notes

1. ai.xprize.org/about/judges.

2. partnershiponai.org.

3. See the Asilomar AI Principles (futureoflife.org/ai-principles).

4. Such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (standards.ieee.org/news/2017/ieee_global_initiative.html).

5. The Royal Society issued a report on machine learning in 2017 (royalsociety.org/topics-policy/projects/machine-learning).

6. The AI100 Project, a collaboration of AI scientists, issued a report in 2016 called *Artificial Intelligence and Life in 2030* (ai100.stanford.edu).

7. See the Council of the European Union, European Parliament, Regulation (EU) 2016/679 of April 27, 2016 (publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en).

8. The AI for Good Global Summit 2017, www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx.

References

Benrimoh, D.; Israel, S.; Perlman, K.; Fratila, R.; and Krause, M. 2018. Meticulous Transparency — An Evaluation Process for an Agile AI Regulatory Scheme. In *The 31st International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems, Special Track on Artificial Intelligence, Law, and Justice,* 1–12. Berlin: Springer.

Bughin, J.; Hazan, E.; Ramaswamy, S.; Chui, M.; Allas, T.; Dahlstrom, P.; Henke, N.; and Trench, M. 2017. *Artificial intelligence — The Next Digital Frontier?* Chicago, IL: McKinsey Global Institute.

Goodman, B., and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine* 38(3): 50–57.

Hanrahan, D.; Ericson, B.; and Caravanos, J. 2016. Protecting Communities by Remediating Polluted Sites Worldwide. In *Proceedings of the Institution of Civil Engineers–Civil Engineering* 169, 33–40. London: Thomas Telford Ltd.

McGregor, S., and Banifatemi, A. Forthcoming. First-Year Results from the IBM Watson AI XPRIZE: Lessons for the "AI for Good" Movement. In *The NIPS '17 Competition: Building Intelligent Systems* edited by S. Escalera and M. Weimer. Berlin: Springer.

Sean McGregor is a technical lead for the IBM Watson AI XPRIZE and a member of the technical staff at Syntiant Corp. His research interests include the optimization and explanation of machine learning systems, including problems in wildfire suppression policy, heliophysics, and low-precision neural network models. He earned his PhD in machine learning from Oregon State University in 2017 and his BA in environment, economics, and politics and computer dcience from Claremont McKenna College in 2004.

Amir Banifatemi is the AI Initiatives lead at XPRIZE Foundation. He oversees the Frontier Technologies Group, including the IBM Watson AI XPRIZE, the ANA Avatar XPRIZE, and the Lunar XPRIZE. He has a background in engineering and research in machine vision, product design, and financial modeling. He holds an MS in electrical engineering from the University of Technology of Compiègne, a PhD in systems design and cognitive sciences from the University Paris Descartes, and an MBA from HEC Paris.

Alexa Prize — State of the Art in Conversational AI

Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Ashwin Ram, Raefer Gabriel, Rohit Prasad

■ *To advance the state of the art in conversational AI, Amazon launched the Alexa Prize, a $2.5 million competition that challenges university teams to build conversational agents, or "socialbots," that can converse coherently and engagingly with humans on popular topics for 20 minutes. The Alexa Prize offers the academic community a unique opportunity to perform research at scale with real conversational data obtained by interacting with millions of Alexa users, along with user-provided ratings and feedback, over several months. This opportunity enables teams to effectively iterate, improve, and evaluate their socialbots throughout the competition. Eighteen teams were selected for the inaugural competition last year. To build their socialbots, the students combined state-of-the-art techniques with their own novel strategies in the areas of natural language understanding and conversational AI. This article reports on the research conducted over the 2017–2018 year. While the 20-minute grand challenge was not achieved in the first year, the competition produced several conversational agents that advanced the state of the art, that are interesting for everyday users to interact with, and that help form a baseline for the second year of the competition.*

Artificial intelligence is becoming ubiquitous. With advances in technology, algorithms, and sheer compute power, it is now practical to utilize AI techniques in everyday applications in the domains of transportation, healthcare, gaming, productivity, and media. Yet one seemingly intuitive task for humans still eludes computers: natural conversation. While simple for humans, voice communication in everyday language continues to be one of the most difficult challenges in AI. Human conversation requires the ability to understand the meaning of spoken language, relate that meaning to the context of the conversation, create a shared understanding and world view between the parties, model discourse and plan conversational moves, maintain semantic and logical coherence across turns, and generate

  ISSN 0738-4602

natural speech. Conversational agents capable of natural human conversation have applicability in both professional and everyday domains.

Voice-based virtual assistants, an important type of conversational agent, have become very popular in the last several years. The first generation of such assistants — Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana — have been focused on short, task-oriented interactions, such as playing music or answering simple questions, as opposed to the longer free-form conversations that occur naturally in social and professional human interaction. Conversational AI is the study of techniques for creating software agents that can engage in natural conversational interactions with humans. Significant advances in this area are needed to make interactions with virtual assistants and other types of AI agents easier and more natural for everyday use, particularly for open-domain conversations, those that are not bounded to a single task or topic.

Conversational AI is still in its infancy and several leading university research teams are actively pushing research boundaries in this area (Serban et al. 2016; Vinyals and Le 2015). Access to large-scale data and real-world feedback can drive faster progress in research. To address this challenge, Amazon announced the Alexa Prize on September 26, 2016, with the goal of advancing the research in conversational AI. Selected university teams were challenged to build conversational agents, known as "socialbots," to converse coherently and engagingly with humans on popular topics such as sports, politics, entertainment, fashion, and technology for 20 minutes. The grand challenge is to conduct coherent and engaging conversations for 20 minutes, with an average rating of 4 or higher on a scale of 1 to 5.

Given the complexity of the challenge, Amazon collaborated with the participating teams to provide them with tools, data, and a unique opportunity to perform iterative research with a live system deployed to millions of Alexa users. Through the Alexa Prize competition, participating universities were able to conduct research by building socialbots, training conversational models, and testing hypotheses at scale. Alexa users interacted with socialbots via the "Alexa, let's chat" experience, engaged in live conversations, and left ratings and feedback for the teams at the end of their conversations. Over 40,000 hours of conversation were logged in the course of the 2017 competition through the finals last November. As users continue to interact with the winning socialbots, this has now become over 130,000 hours.

In this article, we describe the scientific problems related to open-domain conversational systems, the state of the art in addressing these problems, how these approaches were used during the inaugural competition, and the results and scientific advances obtained. We present the technical setup of the Alexa Prize Finals event along with the process of selecting the winner. We conclude with a summary of the work that we plan to address in the second year of the competition.

The Alexa Prize Experience

The Alexa Prize competition received hundreds of applications from interested universities. After a detailed review of the applications, Amazon announced 12 sponsored and 6 unsponsored teams as the inaugural cohort for the Alexa Prize. The teams that went live for the 2017 competition, listed alphabetically by university, were DeisBot (Brandeis University), Magnus (Carnegie Mellon University), RubyStar (Carnegie Mellon University), Alquis (Czech Technical University in Prague), Emerson (Emory University), What's Up Bot (Heriot-Watt University), Pixie (Princeton University), Wise Macaw (Rensselaer Polytechnic Institute), Chatty Chat (Seoul National University), Eigen (University of California, Berkeley), SlugBot (University of California, Santa Cruz), Edina (University of Edinburgh), MILA Team (University of Montreal), Roving Mind (University of Trento), and Sounding Board (University of Washington).

The university teams built socialbots using the Alexa Skills Kit (ASK) (Kumar et al. 2017). The Amazon team provided automatic speech recognition (ASR) to convert user utterances to text for the socialbots and text to speech (TTS) to render text responses from the socialbots to speech for the users. All the intermediate steps — natural language understanding (NLU), dialogue modeling, and conversational user experience (CUX) — were handled by the university teams through their socialbots. Teams were allowed to leverage the standard NLU system that is provided with ASK. We also provided live news feeds to enable socialbots to stay current with popular topics and news events that users might want to talk about, and other tools and data as described in this article.

While the Alexa Prize had clear scientific goals and objectives, Alexa users played the key role of providing feedback on the socialbots, helping teams improve their systems and helping us determine which socialbots were the most coherent and engaging. Because users helped drive the direction and result of the competition, it was important for us to ensure an easy and compelling hook into the Alexa Prize socialbots to obtain a statistically significant number of data points for the ratings and feedback needed to improve the socialbots.

Eighteen teams were selected, although only 15 went live. To allow us to randomize traffic to all 15 socialbots without revealing their identity and to set user expectations about the socialbots being early-stage systems, we designed and implemented the Alexa Prize skill with a natural invocation phrase that was easy to remember ("Alexa, let's chat," "Alexa,

let's chat about <topic>," and common variants). The user heard a short editorial that educated them about the Alexa Prize and instructed them on how to end the conversation and provide ratings and feedback. We kept it succinct and interesting so we would not lose the user's attention before they had a chance to speak with a socialbot. The editorial and instructions changed as needed to keep the information relevant to the different phrases. For example, at the initial public launch on May 8, 2017, when the 15 socialbots were still in their infancy, the Alexa Prize skill started with the following editorial: "Hi! Welcome to the Alexa Prize Beta. I'll get you one of the socialbots being created by universities around the world. When you're done chatting, say stop."

After listening to the editorial, the user was handed off to one of the competing socialbots selected at random. Socialbots began the conversation with a common introduction phrase ("Hi, this is an Alexa Prize socialbot") without revealing their identity. The user could exit at any time, and thereafter was prompted to provide a verbal rating ("On a scale from one to five stars, how do you feel about speaking with this socialbot again?"). Finally, the user could offer free-form verbal feedback without knowing which socialbot they had interacted with. Ratings and feedback were provided to individual teams to help them improve their socialbots.

Challenges with Conversational AI

Current state-of-the-art systems are still a long way from engaging in truly natural everyday conversations with humans (Levesque 2017). There are a number of major challenges associated with building conversational agents: conversational automatic speech recognition for free-form multiturn speech; conversational natural language understanding for multiturn dialogues; conversational datasets and knowledge ingestion for comprehension; commonsense reasoning for understanding concepts; context modeling for relating past concepts; dialogue planning for driving coherent and engaging conversations; response generation and natural language generation for generating relevant, grammatical, and nongeneric responses; sentiment detection for systematically identifying, extracting, quantifying, and studying affective states and for handling sensitive content (such as profanity, inflammatory opinions, inappropriate jokes, hate speech detection), driving quality conversations; personalization for addressing user preferences; conversation evaluation for evaluating the quality of the conversations and the artificial agent; and conversational experience design for maintaining a great experience for the interactors.

Most voice-based conversational agents follow a similar architecture. First, the agent comprehends speech signals and converts them to text by a process called automatic speech recognition (ASR). After obtaining the text, the agent tries to understand the meaning and intention of the user using existing knowledge by a process called natural language understanding (NLU). Once the concepts and intents are identified, the agent starts the process of response generation (RG), which involves identifying relevant responses based on context, knowledge, personalization, and some form of planning. Planning may involve optimizing for some reward such as sentiment, serving the goal in a goal-directed dialogue, or increasing user engagement. This process can be managed by a dialogue manager (DM), which acts as an engine for maintaining the state and flow of the conversation. Finally, once the output response is produced in textual form, the agent converts it to speech by a process called text to speech (TTS).

Using these techniques, academia and industries have created virtual assistants to support short, task-oriented dialogues such as playing music or asking for information. Some assistants are capable of longer multiturn dialogues, although most of these are goal directed or designed for specific tasks such as customer support or shopping (for example, eBay's ShopBot). Furthermore, these systems tend to be text based and not capable of natural voice conversation. Long, free-form voice conversations that occur naturally in social and professional human interactions are often open domain. In natural conversations, intents and topics change with time based on the interest of the interactors and the state of the conversation. Furthermore, natural conversations feature many plausible responses at each turn and are highly path dependent: even if two sets of interactors have a similar background and share a similar set of knowledge, they may end up having completely different conversations.

The remainder of this article describes how the challenges listed at the beginning of this section were addressed and what results were obtained.

Addressing Problems in Creating Conversational Agents

The Alexa Prize team developed components to provide conversational speech recognition, conversational intent tracking, conversational topic tracking, inappropriate and sensitive content detection, and conversational quality evaluation. In addition, the team addressed engineering challenges such as traffic allocation, socialbot scalability, socialbot invocation, and a feedback framework. These components and solutions enabled live deployment to a large user base consisting of millions of Alexa customers. To build the socialbots, university teams combined state-of-the-art techniques with their own novel strategies in the areas of natural language understanding, context modeling, dialogue management, response generation, ranking and selection, sentiment analysis, and knowledge acquisition. Subse-

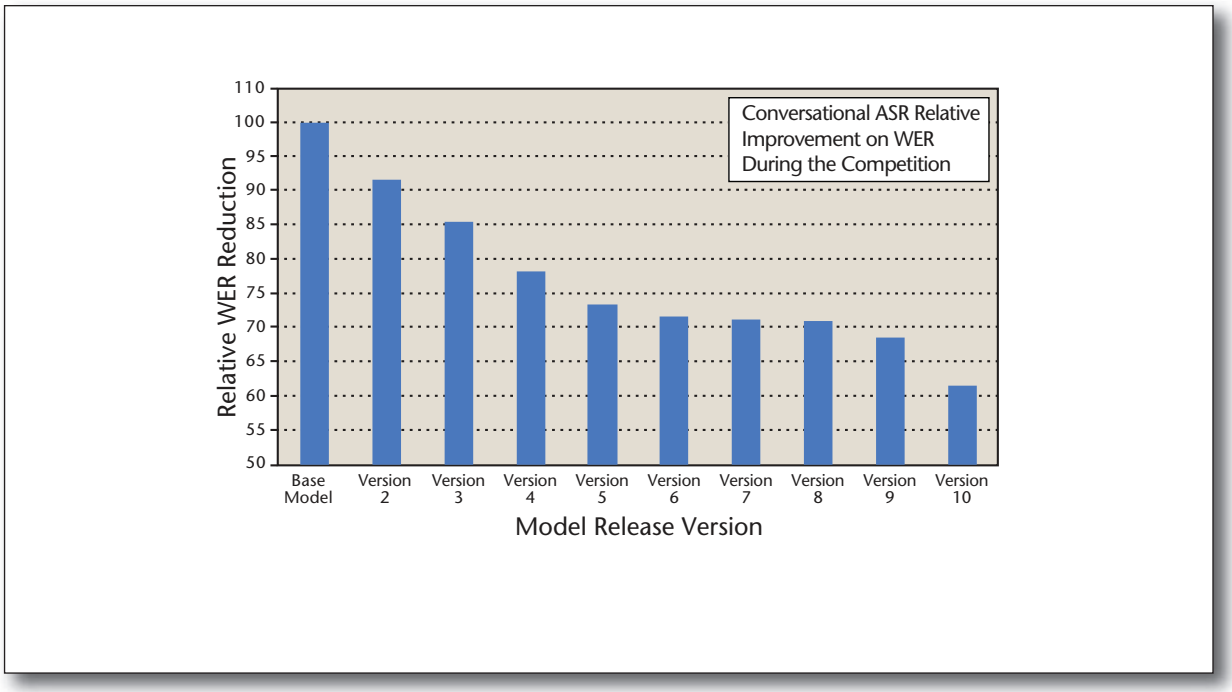


Figure 1. Conversational ASR Performance Improvement.

Relative reduction in WER with respect to base model.

quent sections describe the state of each of these problems and the advancements produced by all these teams.

Conversational Automatic Speech Recognition (ASR)

Speech is the gateway to voice-based agents, and errors in speech recognition can get propagated to later stages such as NLU and the dialogue manager, leading to incorrect or incoherent responses.

ASR is even more difficult in open-domain and non-task-oriented conversational agents. Free-form speech does not necessarily fit a command-like structure. It typically contains longer sentences, and the space of plausible word combinations is much larger. In addition, social conversations are informal, and open ended, they contain many topics, and they have a high out-of-vocabulary rate. Furthermore, production-grade ASR approaches must deal with a much wider array of noise and environmental conditions than the conditions in the normalized research datasets often reported in the literature. All of these make conversational ASR a challenging problem.

We developed a custom language model (LM) targeted specifically at open-ended conversations with socialbots. We initially used publicly available conversational datasets such as Fisher, Switchboard, Reddit comments, Linguistic Data Consortium (LDC) news, OpenSubtitles, and Yelp reviews, along with the Washington Post news data, for building this conversational language model. As the competition progressed, we incrementally added data from Alexa Prize utterances collected over the course of the competition to our dataset. Performance of the model improved significantly on conversational test sets over the course of the competition. Figure 1 provides the relative improvement in word error rate (WER). We see a reduction by nearly 40 percent relative to the base model.

Conversational Natural Language Understanding (NLU)

Understanding the input of an interactor during a conversation is critical for dialogue systems. If machines cannot comprehend a user's intent or the topics and entities mentioned in an utterance, then the machine will not be able to respond well, which will lead to a poor customer experience. NLU in goal-oriented dialogue systems follows a domain, an intent, and a slot-like approach (Kumar et al. 2017). For example, in the utterance "Play Havana from Camila Cabello," the domain is music, the intent is to play a song, and the entities (modeled as slots) are "song_name: Havana, artist: Camila Cabello." These approaches work well in goal-oriented dialogue systems; however, open-domain dialogue systems are

free-form and are not confined to a predefined set of domains, intents, or entities. The intent or domain may be unclear and the slots may not be well defined. For example, consider the following utterance: "Last night, I went to Justin Bieber's show. He was great, but the crowd was not. Do you think should I go next time?" The goal of the utterance is not well defined. The utterance consists of multiple intents: information delivery, opinion sharing, and opinion request. Furthermore, there are multiple slots in the same utterance. Traditional NLU systems do not work well for natural conversations. The following NLU components were developed during the Alexa Prize competition to address these challenges.

Conversation Intent

To connect an Alexa user with a socialbot, we first needed to identify whether the user's intent was to have a conversation with Alexa. We introduced a "conversation intent" within the Alexa NLU model to recognize a range of utterances such as "let's chat," "let's talk," "let's chat about <topic>," and so forth, using a combination of grammars and statistical models. We further expanded the experience to other natural forms of conversational initiators such as "what are we going to talk about," "can we discuss politics," "do you want to have a conversation about the Mars Mission," and so on.

In the production system, if an utterance from an Alexa user is identified as a conversation intent, then one of the Alexa Prize socialbots is invoked and the user interacts with that socialbot until the user says stop. Following the detection of conversational intent, the entire conversation is controlled by the socialbots. Teams used a combination of Alexa Skills Kit NLU along with their own NLU approaches, as will be described.

NLU Techniques Adopted by the Participating Teams

After a user has initiated a conversation, the socialbot requires an NLU system to identify semantic and syntactic elements from the user utterance, including user intent (such as opinion, chit-chat, knowledge, and others), entities and topics (for example, the entity "Mars Mission" and the topic "space travel" from the utterance "what do you think about the Mars Mission"), user sentiment, as well as sentence structure and parse information. A certain level of understanding is needed to generate responses that align well with user intent and expectation and to maximize user satisfaction. Although conversational utterances do not generally follow intent-slot structure, teams brought several workarounds to address this problem.

NLU is difficult because of the inherent complexities within the human language, such as anaphora, elision, ambiguity, and uncertainty, which require contextual inference in order to extract the necessary information to formulate a coherent response. These problems are magnified in conversational AI since it is an open domain problem where a conversation can be on any topic or entity and the content of the dialogue can also change rapidly. Some specific techniques used by teams are listed in the following paragraphs.

Named Entity Recognition (NER)

Identifying and extracting entities (names, organizations, locations) from user utterances. Teams used various libraries such as StanfordCoreNLP (Manning et al. 2014), pacy,[1] and Alexa's ASK NLU to perform this task. NER is helpful for retrieving relevant information for response generation, as well as for tracking conversational context over multiple turns.

Intent Detection

Intents represent the goal of a user for a given utterance, and the dialogue system needs to detect it to act and respond appropriately to that utterance. Some of the teams built rules for intent detection or trained models in a supervised fashion by collecting the data from Amazon Mechanical Turk or by using open source datasets, such as Reddit comments, with a set of intent classes. Others utilized Alexa's ASK NLU engine for intent detection.

Anaphora and Coreference Resolution

Finding expressions that refer to the same entity in past or current utterances. Anaphora resolution is important for downstream tasks such as question answering and information extraction in multiturn dialogue systems. Most of the teams used StanfordCoreNLP's Coreference Resolution System (Manning et al. 2014) to perform this task.

Sentence Completion

Some teams expanded user utterances with contextual information. For example, "Yes" can be transformed to "Yes, I like Michael Jackson" when uttered in the context of a question about the singer, or "I like Michael Jackson" can be extended to "I like Michael Jackson, singer and musician" in a conversation where this entity needs disambiguation. Teams wrote customized wrappers for preforming sentence completion, which also involves querying knowledge bases to obtain more information about the entities, as described in the entity-linking section.

Topic and Domain Detection

Classifying the topic (for example, Seattle Seahawks) or domain (such as sports) from a user utterance. Teams used various datasets to train topic detection models, including news datasets, Twitter, and Reddit comments. Some teams also collected data from Amazon Mechanical Turk to train these models.

Entity Linking

Identifying information about an entity. Teams generally used publicly available knowledge bases such as Evi,[2] FreeBase (Bollacker et al. 2008), and Wikidata.[3] Some teams also used these knowledge bases to identify related entities.

Text Summarization
Extracting or generating key information from documents for efficient retrieval and response generation. Some of the teams adopted this technique for summarizing the articles or potential responses for efficient response generation.

Sentiment detection
Identifying user sentiment. Some teams developed sentiment detection modules to help with generating engaging responses. This approach also helped them to better understand a user's intent and generate appropriate responses.

Knowledge Ingestion and Common Sense Reasoning

Currently, available conversational data is limited to datasets that have been produced from online forums (for example, Reddit), social media interactions (for example, Twitter), and movie subtitles (for example, OpenSubtitles, Cornell Movie-Dialogs Corpus). While these datasets are useful at capturing the syntactic and semantic elements of conversational interactions, they also have many issues with data quality, content (profanity, offensive data), query-response pair tracking, context tracking, multiple users interacting without a specific order, and short and ephemeral conversations. In the absence of better alternatives, teams still used these datasets. To address offensive content and profanity, teams built classifiers to detect this content. Furthermore, we shared *Washington Post* (WaPo) live comments, which are conversational in nature and also highly topical. Several teams made use of these comments.

The teams also used various knowledge bases, including Amazon's Evi, Freebase, and Wikidata for retrieving general knowledge, facts, and news, and for general question answering. Some teams also used these sources for entity linking, sentence completion, and topic detection. Ideally, a socialbot should be able to ingest and update its knowledge base automatically; however, this is an unsolved problem and an active area of research. Finally, teams also ingested information from news sources such as the Washington Post and CNN to keep current with news events that users may want to chat about.

For commonsense reasoning, several teams built modules to understand user intent. Some teams preprocessed open source and Alexa Prize datasets and extracted information about trending topics and opinions on popular topics, integrating them within their dialogue manager to make the responses seem as natural as possible. To complement commonsense reasoning, some of the top teams added user satisfaction modules to improve both engagement and conversational coherence.

To make sure that teams were leveraging relevant datasets and knowledge bases, we emphasized early availability of live user interactions to the socialbots, which helped the teams in identifying relevant data sources before the competition went live.

Dialogue and Context Modeling

A key component of any conversational agent is a robust system to handle dialogues effectively. The system should accomplish two main tasks: help break down the complexity of the open domain problem to a manageable set of interaction modes, and be able to scale as the diversity and breadth of topics expands. A common dialogue strategy used by teams was a hierarchical architecture with a main dialogue manager (DM) and multiple smaller DMs corresponding to specific tasks, topics, or contexts.

Some teams, such as Sounding Board, used a hierarchical architecture and added additional modules such as an error handler to handle cases such as low-confidence ASR output or low-confidence response candidates (Fang et al. 2017).[4] Other teams, such as Alquist, (Pichl, J. et al. 2017)[4] used a structured topic-based dialogue manager, where components were broken up by topics, along with intent-based dialogue modules broken up by intents. Generally, teams also incorporated special-purpose modules such as a profanity or offensive content module to filter a range of inappropriate responses and modules to address feedback and acknowledgement and to request clarity or rephrasing from users. Teams experimented with approaches to track context and dialogue states, and corresponding transitions to maintain dialogue flow. For example, Alquist and Slugbot (Bowden et al. 2017)[4] modeled dialogue flow as a state graph. These and other techniques helped socialbots produce coherent responses in an ongoing multiturn conversation and guided the direction of the conversation as needed. A few teams, such as Magnus (Prabhumoye et al. 2017),[4] built finite-state machines (FSMs) (Wright 2005) for addressing specific modules such as movies, sports, and others. One challenge in using this technique for dynamic components is scaling and context switching; however, for small and static modules, FSMs can be useful.

The top teams focused not only on response generation but also on customer experience, and experimented with conversational strategies to increase engagement as discussed in the next section.

Conversational User Experience

Participating teams built several conversational user experience (CUX) modules, which included engagement, personalization, and other user experience–related aspects. CUX modules are relatively easy to build, but such modules may lead to significant gains on ratings and duration. CUX is an essential component, and the teams that focused most of their efforts on NLU and DM, with less emphasis on CUX, were not received as top performers by Alexa users. Following are the five main components built by various teams.

Personalization

Socialbots received an obfuscated (by one-way hash function) user ID to enable personalization for repeat users while maintaining user privacy. Alquist built a personalization module to remember past interactions and users. This module was used to give a more natural and personal touch in initiating conversations. Sounding Board tried multiple strategies to get more information about user preferences on topics. For example, they developed a personality quiz that enabled them to tailor topics based on the user's personality. They found that extroverts, as determined by the personality quiz, correlated with higher ratings and longer turns. Edina added a level of personalization to their DM to track whether a certain topic or type of response is doing particularly well with a user (Krause et al. 2017).[4]

Topic Switching

Alana used a multibot strategy consisting of data-driven bots and rule-based bots, including Eliza/Template, Persona, Quiz Game, NewsBot, Factbot, Evi, and Weatherbot. When presenting the information from a submodule, for example, NewsBot, Alana determined whether to keep or change the topic based on user feedback. Edina identified topical drift to recognize when the customer wants to set a new topic or keep the current one. Emerson focused on machine-driven conversation through a topic-recommendation mechanism for conversation topic transition.

Initiative

Initiative in conversation is another key variable — should the bot guide a user to certain topics of conversation, let the user steer topics, or mix these approaches? SlugBot designed a system-initiative module to direct the conversation through stories, games, and informing the user of various headlines. Edina collected data on various topics through Amazon Mechanical Turk (AMT) and built proactive modules to drive conversation with users. Emerson investigated four levels of initiative: passive, less passive, active, and less active. Based on their experiments, the Emerson team concluded that conversations actively driven by bots give the best performance on rating and duration. ChattyChat used machine-initiated dialogues to drive conversation by asking yes or no questions along with topic suggestion (Yi and Jung 2017).

Sentiment-Based Modules

Sounding Board gauged user reaction through sentiment analysis and user decisions on certain cases, particularly opinion-related utterances. They developed a dissatisfaction detector by training a classifier to detect when a user expresses discontent, which was used to trigger a change of topic. Magnus built a classifier to filter abusive content. RubyStar had a module to avoid explicit topics such as pornography, as well as other sensitive subjects. In such cases, RubyStar responded with predefined templated responses.

Engagement and Greeting Modules

Responses that engage and captivate a user, entertaining them and leaving them wanting to keep talking, are critical to developing great conversation. RubyStar used an approach called engagement reranking that was trained on Reddit comments (Liu et al. 2017).[4] They split comments into engaging (high number of upvotes) and nonengaging (low number of upvotes). Roving Mind considered three semantic dimensions (Cervone et al. 2017)[4] proposed in Dialogue Act Markup Language (Bunt et al. 2010): social obligation (addressing basic social conventions such as greetings and feedback to user); addressing user feedback to a machine's statement; and task (addressing user actions). WiseMacaw developed a two-layer chatbot framework that can be described as a metagame, where the user could walk to multiple gaming modules (Ji et al. 2017).[4]

Several teams started adding games, quizzes, and related modules in their interactions, which led to significant increase in ratings and duration with certain segments of users. However, such interactions are not necessarily conversational; furthermore, they did not advance the state of conversational AI, which was the main objective of this competition. To address these issues, we guided teams to remove such modules and eliminated their inclusion in the final round of the competition.

Response Generation

There are four main types of approaches for response generation in dialogue systems: template/rule-based, retrieval, generative, and hybrid. A functional system can be an ensemble of these techniques. It can follow a waterfall structure, for example, rules —> retrieval —> generative. Or it can use a hybrid approach with complementary modules, for example, generative models for retrieval, or generative models to create templates for a retrieval or rule-based module. Some of the teams used AIML, ELIZA (Weizenbaum 1966), or Alicebot[5] for rule-based and templated responses. Teams also built retrieval-based modules that tried to identify an appropriate response from the dataset of dialogues available. Retrieval was performed using techniques such as n-gram matching and entity matching, or using similarity metrics based on vectors such as TF-IDF, word/sentence embeddings, skip-thought vectors, and dual-encoder systems.

Hybrid approaches leveraging retrieval in combination with generative models are fairly new and have shown promising results in the past couple of years, usually with sequence-to-sequence approaches with some variants. Some of the Alexa Prize teams created novel techniques along these lines and demonstrated scalability and relevance for the open-domain, response-generation models deployed in production systems. MILABot, for example, devised a hierarchical latent variable

encoder-decoder (VHRED) (Serban et al. 2016) model, in addition to other neural network models such as skip thought (Kiros et al. 2015) to produce hybrid retrieval-generative candidate responses. Some teams (such as Pixie)(Adewale et al. 2017)[4] used a two-level, long short-term memory (LSTM) model (Hochreiter and Schmidhuber 1997) for retrieval. Eigen (Guss et al. 2017)[4] and RubyStar, on the other hand, used dynamic memory networks (Sukhbaatar et al. 2015) and character-level recursive neural networks (RNN) (Sutskever, Vinyals, and Le 2014) for generating responses. Alquist used a sequence-to-sequence model (Sutskever, Vinyals, and Le 2014) specifically for their chit-chat module. While these teams deployed the generative models in production, other teams also experimented with generative and hybrid approaches offline.

Ranking and Selection Techniques

Open-domain social conversations do not always have a specific goal or target, and the response space can be unbounded. There may be multiple valid responses for a given utterance. As such, identifying the response that will lead to the highest customer satisfaction and help drive the conversation forward is a challenging problem. Socialbots need mechanisms to rank possible responses and select the response that is most likely to achieve the goal of a coherent and engaging conversation in that particular dialogue context. Alexa Prize teams attempted to solve this problem with either rule-based or model-based strategies.

For teams that experimented with rule-based rankers, a ranker module chose a response from the candidate responses obtained from submodules (such as topical modules or intent modules) based on some logic. For model-based strategies, teams utilized either a supervised or reinforcement learning approach, trained on user ratings (Alexa Prize data) or on predefined large-scale dialogue datasets such as Yahoo! Answers, Reddit comments, Washington Post Live comments, and OpenSubtitles. The ranker was trained to provide higher scores to correct responses (for example, follow-up comments on Reddit are considered correct responses) while ignoring incorrect or noncoherent responses obtained by sampling. Alan (Papaioannou et al. 2017),[4] for example, trained a ranker module on Alexa Prize ratings data and combined that with a separate ranker function that used hand-engineered features. Teams using a reinforcement learning approach developed frameworks where the agent was a ranker, the actions were the candidate responses obtained from submodules, and the agent was trying to maximize the trade-off between selecting a response to satisfy the customer immediately and selecting one that takes into account some long-term reward. MILABot, for example, used this approach and trained a reinforcement learning ranker function on conversation ratings.

The afore-mentioned components form the core of socialbot dialogue systems. In addition, we developed the following components to support the competition.

Conversational Topic Tracker

To understand the intent of a user, it is important to identify the topic of the given utterance and corresponding keywords. Alexa Prize data is highly topical because of the nature of the social conversations. Alexa users interacted with socialbots on hundreds of thousands of topics in various domains such as sports, politics, entertainment, fashion, and technology. This is a unique dataset collected from millions of human–conversational agent interactions. We identified the need for a conversational topic tracker for various purposes such as conversation evaluation (for example, coherence, depth, breadth, diversity), sentiment analysis, entity extraction, profanity detection, and response generation.

To detect conversation topics in an utterance, we adopted deep average networks (DAN) and trained a topic classifier on interaction data categorized into multiple topics. We proposed a novel extension by adding topic-word attention to formulate an attention-based DAN (ADAN) (Guo et al. 2017) that allows the system to jointly capture topic keywords in an utterance and perform topic classification. We fine-tuned the model on the data collected during the course of the competition. The accuracy of the model was obtained to be 82.4 percent on 26 topical classes (sports, politics, movies_TV, and so on). Furthermore, the topic model was also able to extract keywords corresponding to each topic. We used the conversational topic tracker to evaluate the socialbots on various metrics such as conversational breadth, conversational depth, and topical and domain coverage (Venkatesh et al. 2017; Guo et al. 2017). We will explore additional details on evaluation later in this article.

Inappropriate and Sensitive Content Detection

One of the most challenging aspects of delivering a positive experience to end users in socialbot interactions is to obtain high-quality conversational data. The datasets most commonly used to train dialogue models are sourced from internet forums (for example, Reddit, Twitter) or movie subtitle databases (for example, OpenSubtitles, Cornell Movie-Dialogs Corpus). These sources are all conversational in structure in that they can be transformed into utterance-response pairs. However, the tone and content of these datasets are often inappropriate for interactions between users and conversational agents, particularly when individual utterance-response pairs are taken out of context. In order to effectively use dialogue models based on these or other dynamic data sources, an efficient mechanism to identify and filter

different types of inappropriate and offensive speech is required.

We identified several (potentially overlapping) classes of inappropriate responses: (1) profanity, (2) sexual responses, (3) racially offensive responses, (4) hate speech, (5) insulting responses, and (6) violent responses (inducements to violent acts or threatening responses). We explored keyword- and pattern-matching strategies, but these strategies are subject to poor precision (with a broad list) or poor recall (with a carefully curated list), as inappropriate responses may not necessarily contain profane or other blacklisted words. We tested a variety of support vector machines and Bayesian classifiers trained on n-gram features using labeled ground truth data. The best accuracy results were in profanity (>97 percent at 90 percent recall), racially offensive responses (96 percent at 70 percent recall), and insulting responses (93 percent at 40 percent recall). More research is needed to develop effective offensive speech filters. In addition to dataset cleansing, an offensive speech classifier is also needed for online filtering of candidate socialbot responses prior to outputting them to ASK for text-to-speech conversion.

Addressing Problems in Evaluating Conversational Agents

Social conversations are inherently open ended. For example, if a user asks the question "What do you think of Barack Obama?," there can be thousands of distinct, valid, and reasonable responses. That is, the response space is unbounded for open-domain conversations. This makes training and evaluating social, non-task-oriented, conversational agents extremely challenging. It is easier to evaluate a task-oriented dialogue system because we can measure systems by successful completion of tasks, which is not the case with open-ended systems. As with human-to-human dialogues, an interlocutor's satisfaction with a socialbot could be related to how engaging, coherent, and enjoyable the conversation was. The subjectivity associated with evaluating conversations is a key element underlying the challenge of building non-goal-oriented dialogue systems.

This problem has been heavily studied but lacks a widely agreed-upon metric. A well-designed evaluation metric for conversational agents that addresses the above concerns would be useful to researchers in this field. There is significant previous work on evaluating goal-oriented dialogue systems. Two of those notable earlier works are TRAINS system and PARADISE (Walker et al. 1997). All of these systems involve some subjective measures that require a human in the loop. Due to the expensive nature of human-based evaluation procedures, researchers have been using automatic machine translation (MT) metrics, such as BLEU, or text summarization metrics, such as ROUGE, to evaluate systems. But as shown by Liu et al. (2017), these metrics do not correlate well with human expectations.

The Turing Test (Turing 1950) is a well-known test that can potentially be used for dialogue evaluation. However, we do not believe that the Turing Test is a suitable mechanism to evaluate socialbots for the following reasons:

Incomparable elements: Given the amount of knowledge an AI has its disposal, it is not reasonable to suggest that a human and an AI should generate similar responses. A conversational agent may interact differently from a human, but may still be a good conversationalist.

Incentive to produce plausible but low-information content responses: If the primary metric is just generation of plausible human-readable responses, it is easy to opt out of the more challenging areas of response generation and dialogue management. It is important to be able to source interesting and relevant content while generating plausible responses.

Misaligned objectives: The goal of the judge should be to evaluate the conversational experience, not to attempt to get the AI to reveal itself.

To address these issues, we propose a comprehensive, multimetric evaluation strategy designed to reduce subjectivity by incorporating metrics that correlate well with human judgement. The proposed metrics provide granular analysis of the conversational agents, which is not captured in human ratings. We show that these metrics can be used as a reasonable proxy for human judgment. We provide a mechanism to unify the metrics for selecting the top performing agents, which has also been applied throughout the Alexa Prize competition. The following objective metrics (Guo et al. 2017; Venkatesh et al. 2017) have been used for evaluating conversational agents. The proposed metrics also align with the goals of a socialbot, that is, the ability to converse coherently and engagingly about popular topics and current events.

Conversational user experience (CUX): Different users have different expectations concerning the socialbots, and so their experiences might vary widely since open-domain dialogue systems involve subjectivity. To address these issues, we used average ratings from frequent users as a metric to measure CUX. With multiple interactions, frequent users have their expectations established and they evaluate a socialbot in comparison to others.

Coherence: We annotated hundreds of thousands of randomly selected interactions for incorrect, irrelevant, or inappropriate responses. With the annotations, we calculated the response error rate (RER) for each socialbot, using that figure to measure coherence.

Engagement: Evaluated through performance of conversations identified as being in alignment with socialbot goals. Measured using duration, turns, and ratings obtained from engagement evaluators (a set of Alexa users who were asked to evaluate socialbots based on engagement).

Algorithm	RMSE	Spearman	Pearsonr
Random	2.211	0.052	0.017
HLSTM	1.392	0.232	0.235
GBDT	1.34	0.352	0.351

Table 1. Correlation of the Regression Model with User Ratings.

Domain coverage: Entropy analysis of conversations against the five socialbot domains for Alexa Prize (Sports, Politics, Entertainment, Fashion, Technology). Performance was targeted on high entropy, while minimizing the standard deviation of the entropy across multiple domains. High entropy ensures that the socialbot is talking about a variety of topics, while a low standard deviation gives us confidence that the metric is applied equally across domains.

Topical diversity: Obtained using the size of topical vocabulary for each socialbot. A higher topical vocabulary within each domain implies more topical affinity.

Conversational depth: We used the topical model to identify the domain for each individual utterance. Conversational depth for a socialbot was calculated as the average of the number of consecutive turns on the same topical domain, where single turn corresponds to user utterance and corresponding bot response pair within a conversation. Conversational depth evaluates the socialbot's ability to have multiturn conversations on specific topics within the five domains.

Selecting Alexa Prize Finalists

The Alexa Prize competition was structured to allow users to participate in the selection of finalists. Two finalists were selected purely on the basis of user ratings averaged over all the conversations with those socialbots. At the end of the conversation, users were asked to rate how coherent and engaging the conversation was.

In addition, one finalist was selected by Amazon based on internal evaluation of coherence and engagement of conversations by over one thousand Amazon employees who volunteered as Alexa Prize judges, on analysis of conversational metrics computed over the semifinals period, and on scientific review of the team's technical papers by senior Alexa scientists. The quality of all the socialbots was also analyzed based on the metrics mentioned above. We observed that a majority of those metrics correlate well with user ratings, frequent ratings, and ratings from Alexa Prize judges, with a correlation coefficient greater than 0.75. A simple combination of the metrics correlated strongly with Alexa user ratings (0.66), suggesting that the "wisdom of crowds" (Surowiecki 2004) is a reasonable approach to evaluation of conversational agents when conducted at scale in a natural setting. The average rating across all socialbots was lower by 20 percent for the judge's pool as compared with the general public.

Teams also evaluated the quality of their socialbots and made necessary improvements during the competition by leveraging the ratings and feedback from users. Alexa users had millions of interactions and over 100,000 hours of conversations with socialbots throughout the duration of the competition.

Automatic Evaluation of Open-Domain Conversations

If we are able to build a model that can predict the rating of an Alexa Prize conversation with reasonable accuracy, then it is possible to remove humans from the loop for evaluating non-task-oriented dialogues.

To automate the evaluation process, we did a preliminary analysis on 60,000 conversations and ratings, and we trained a model to predict user ratings. We observed the Spearman and Pearson correlations of 0.352 and 0.351 respectively (table 1) with significantly low p-value with a model trained using a gradient-boosted decision tree (GBDT). Although the results for GBDT are significantly better than random selection for five classes and the model trained using hierarchical LSTM, there is a need to extend this study to millions of Alexa Prize interactions. Furthermore, some of the evaluation metrics (coherence, topical depth, topical breadth, domain coverage) obtained at conversation level can also be used as features. With a significantly higher number of conversations combined with topical features, we hypothesize that the model would perform much better than the results obtained in the preliminary analysis shown in table 1. Given subjectivity in ratings, we appropriately found interuser agreement to be quite low for ratings analysis. Users may have their own

criteria to evaluate the socialbots. Therefore, as a part of the future work, we will train the model with user-level features as well.

The Alexa Prize Finals

Following the conclusion of the semifinals, three finalists were identified: Alquist (Czech Tech University), Alana (Heriot-Watt University), and Sounding Board (University of Washington). These finalists remained online for the remainder of the competition. They entered the last phase of the competition with an average rating of 2.77. After two months of additional interaction with Alexa users, they went to the finals with an average rating of 3.48, an improvement of 26 percent from their earlier ratings. The Alexa Prize finals were conducted in November as a closed-door event with three rounds of conversations with each socialbot. We invited professional judges well versed in conversation: talk show hosts, authors, speakers, and AI experts. Three judges were "conversationalists" and asked to engage in conversations with the socialbots; five judges were "evaluators" and asked to listen to and rate the conversations. The winner was Sounding Board from the University of Washington with an average score of 3.17 on a 5-point scale and an average conversation duration of 10 minutes and 22 seconds.

The objective of the finals was to evaluate socialbots on their ability to conduct coherent and engaging voice conversations with humans and to chain relevant entities into a dialogue. Highlights of these conversations are available in a short film.[6] The best interactions combined many of the techniques discussed in this article.

For example, a conversation that started off with a request to discuss the book *The Dark Tower* evolved in the following manner: The Dark Tower (book) → The Dark Tower (movie) → Matthew McConaughey's performance in the movie → Other movies that he's acted in and that the user may find interesting → a topical humorous suggestion, "Someone needs to make a mashup of *Interstellar* and *Contact* where Matthew McConaughey tries to prevent Matthew McConaughey from going into space." Such relevant transitions and natural injection of humor give us a glimpse into the future of conversational AI.

Another example that combines entity resolution and reference disambiguation in the context of an introductory ice-breaker experience:

Socialbot: Hi. This is an Alexa Prize socialbot. How's it going?

User: It's going great, how about you?

Socialbot: Awesome. I'm glad you're doing well today. I'm having a good day, thanks for asking. Ready to enjoy our chat. If you like, you can tell me your name?

User: My friends call me John, but you can call me Mr. Pickles.

Socialbot: Excellent, I'm happy to meet you Mr. Pickles.

There were some examples of failure points as well (such as ASR failures). The field has a long way to go yet, but there is definite promise and improvement in the accomplishments of this first year, and we look forward to seeing more gains made through Alexa Prize 2018.

The finalist socialbots will continue to be a part of the "let's chat" experience until the Alexa Prize 2018 socialbots become available to the public in May 2018. To speak with the socialbots yourself, simply say "Alexa, let's chat" to any Alexa-enabled device.

Results

The Alexa Prize was designed as a framework to support research on conversational AI at scale in a real-world setting. The scientific advances described above (and detailed in individual team papers) resulted in significant improvements in socialbot quality and a significant amount of user engagement.

User Engagement

Customer engagement remained high throughout the competition. Alexa Prize ranked in the top 10 Alexa skills by usage, with over 40,000 hours of conversations spanning millions of utterances by the end of the finals. Customers chatted on a wide range of popular and current topics with movies/TV, music, politics, celebs, business, and scitech being the highest frequency (most popular) topics. The most popular topics from the post-semifinals feedback phase were movies/TV (with an average rating of 3.48), scitech (3.60), travel/Geo (3.51), and business (3.48). Based on user ratings, the three lowest rated topics were arts (with an average rating of 2.14), shopping (2.63), and education (3.03).

It is still early in the Alexa Prize journey towards natural human conversation, but the high level of engagement and feedback (over 130,000 hours of conversation to date) demonstrates that users are interested in chatting with socialbots and supporting their development.

Socialbot Quality

Over the course of the competition, socialbots showed a significant improvement in customer experience. The three finalists improved their ratings by 29.6 percent (from 2.77 to 3.59) over the duration of competition. All 15 socialbots had an average customer rating of 2.87, with a median conversation duration of 1:35 minutes and a 90th percentile of 5:43 minutes by the end of the semifinal phase. The conversation duration of finalists across the entire competition was 1:41 minutes (median) and 8:02 minutes (90th percentile), improving 19.4 percent and 58.26 percent respectively from the start of the competition, with 10 turns (median) per conversation.

We measured response error rate (RER) through the

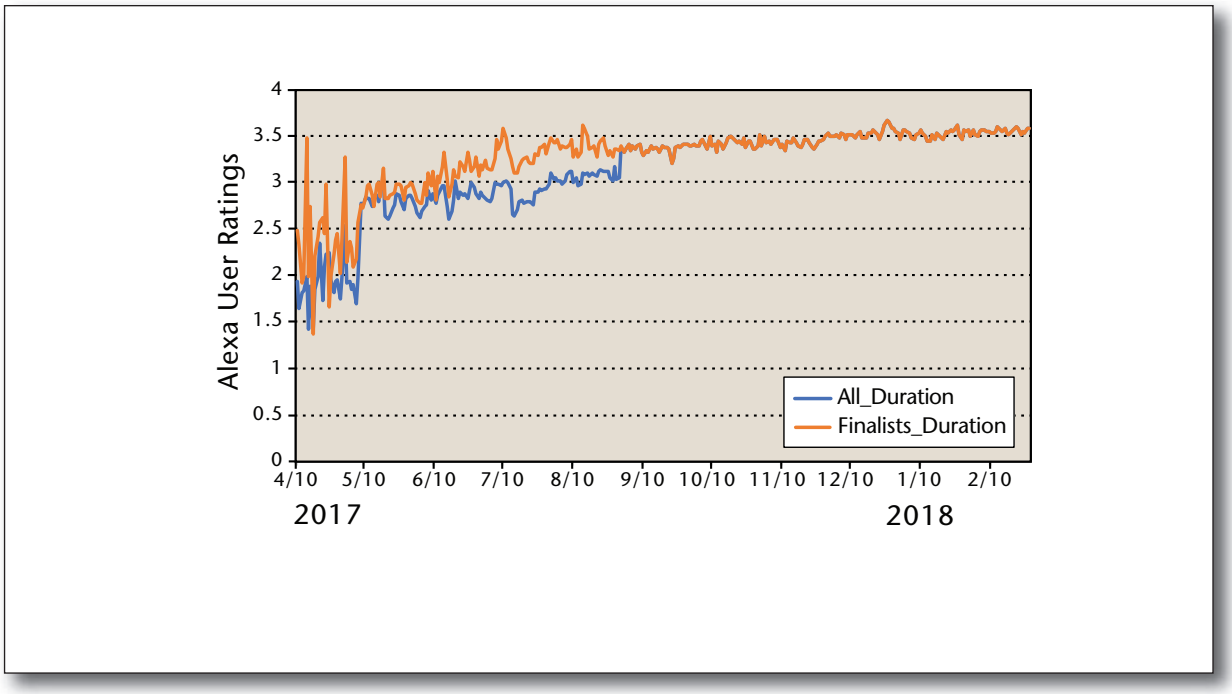


Figure 2. Daily Ratings for Socialbots During the Competition.

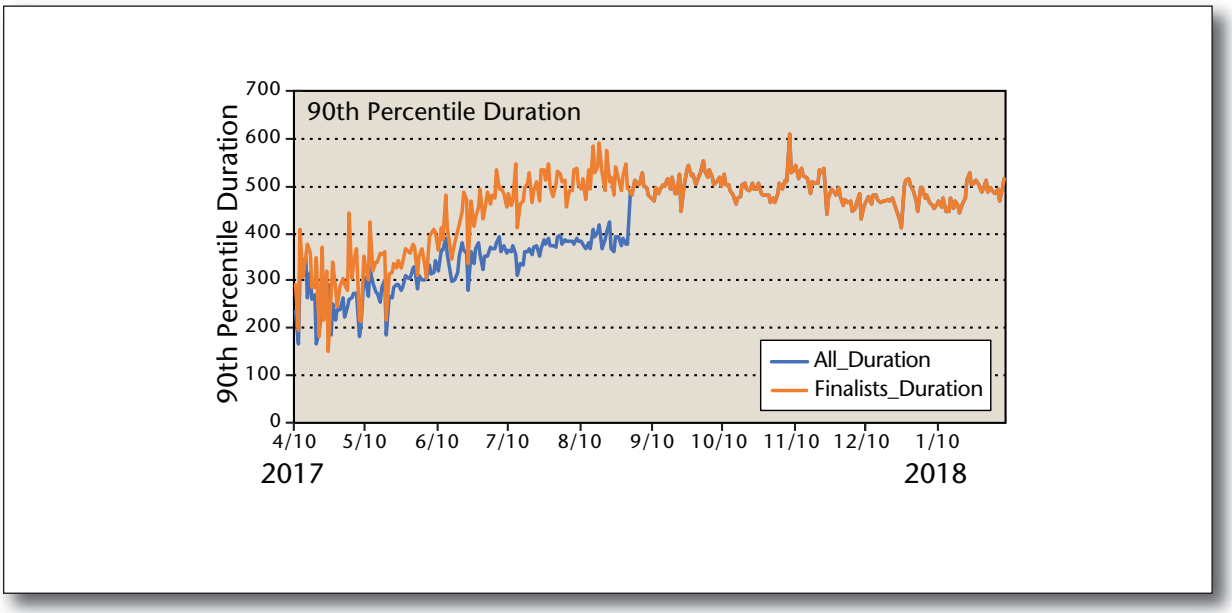


Figure 3. Conversation Duration Median and 90th Percentile for Socialbots During the Competition.

manual annotation of a large fraction of user utterance-social response pairs by trained data analysts to identify incorrect, irrelevant, or inappropriate responses. RER was quite irregular during the first 30 days of the launch, from May 8 to June 8, 2017, fluctuating between 8.5 percent and 36.7 percent. This fluctuation was likely due to rapid experimentation by teams in response to initial user data. During the semifinals, from July 1 to August 15, 2017, RER was in the range of 20.8 percent to 28.6 percent. The three finalists improved further over the post-semifinals feedback phase, and their average RER was at 11.21 percent (L7D: last 7 days) as they went into finals.

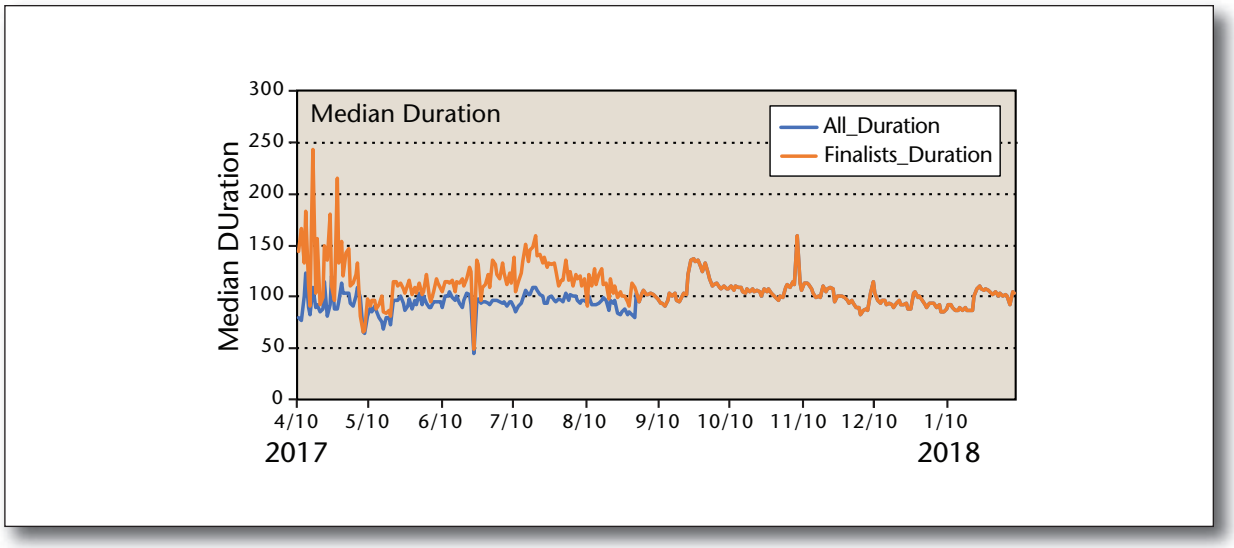


Figure 4. Conversation Duration 90th Percentile for Socialbots During the Competition.

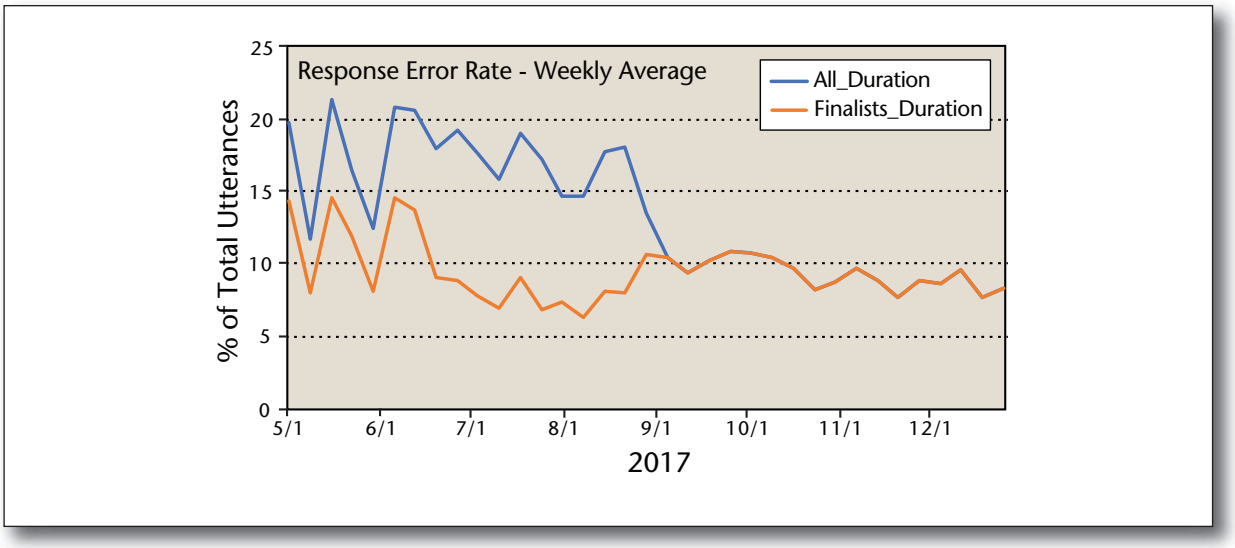


Figure 5. Response Error Rate for Individual Utterance-Response Pairs for All Socialbots and Finalists During the Competition.

Conclusion and Future Work

Fifteen teams went live in the inaugural Alexa Prize, and customer ratings improved by about 24 percent over the duration of the competition (May 8 to Nov 13). An analysis of the technical and scientific detail for each team in relation to their performance in the competition led to the following findings:

The following components are key for building an effective socialbot: (1) dialogue manager (DM), (2) natural language understanding (NLU) module, (3) knowledge module, (4) response generation, (5) conversational user experience (CUX) handler, (6) ranking and model selection policy module.

Teams that focused on building CUX modules saw significant gains on ratings and duration. CUX is an essential component, and the teams who focused most of their efforts on NLU and DM, with less emphasis on CUX, were not received as top performers by Alexa users.

A robust NLU system supported by strong domain coverage leads to high coherence. Teams who invest-

ed in building a strong NLU and knowledge components had the lowest response error rates leading to higher user ratings.

Different conversational goals call for different response-generation techniques, suggesting that retrieval, generative, and hybrid mechanisms may all be required within the same system. When the performance of a socialbot has converged, generative and hybrid modules combined with a robust ranking and selection module can lead to a better conversational agent.

A response ranking and selection model greatly impacts socialbot quality. The teams who built a strong model-selection policy had significant improvements in ratings and average number of dialogue turns.

Even if a socialbot has strong response-generation and ranker modules, lack of good NLU and DM components adversely affect user ratings.

We expected that the grand challenge of 20-minute conversations would take many years to achieve — the Alexa Prize was set up as a multiyear competition to enable sustained research on this problem. Despite the difficulty of the challenge, it is extremely encouraging to see the work that the inaugural cohort of the Alexa Prize has achieved in year one of the competition. We have seen significant advancements in research, and in the quality of socialbots as observed through the customer ratings, but much remains to be achieved. With the help of Alexa users and the science community, Alexa Prize 2018 will continue to work towards the goal of 20-minute-long coherent and engaging social conversations, and continue to advance the state of conversational AI.

Acknowledgments

We would like to thank all the university students and their advisors (Alexa Prize Teams 2017) who participated in the competition. We would also like to thank the entire Alexa Prize team (Eric King, Kate Bland, Qing Liu, Jeff Nunn, Ming Cheng, Ashish Nagar, Yi Pan, Han Song, SK Jayadevan, Amanda Wartick, Anna Gottardi, Gene Hwang, Art Pettigrue, and Nate Michel) for their contribution in making the Alexa Prize competition a success. We would also like to thank Amazon leadership and Alexa principals for their vision and support through this entire program; the marketing, public relations, and legal departments for helping drive the right messaging and a high volume of traffic to the Alexa Prize skill, ensuring that the participating teams received real-world feedback for their research; Alexa engineering for all the support and work on enabling the Alexa Prize skill and supporting a custom Alexa Prize ASR model, while always maintaining operational excellence; and Alexa machine learning for continued support with NLU and data services, which allowed us to capture user requests to initiate conversations and also provide high-quality annotated feedback to the teams. We also want to thank ASK leadership and the countless teams in ASK who helped us with the custom APIs for Alexa Prize teams, enabling skill beta testing for the Alexa Prize skills before it went general availability, and who further supported us with skill management, QA, certification, marketing, operations, and solutions. We would also like to thank the Alexa experiences organization for exemplifying customer obsession by providing us with critical input to share with the teams on building the best customer experiences and driving us to track our progress against customer feedback.

Finally, thank you to the Alexa customers who engaged in tens of thousands of hours of conversations spanning millions of interactions with the Alexa Prize socialbots and who provided the feedback that helped teams improve over the course of the year.

Notes

1. pacy.io.
2. www.evi.com.
3. www.wikidata.org/wiki/Wikidata:Main_Page.
4. See the Alexa Prize Proceedings, developer.amazon.com/alexaprize/proceedings.
5. alicebot.sourceforge.net/alice_page.htm.
6. www.youtube.com/watch?v=WTGuOg7GXYU&feature=youtu.be.
7. developer.amazon.com/alexaprize/2017/teams.

References

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1247–50. New York: Association for Computing Machinery.

Adewale, O.; Beatson, A.; Buniatyan, D.; Ge, J.; Khodak, M.; Lee, H.; Prasad, N.; Saunshi, N.; Seff, A.; Singh, K.; Suo, D.; Zhang, C.; Arora, S. 2017. Pixie: A Social Chatbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Bowden, K. K.; Wu, J.; Oraby, S.; Misra, A.; Walker, M. 2017. Slugbot: An Application of a Novel and Scalable Open Domain Socialbot Framework. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; Soria, C.; and Traum, D. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the International Conference on Language Resources and Evaluation*. Luxembourg: European Language Resources Association.

Cervone, A.; Tortoreto, G.; Mezza, S.; Gambi, E.; Riccardi, G. 2017. Roving Mind: A Balancing Act between Open–Domain and Engaging Dialogue Systems. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Fang, H.; Cheng, H.; Clark, E.; Holtzman, A.; Sap, M.; Ostendorf, M.; Choi, Y.; Smith, N. 2017. Sounding Board – University of Washington's Alexa Prize Submission. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Guo, G.; Metallinou, A.; Khatri, C.; Raju, A.; Venkastech, A.; and Ram, R. 2017. Topic-Based Evaluation for Conversational Bots. Paper presented at the NIPS 2017 Workshop on Conversational AI — Today's Practice and Tomorrow's Potential. Long Beach, CA, December 8.

Guss, W. H.; Bartlett, J.; Kuznetsov, P.; Patil, P. 2017. Eigen: A Step Towards Conversational AI. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–80.

Ji, J.; Wang, Q.; Battad, Z.; Gou, J.; Zhou, J.; Divekar, R.l; Carlson, C.; Si, M.. 2017. A Two-Layer Dialogue Framework for Authoring Social Bots. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Skip-Thought Vectors. In *Annual Conference on Neural Information Processing Systems*. Advances in Neural Information Processing Systems 28. Red Hook, NY: Curran Associates, Inc.

Krause, B.; Damonte, M.; DobreM.; Duma, D.; Fancellu, F.; Kahembwe, E.; Cheng, J.; Fainberg, J.; Webber, B. 2017. Edina: Building an Open Domain Socialbot with Self-Dialogues. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Kumar, A.; Gupta, A.; Chan, J.; Tucker, S.; Hoffmeister, B.; Dreyer, M; Peshterliev, S.; Gandhe, A.; Filiminov, D.; Rastrow, A.; Monson, C.; and Kuymar, A. 2017. Just ASK: Building an Architecture for Extensible Self-Service Spoken Language Understanding. arXiv preprint arXiv:1711.00549 [cs.CL]. Ithaca, NY: Cornell University Library.

Levesque, H. J. 2017. *Common Sense, the Turing Test, and the Quest for Real AI: Reflections on Natural and Artificial Intelligence*. Cambridge, MA: The MIT Press.

Liu, H.; Lin, T.; Sun, H.; Lin, W.; Chang, C.-W.; Zhong, T.; Rudnicky, A. 2017. RubyStar: A Non-Task-Oriented Mixture Model Dialog System. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations,* 55–60. Stroudsburg, PA: Association for Computational Linguistics.

Papaioannou, I.; Curry, A.; Part, J.; Shalyminov, I.; Xu, X.; Yu, Y.; Dusek, O.; Rieser, V.; Lemon, O. 2017. Alana: Social Dialogue Using an Ensemble Model and a Ranker Trained on User Feedback. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Pichl, J.; Marek, P.; Konrád, J.; Matulík, P.; Nguyen, H. L.; Sedivy, J. 2017. Alquist: The Alexa Prize Socialbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Prabhumoye, S.; Botros, F.; Chandu, K.; Choudhary, S.; Keni, E.; Malaviya, C.; Manzini, T.; Pasumarthi, R.; Poddar, S.; Ravichander, A.; Yu, Z.; Black. A. 2017. Building CMU Magnus from User Feedback. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Serban, I.; Sankar, C.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Chandar, S.; Ke, N. R.; Rajeswar, S.; de Brebisson, A.; Sotelo, J. M. R.; Suhubdy, D.; Michalski, V.; Nguyen A.; and Bengio, Y. 2017. The Octopus Approach to the Alexa Competition: A Deep Ensemble-Based Socialbot. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. arXiv preprint arXiv:1605.06069[cs.CL]. Ithaca, NY: Cornell University Library.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. Weakly Supervised Memory Networks. arXiv preprint arXiv:1503.08895[cs.NE]. Ithaca, NY: Cornell University Library.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business*. New York: Doubleday.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Annual Conference on Neural Information Processing Systems,* Advances in Neural Information Processing Systems 27, 3104–12. Red Hook, NY: Curran Associates, Inc.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460.

Venkatesh, A.; Khatri, C.; Ram, A.; Guo, F.; Gabriel, R.; Nagar, A.; Prasad, R.; Cheng, M.; Hedayatnia, B.; Metallinou, A.; Goel, R.; Yang, S.; and Raju, A. 2017. On Evaluating and Comparing Conversational Agents. Paper presented at the NIPS 2017 Workshop on Conversational AI — Today's Practice and Tomorrow's Potential. Long Beach, CA, December 8.

Vinyals, O., and Le, Q. 2015 A Neural Conversational Model. Paper presented at the ICML Deep Learning Workshop. Lille, France, July 10.

Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. arXiv preprint arXiv:cmp-lg/9704004 [cs.CL]. Ithaca, NY: Cornell University Library.

Weizenbaum, J. 1966. Eliza — A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9(1): 36–45.

Wright, D. R. 2005. Finite State Machines: CSC215 Class Notes. Unpublished Class Notes. Raleigh, NC: North Carolina State University.

Yi, S.; Jung, K.. 2017. A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy. A Two-Layer Dialogue Framework for Authoring Social Bots. *Alexa Prize Proceedings*. Seattle, WA: Amazon.

Chandra Khatri is an AI scientist at Amazon Lab126 with a research and development team responsible for making Alexa conversational. Currently, he is the leading scientist for Alexa Prize competition. Some of his recent work involves open-domain dialogue planning and evaluation, conversational speech recognition, conversational natural language understanding, and topic modeling. Prior to Alexa, Khatri was a research scientist at eBay in an applied science group. At eBay, he led various deep learning and NLP initiatives, such as automatic text summarization and automatic content generation within the ecommerce domain. He holds degrees in machine learning and computational science and engineering from the Georgia Institute of Technology and the Birla Institute of Technology and Science.

Anu Venkatesh is a technical program manager on the Alexa AI team and is responsible for execution of the Alexa Prize, a university competition to advance conversational

AI, and additional technical initiatives targeted at making Alexa more conversational. Her background is in speech, dialogue, and natural language understanding. Prior to Alexa, Venkatesh was a research scientist at Nuance Communications working on their speech recognition and virtual assistant technologies. Venkatesh received her master's degree in AI from the Georgia Institute of Technology in 2009 and a bachelor's degree in computer science from the PES Institute of Technology, India in 2007.

Behnam Hedayatnia is an applied scientist on the Alexa AI team. He has been working on conversational AI since 2017. His main focus is on driving conversational ASR and dialogue for the Alexa Prize. He received his MS in electrical engineering at the University of California, San Diego, in 2017, while working as a research scientist at the San Diego Super Computer Center. He received his BS in electrical engineering from the University of California, San Diego, in 2015.

Ashwin Ram is technical director of AI in the office of the CTO for Google Cloud. He focuses on bringing Google AI to the world through deep personalized engagement with the leadership of top companies to reimagine their businesses by leveraging the power of AI. He also works with Google's AI teams to drive new technologies and capabilities that address customer needs. Prior to Google, Ram was senior manager of AI science for Amazon Alexa. He led cross-functional R&D initiatives to create advanced conversational AI technologies for intelligent agents, including the university-facing Alexa Prize competition. Ram received his PhD from Yale University in 1989, his MS from University of Illinois in 1984, and his BTech from IIT Delhi in 1982.

Raefer Gabriel is the lead for the Alexa Prize, and manager of AI software development at Alexa AI. He focuses on driving the next generation of natural and social interaction on Alexa, and on leading the teams building models, infrastructure, and development toolkits to support open and multidomain conversation. Prior to Amazon, Gabriel was the CEO of machine intelligence studio Delvv, Inc., and chief scientist and cofounder of Reputation.com. He has founded other companies, including TruExchange, a futures trading software company, and worked in the hedge fund industry as a quantitative analyst. Gabriel received his MBA from Columbia Business School in 2007. He graduated from Harvard University with an AB in physics in 2000.

Rohit Prasad is vice president and head scientist of Alexa Artificial Intelligence. Prasad leads Alexa research and development in speech recognition, natural language understanding, and machine learning technologies. Prior to Amazon, Prasad was deputy manager and senior director of the speech, language, and multimedia business unit at Raytheon BBN Technologies. In that role, he directed US Government–sponsored research and development initiatives in speech-to-speech translation, psychological health analytics, document image translation, and STEM learning. Prasad earned his master's degree in electrical engineering at the Illinois Institute of Technology, Chicago, and a bachelor's degree in electronics and communications engineering from Birla Institute of Technology, India.

On Reproducible AI:
Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications

Odd Erik Gundersen, Yolanda Gil, David W. Aha

■ *Artificial intelligence, like any science, must rely on reproducible experiments to validate results. Our objective is to give practical and pragmatic recommendations for how to document AI research so that results are reproducible. Our analysis of the literature shows that AI publications currently fall short of providing enough documentation to facilitate reproducibility. Our suggested best practices are based on a framework for reproducibility and recommendations for best practices given by scientific organizations, scholars, and publishers. We have made a reproducibility checklist based on our investigation and described how every item in the checklist can be documented by authors and examined by reviewers. We encourage authors and reviewers to use the suggested best practices and author checklist when considering submissions for AAAI publications and conferences.*

Reproducibility is a cornerstone of the scientific method. The ability and effort required from other researchers to replicate experiments and explore variations depends heavily on the information provided when the original work was published. Reproducibility is challenging for many sciences, for example when the variability of physical samples and reagents can significantly affect the outcome (Begley and Ellis 2012; Lithgow, Driscoll, and Phillips 2017). In computer science, a large portion of the experiments are fully conducted on computers, making the experiments more straightforward to document (Braun and Ong 2014). Most AI and machine learning research also falls under this category of computational experimentation. However, reproducibility in AI is not easily accomplished (Hunold and Träff 2013; Fokkens et al. 2013; Hunold 2015). This may be because AI research has its own unique reproducibility challenges. Ioannidis (2005) suggests that the use of analytical methods which are still a focus of active investigation is one reason it is comparatively difficult to ensure that computational research is reproducible. For

  ISSN 0738-4602

Factor	Variable	Description
Method	Problem	Is there an explicit mention of the problem the research seeks to solve?
	Objective	Is the research objective explicitly mentioned?
	Research method	Is there an explicit mention of the research method used (empirical, theoretical)?
	Research questions	Is there an explicit mention of the research question(s) addressed?
	Pseudocode	Is the AI method described using pseudocode?
Data	Training data	Is the training set shared?
	Validation data	Is the validation set shared?
	Test data	Is the test set shared?
	Results	Are the relevant intermediate and final results output by the AI program shared?
Experiment	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Method source code	Is the AI system code available open source?
	Hardware	Is the hardware used for conducting the experiment specified?
	Software dependencies	Are software dependencies specified?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
	Experiment source code	Is the experiment code available open source?

Table 1. Description of All Variables and Their Factors.

example, Henderson et al. (2017) show that problems due to nondeterminism in standard benchmark environments and variance intrinsic to AI methods require proper experimental techniques and reporting procedures. Acknowledging these difficulties, computational research should be documented properly so that the experiments and results are clearly described.

The AI research community should strive to facilitate reproducible research, following sound scientific methods and proper documentation in publications. Concomitant with reproducibility is open science, which involves sharing data, software, and other science resources in public repositories using permissive licenses. Open science is increasingly associated with FAIR principles to ensure that science resources have the necessary metadata to make them findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Modern digital scholarship promotes proper credit to scientists who document and share their research products through citations of datasets, software, and innovative contributions to the scientific enterprise.

The focus in this article is on best practices for documentation and dissemination of AI research to facilitate reproducibility, support open science, and embrace digital scholarship. We begin with an analysis of recent AI publications that highlights the limitations of their documentation in support of reproducibility.

State of the Art: How AI Research Is Currently Documented

Gundersen and Kjensmo (2018) analyzed how well empirical AI research is documented to facilitate reproducibility. Empirical AI research involves evaluating how well computational AI methods solve a problem. An AI method refers to an abstract method for solving such problems. Examples include agent systems that perform collaborative tasks and neural network architectures trained using backpropagation.

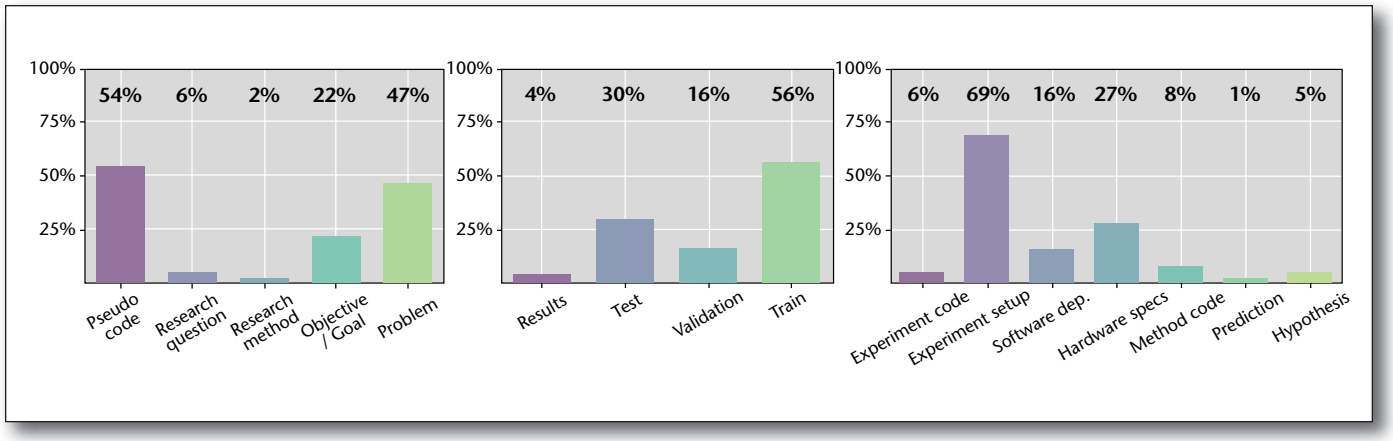


Figure 1. Percentage of Papers Documenting Each Variable for the Three Factors

Method (left), data (middle), and experiment (right). Gundersen and Kjensmo (2018).

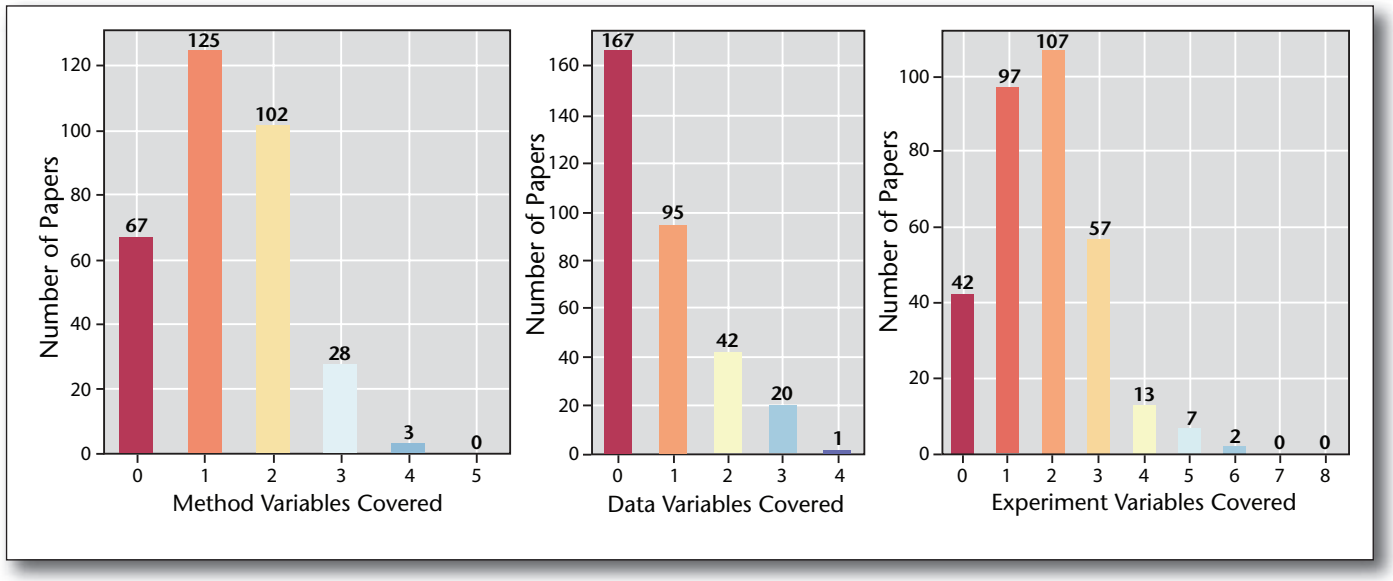


Figure 2. The Number of Variables for the Three Factors as Documented for All the Papers Describing Empirical Research in the Study.

Method (left), data (middle), and experiment (right). Gundersen and Kjensmo (2018).

The analysis by Gundersen and Kjensmo (2018) is based on a literature review and a framework for reproducibility. Their framework divides documentation into three factors: (1) method, which specifies the AI method under investigation and the problem to be solved; (2) data, which describes the data used for conducting the empirical research; and (3) experiment, which documents how the experiment was conducted. How well these three factors are documented is indicated by 16 yes/no variables (see table 1) that are directly relevant for facilitating reproducibility.

A publication that documents an empirical research study can be scored using these variables. Three reproducibility metrics are proposed. The three metrics are: (1) R1D, which calculates the average of all variables for all three factors (method, data, and experiment); (2) R2D, which computes the average of the variables for the method and data factors; and (3) R3D, which calculates the average of all variables for the method factor. These three metrics provide an indication of how well the scored papers document the research for three different degrees of reproducibility (we provide more detail on these degrees of reproducibility later on in the article).

In total, Gundersen and Kjensmo sampled 400

papers from the AAAI 2014, AAAI 2016, IJCAI 2013, and IJCAI 2016 conferences. Among these, 325 papers describe empirical studies, while the remaining 75 papers do not. Figure 1 displays the percentage of the surveyed papers that documented the different variables, while figure 2 summarizes how many of the variables were documented for each factor per paper.

We make a few observations. As seen in figure 1, few of the papers explicitly mention the research method that is used, and only around half explicitly mention which problem is being solved. Only about a third of the papers share the test dataset and only 4 percent share the result produced by the AI program. Only 8 percent of the papers share the source code of the AI method that is being investigated, while only 5 percent explicitly specify the hypothesis and 1 percent specify their prediction. Figure 2 shows that 67 papers do not explicitly document any of the variables for the factor method; only one paper documents and shares training, validation, and test sets as well as the results; and approximately 90 percent of the papers document no more than three of the seven variables of the factor experiment.

As seen in figure 3, the trends are unclear. Statistical analysis showed that only two of the metrics, R1D and R2D, for IJCAI had a statistically significant increase over time. While R2D and R3D for AAAI decrease over time, the decrease is not statistically significant.

The study by Gundersen and Kjensmo (2018) has some limitations. For example, the study required that for the variable problem to be set to yes (true), the paper must explicitly state the problem that is being solved. Another shortcoming is that all the AI methods that are documented in the research papers are not necessarily described better with pseudocode than without, but this fact was not given any consideration. If a paper described an AI method and pseudocode was not provided, the pseudocode variable was set to false for that paper. Finally, some of the variables might be redundant (for example, problem, goal, or research questions). Still, despite these shortcomings, the findings indicate that computational AI research is not documented systematically and with enough information to support reproducibility.

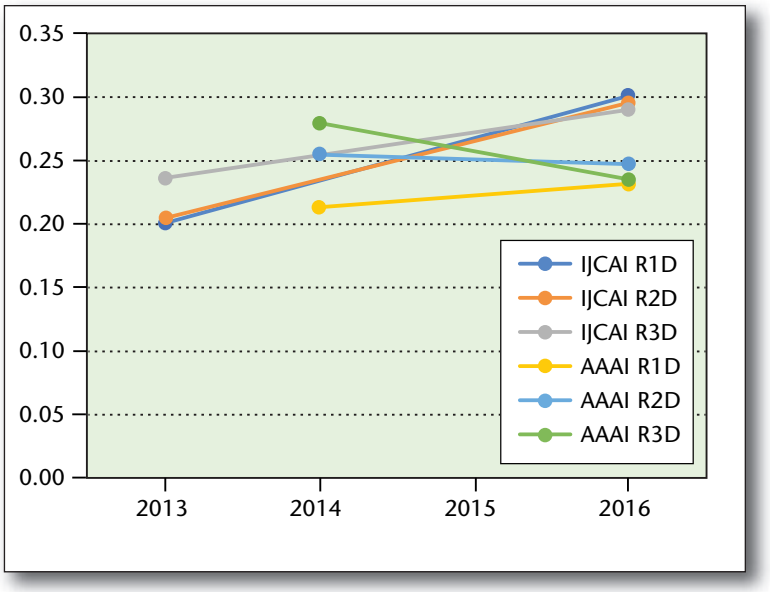


Figure 3. Change Over Time of the Three Reproducibility Metrics for Selected Years of the Two Conferences AAAI and IJCAI.

Gundersen and Kjensmo (2018).

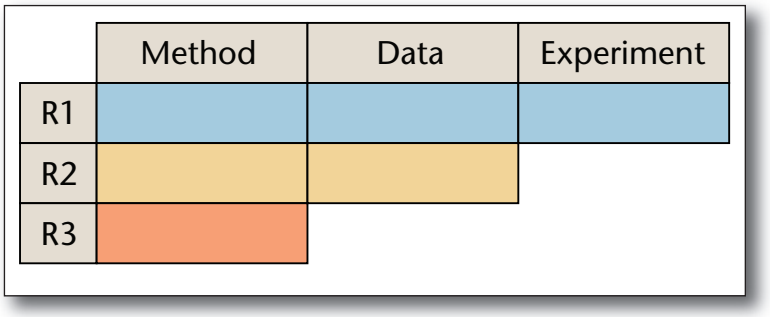


Figure 4. The Three degrees of Reproducibility Are Defined by Which Documentation Is Used to Reproduce the Results.

The three degrees of reproducibility each require a different set of factors to be documented.

Degrees of Reproducibility

Gundersen and Kjensmo (2018) distinguish between three degrees of reproducibility, which are defined as follows:

R1: Experiment Reproducible. The results of an experiment are experiment reproducible when the execution of the same implementation of an AI method produces the same results when executed on the same data. This is often called *repeatability.*

R2: Data Reproducible. The results of an experiment are data reproducible when an experiment is conducted that executes an alternative implementation of the AI method that produces the same results when executed on the same data. This is often called *replicability.*

R3: Method Reproducible. The results of an experiment are method reproducible when the execution of an alternative implementation of the AI method produces consistent results when executed on different data. This is often called *reproducibility.*

Empirical research that is R1 (experiment reproducible) must document the AI method, the data used to conduct the experiment, and the experiment itself including the source code for the AI method and the experiment setup, while R2 (data reproducible) research must only document the AI method and the data. R3 (method reproducible) research must only document the AI method. Figure 4 illustrates the different factors that must be documented for the three reproducibility degrees.

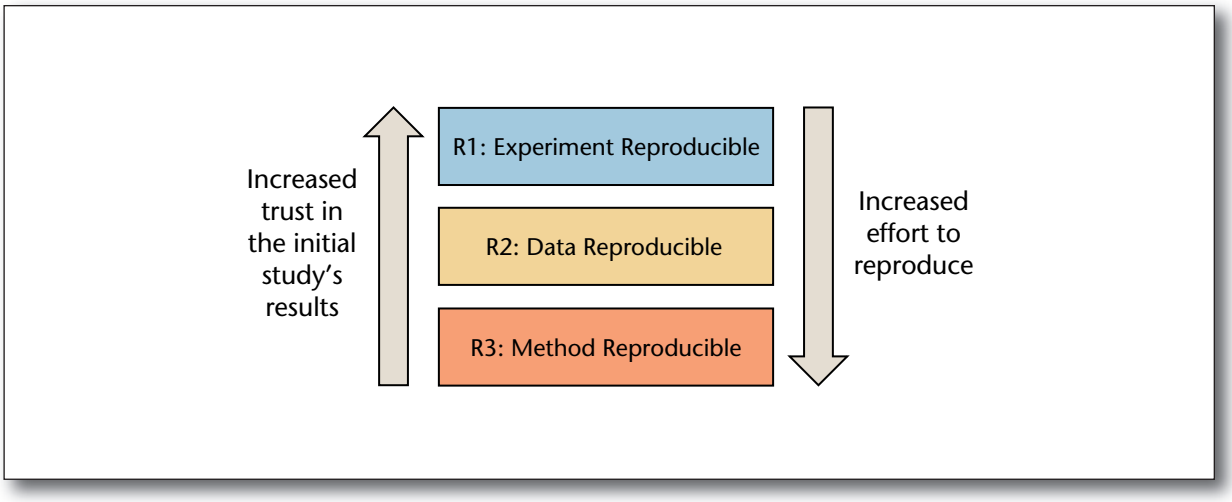


Figure 5. Effects of Documentation as Seen from the Perspective of Independent Researchers.

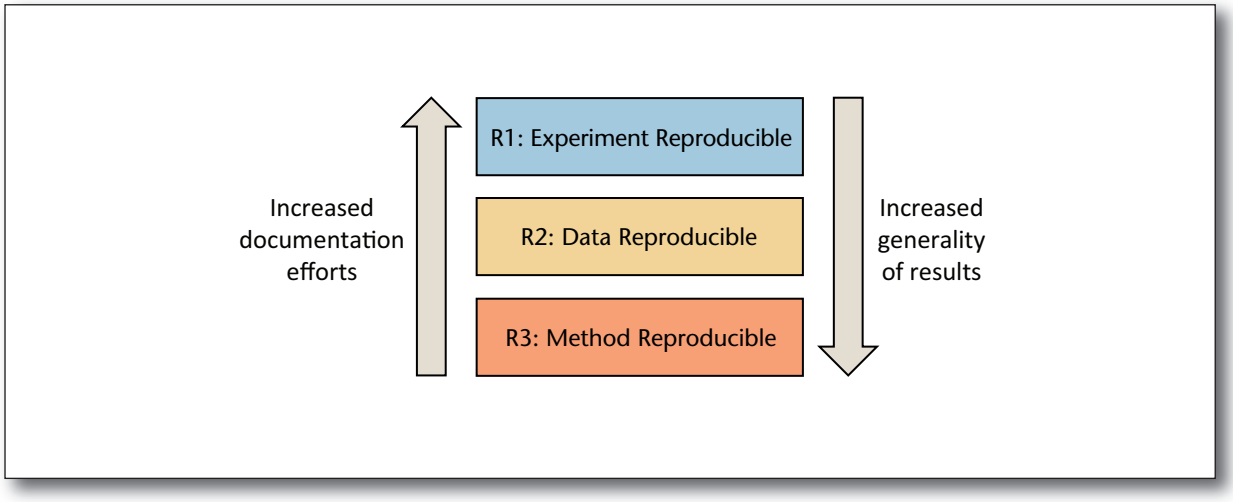


Figure 6. Effects of Documentation as Seen from the Perspective of the Original Researchers.

If an independent team reproduces research and gets results that are consistent with the original results, the generality of the results depends on the level of documentation that was provided to the independent team. If the original research was R1 (experiment reproducible), the independent team has confirmed that the specific implementation of the AI method provided by the original research team achieved the reported results on the specific data that also was provided by the original research team. Hence, the generality of the results is limited to that specific implementation and that specific data. However, if the independent team reproduces the results of some research that was R3 (method reproducible) and gets consistent results, the results are more general, as they apply to a reimplementation and other data. This factor leads to different incentives for the researchers who conducted the initial empirical study and the independent researchers who attempt to reproduce the results.

Independent researchers trust an empirical study's results increasingly with the amount of documentation that is shared with them, while the effort to reproduce the results increases when the amount of documentation is reduced. This situation is illustrated in figure 5. Hence, independent researchers would prefer R1 (experiment reproducible) research.

On the other hand, the effort to document the research increases for the original researchers with the amount of documentation that needs to be shared, while the generality of the method is increased if independent researchers reproduce the results given less documentation. Hence, the original researchers may prefer to document their research to be R3 (method reproducible) (figure 6).

Combine this conflict of incentives for the origi-

Recommendations	*Data mentioned in a publication should:*
1.	Be available in a shared community repository, so anyone can access it
2.	Include basic metadata, so others can search and understand its contents
3.	Have a license, so anyone can understand the conditions for reuse of the data
4.	Have an associated digital object identifier (DOI) or persistent URL (PURL) so that the data is available permanently
5.	Be cited properly in the prose and listed accurately among the references, so readers can identify the datasets unequivocally and data creators can receive credit for their work

Table 2. Author Checklist Part I.

Recommendations for data in publications.

nal and independent researchers with the pressure to publish, and it is easy to see how this situation can lead to research being documented less vigorously. However, by following the recommendations given here, authors can increase the trustworthiness and reproducibility of research results with relatively little effort. Still, changes cannot be expected solely from individual researchers alone. The research community, funding sponsors, employers of researchers, and publishers should each, in their respective roles, incentivize and reward reproducible research.

Best Practices and Recommendations

The recommendations we introduce are based on best practices put forward by scientific organizations such as the Research Data Alliance;[1] the Federation of Earth Science Information Partners;[2] DataCite;[3] the National Research Council (2012); the Task Group on Data Citation Standards and Practices (2013); the Data Citation Synthesis Group (2014); and scholars such as Ball and Duke,[4] Wilkinson et al. (2016), Stodden et al. (2016), Gil et al. (2016), Nosek et al. (2015), Starr et al. (2015), Downs et al. (2015), Mooney and Newton (2012), Goodman et al. (2014), Garijo et al. (2013), and Altman and King (2007), as well as earth and space science publishers[5] Hanson et al. (2015).

Strong momentum is building in support of FAIR practices, that is, to make data findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Our recommendations support FAIR principles and extend them to promote reproducible research, open science, and digital scholarship.

Implementing these recommendations requires extra space in publications. We suggest including this additional content in appendices that technical reviewers will not be required to assess but can quickly check. For electronic publications, there should not be any space limitations imposed for such appendices.

When these recommendations cannot be met, a brief explanation should be included about the reasons. Possible reasons may be restricted access (for example, proprietary or sensitive data), ownership by close collaborators who do not wish to disclose certain details, inadequate resources (for example, to house large datasets), or an unreasonable burden on authors.

We begin with recommendations for data and source code as the basic ingredients of a computational experiment. Then we describe recommendations to document AI methods and the experiments themselves. If all recommendations for AI methods (table 4) are implemented, then the publication should in theory be R3 (method reproducible), while if all recommendations for data (table 2) are also implemented, then the research should be R2 (data reproducible). Finally, all four sets of recommendations (tables 2–5) must be implemented for the research to be fully R1 (experiment reproducible).

We will refer to the complete set of 20 recommendations as an author checklist, we provide examples to demonstrate that they are synergistic, and we argue that they can be easily implemented.

Recommendations for Data

Table 2 summarizes our recommendations for documenting data, which concern (1) repository use, (2) metadata, (3) licenses, (4) persistent unique identifiers, and (5) citations. These recommendations can be easily implemented if researchers use community data repositories that support recommended best practices.

Data Repositories

Data repositories exist for many domains, and as such they are available to the AI community. Examples of these general repositories are Zenodo,[6] figshare,[7] and Dataverse.[8] These repositories will automatically assign a DOI to any uploaded data and will also accept software, figures, movies, and slide

Recommendations	*Source code used for implementing an AI method and executing an experiment should:*
6.	Be available in a shared community repository, so anyone can access it
7.	Include basic metadata, so others can search and understand its contents
8.	Include a license, so anyone can understand the conditions for use and extension of the software
9.	Have an associated digital object identifier (DOI) or persistent URL (PURL) for the version used in the associated publication so that the source code is permanently available
10.	Be cited and referenced properly in the publication so that readers can identify the version unequivocally and its creators can receive credit for their work

Table 3. Author Checklist Part II.

Recommendations for source code implementing AI methods and experiments in publications.

presentations. They will also inquire about choosing a license and about specifying a descriptive name and authors for a submitted dataset. AAAI could, as a service, provide a list of recommended data repositories. This list could be modeled on a service provided by COPDESS, which is a large coalition for publishing data in the earth and space sciences.[5] Universities also offer general repositories, whether developed in-house or as installations of general infrastructure such as Dataverse. University repositories are typically maintained by library departments, and always offer DOIs, licenses, and citations.

We encourage maintainers of data repositories that serve the AI community to adopt mechanisms for assigning DOIs or persistent URLs (PURLs) to datasets that they provide. The management of PURLs or DOIs can be complex. We suggest consulting with organizations such as FORCE11 and the Research Data Alliance, which have working groups with extensive and detailed recommendations on this topic.

Metadata

Basic metadata includes a descriptive title, the dataset's authors, and creation date. Additional metadata is always valuable to others in terms of understanding and reusing the dataset.

Licenses for Data

Recommended licenses for data are Creative Commons licenses,[9] preferably CC-BY (unlimited reuse as long as there is attribution) or CC0 (unlimited reuse without conditions).

Permanent Unique Identifiers for Data

Many authors make data available by providing a URL to their personal or lab pages. These references may not last long due to changes in sites and in author affiliations (Klein et al. 2014). Instead, we encourage authors to use persistent unique identifiers so that their data is always available. DOIs are managed by data repositories and given to individual datasets or to collections (DeRisi et al. 2003). Most data repositories provide DOIs, and for this they forge an agreement with a DOI authority. Another option that anyone can use is PURLS. PURLs can be assigned by anyone to any web resource using a trusted service such as the W3C's w3id.[10] Data repositories also have the option of using PURLs.

Data Citation

A data citation can be directly provided by a data repository, or it can be constructed by hand. A citation for a dataset consists of a descriptive name (or title) for the dataset, its creators, the name of the repository where it can be accessed, and the permanent URL. For example, a citation for a dataset in Gil et al. (2017) is:

> Adusumilli, Ravali. (2016). Sample datasets used in (Gil et al. 2017) for AAAI 2017 (Data set). Zenodo. http://doi.org/10.5281/zenodo.180716.

Note that by simply uploading the dataset to the Zenodo repository, we obtained the DOI and the citation. Specifying the authors, the name, and the license takes negligible effort. The author checklist for data required little time to implement.

Recommendations for Source Code

We refer to source code as the human-readable computer instructions written in plain text and software as computer programs that are executable by a computer. Typically, source code is compiled to software for a computer to run it. Our recommendations for source code are summarized in table 3.

Source Code Repositories

Source code repositories can be used by any scientists to share code, and as such they are available to the AI community. These code repositories include general repositories such as GitHub and BitBucket, and language-specific repositories such as CRAN for R code or File Exchange in MATLAB Central. General data repositories such as those mentioned above accept source code as an entry, and as with any dataset they

Recommendations	*AI methods used in a publication should be:*
11.	Presented in the context of a problem description that clearly identifies what problem they are intended to solve
12.	Outlined conceptually so that anyone can understand their foundational concepts
13	Described in pseudocode so that others can understand the details of how they work

Table 4. Author Checklist Part III.

Recommendations for AI methods in publications.

always offer DOIs, licenses, and citations.

For a specific publication, the version of the source code that is being used should be clearly specified, and the source code repository should support the identification and future access of specific versions.

Source Code Metadata

Basic metadata includes a descriptive title, the source code's authors, and the creation date. Additional metadata is always valuable to others in terms of understanding and reusing the source code.

Licenses for Source Code

Recommended licenses for source code are the standard licenses from the Open Source Initiative. Licenses such as Apache v2 or MIT are recommended because they provide unlimited reuse (as long as there is attribution). Other more restrictive licenses are available to limit commercial use or impose licensing conditions on extensions of the original source code.

Permanent Unique Identifiers for Source Code

A separate DOI should be assigned to meaningful versions of the source code, such as a version used for a publication. GitHub offers an option to obtain a DOI for a source code version, which is done by storing that version permanently in the Zenodo data repository. Any source code can be uploaded manually to community data repositories such as Zenodo, figshare, and Dataverse. PURLS can be assigned by anyone to any source code version that has a URL on the web, using a trusted service such as w3id.org.

Source Code Citation

Source code citation can be directly provided by a source code repository, or it can be constructed by hand. A citation for a source code version consists of a descriptive name (or title) for the source code, its creators, the name of the repository where it can be accessed, the version, and the permanent URL. For example, a citation for GitHub code in (Gil et al. 2017) is:

Ratnakar, Varun. "DISK software" (v1.0.0). Zenodo. 2016. http://doi.org/10.5281/zenodo.168079.

By uploading the source code into the GitHub code repository, we obtained a persistent identifier for the version used in the publication as well as the citation. Specifying the authors, the name, and the license takes negligible effort. Implementing the author checklist for source code required little time.

Recommendations for AI Methods

Our recommendations for AI methods are listed in table 4.

Problem Description

The problem that a conceptual AI method solves should be explicitly described in the publication. In De Weerdt et al. (2013) the following example can be found: "To address this problem, we propose a novel navigation system ..." The authors explicitly describe the problem that they address. Another good example of this practice can be found in He et al. (2016). Here the authors state the problem explicitly: "In this paper, we address the degradation problem by introducing a deep residual learning framework." The degradation problem is also properly described in their publication.

Conceptual Method

A high-level, textual description of the AI method should be provided to readers to allow them to gain an understanding of it. This description should include a broad overview of how the AI method works and specify input variables and the resulting output. In general, the AI research community excels at providing this information in publications.

Pseudocode

Pseudocode for the AI method should also be provided. In cases where detailed pseudocode cannot be provided due to the complexity of the proposed algorithm or system, a more abstract pseudocode summary can be provided that illustrates the AI method's flow.

Recommendations	*Descriptions of experiments in a publication should:*
14.	Explicitly present the hypotheses to be assessed, before other details concerning the empirical study are presented
15.	Present the predicted outcome of the experiment, based on beliefs about the AI method and its application
16.	Include the experiment design (parameters and the conditions to be tested) and its motivation, such as why a specific number of tests or data points are used based on the desired statistical significance of results and the availability of data
17.	Identify and describe the measure and metrics
18.	Provide the evaluation protocol
19.	Share the results
20.	Describe the results and the analysis
21.	Be described as a workflow that summarizes how the experiment is executed and configured
22.	Include documentation on workflow executions or execution traces that provide parameter settings and initial, intermediate, and final data
23.	Specify the hardware used to run the experiments
24	Be cited and published separately when complex, so that others can unequivocally refer to the individual portions of the method that they reuse or extend

Table 5. Author Checklist Part IV.

Recommendations for experiments described in publications.

Both a high-level description and pseudocode help independent researchers to decide whether their own implementation of the method is correct. If these are not presented carefully, then the empirical study cannot always be easily reproduced.

Recommendations for Experiments

Authors should, to the degree possible, detail how their experiments are designed, and indicate the rationale for their design. Our recommendations for documenting experiments are summarized in table 5.

Hypotheses and Predictions

Hypotheses and predictions should be stated explicitly before descriptions of the other components of an empirical study to ensure that the results analysis is meaningful (Baker 2016).

Experiment Design

A textual description and justification of the experiment's design should be provided, to include a description of each test condition. This description should also explain, for example, why a specific number of tests or data points are used, based on the desired statistical significance of the results and the availability of data.

Measure and Metrics

Identify/define the measures and metrics to be used for the results analysis.

Evaluation Protocol

A justification should be provided for the chosen protocol when documenting an experiment. To avoid hypothesis myopia, this experiment should not be designed to collect only evidence that is guaranteed to support the stated hypotheses. Instead, to encourage an insightful study, it should include conditions that could lead to the rejection of these hypotheses. Why are the datasets used appropriate for the experiment? Why is the chosen empirical design appropriate for assessing the hypothesis, and why are the metrics and measures appropriate for assessing the results?

Results

In order for an independent research team to be able to fully evaluate their reproduction, they would need to compare with the actual results. Hence, the results (the actual output) of the experiment should be shared.

Result Descripton and Analysis

The results should be presented, along with an in-depth analysis of the results based on the specified

measure and metrics. The documenattion should provide an explicit indication of whether the analyses support the hypotheses.

Workflow

This workflow should describe, in a machine-readable way, how software and data are used to implement the evaluation protocol. A workflow step is an invocation of the software. Each step has input data and parameters as well as output data. Input data and the output of any step can be used as input to subsequent steps. The simplest workflow languages capture methods that are directed acyclic graphs, while other languages can represent iterations and conditionals. A publication that simply mentions what software was used usually leaves out critical information about how the software was configured or invoked.

Scripts or electronic notebooks can be an effective way to document workflows, although the organization of source code is more modular in a workflow structure.

Executions

A general workflow can be run many times with different datasets or parameter settings and generate different results. Execution traces of executed workflows provide a complete provenance trail of how each result was generated.

Hardware Specification

The hardware that is used should be specified if it is important for the experiment. This documentation may include specification of the processor type, the number of cores and processors, and RAM and hard disk requirements. Also, the provider of the cloud solution that is used, if any, should be specified. The machine architecture and operating system may need to be specified, so that any discrepancies in results can be properly diagnosed. Library dependencies should also be described. Virtualization technologies, such as Docker and Kubernetes, facilitate these specifications through artifacts called containers. Containers can be provided as appropriate to share the experiment hardware setup.

Workflow Citation

Citing a publication does not make explicit whether the citation is to its AI method, source code, data, empirical design, workflows, execution traces, results, or a general body of work or contributions. If it is important for others to be explicit about what aspects of the work are being reused, then separate citations should be given to each, as appropriate. Although workflow repositories are not as common as data and software repositories, many general data repositories accept any research product and can be used for this purpose.

For example, a citation for a bundle containing workflows and execution details for Gil et al. (2017) is:

Adusumilli, Ravali, Ratnakar, Varun, Garijo, Daniel, Gil, Yolanda, and Mallick, Parag. (2016). Additional materials used in the paper "Towards Continuous Scientific Data Analysis and Hypothesis Evolution" on the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) (Data set). Zenodo. http://doi.org/10.5281/zenodo.190374.

By organizing the workflows and executions described into this publication and bundling them to upload to a general data repository, these authors obtained a persistent identifier as well as a citation. The author checklist for experiments was implemented quickly.

Benefits to Authors

Recognizing that our recommendations will require effort from authors, we want to highlight the following 10 benefits: (1) Practice open science and reproducible research. This approach ensures the kinds of checks and balances that lead to better science. (2) Receive credit for all your research products (that is, through citations for software, datasets, and other products). (3) Increase the number of citations to your publications. Studies have shown that well-documented articles receive more citations (Piwowar et al. 2007). (4) Improve your chances of being funded (that is, by writing coherent and well-motivated empirical study and data management plans). (5) Extend your curriculum vitae. Include data and software sections with citations. Maintaining datasets and writing code are important contributions to the field of AI. (6) Improve the management of your research assets (for example, so your new students, and others, can more easily locate materials generated by your earlier students). (7) Allow for the reproduction of your work (for example, so you and others can leverage it in new studies, even if it was conducted many years ago). (8) Address new sponsor and journal requirements. They are steadfastly driving research towards increased reproducibility and open science. (9) Attract transformative students. They strive for a rigorous research methodology. (10) Demonstrate leadership. Step into the future.

By explicitly citing datasets and source code, and by providing workflows that are machine readable, we create the structure needed for the development of AI systems that can analyze and reason about our literature (Gil 2017). These AI systems would have access to a vast amount of structured scientific knowledge with comprehensive details about experimental design and results. This change could revolutionize how we approach the scientific research process.

Discussion

It is reasonable to expect a limited release of data and source code until the creator has completed the research for which the data was collected, or for

which the source code was written, or until their draft is published. Many journals impose this, such as *Science* and *Nature*. See Joly et al. (2012) for a review of data retention policies.

The creation and documentation of additional information we recommend should be done by researchers who publish their studies. Documenting and sharing code and data in such a way that this information can be easily used and cited by others gives researchers credit for a larger portion of their research effort. For academic researchers, we advocate that tenure committees give weight to the publication of data and source code when evaluating candidates for tenure. Thus, the publication velocity should not be reduced, but include research products other than publications.

The recommendations we suggest should be a part of daily research practices. According to Irakli Loladze, despite increasing the work load by 30 percent, "Reproducibility is like brushing your teeth. It is good for you, but it takes time and effort. Once you learn it, it becomes a habit" (Baker 2016).

Another recommendation for improving the readability and comparability of research papers is to require structured abstracts, which are commonly used in medical journals. Structured abstracts can be used to efficiently communicate a research objective, the motivation for and process by which an empirical study was conducted, and what results were achieved. Structured abstracts also require researchers to structure their own thoughts about their research. We suggest a five-part structured abstract containing (1) the research motivation, (2) the research objective, (3) the method used to conduct any empirical studies, (4) the results of the research, and (5) the conclusion. This structure enforces a coherent research narrative, which is not always the case for unstructured abstracts. The abstract for this article is an example of the proposed structure, while Gundersen and Kjensmo (2018) provides an abstract for empirical research that follows these recommendations and includes an explicit description of the hypothesis and an interpretation of the results.

Call to Arms

As a community, we should ensure that the research we conduct is properly documented. To make AI research reproducible and more trustworthy, we proposed best practices that should be adopted by editors and program chairs and incorporated into the review forms of AAAI publication venues.

Publishers should provide extra space to document and cite data, source code, and empirical study designs. AAAI leadership should encourage AI researchers to increase the reproducibility of their published work. This support could include providing structured templates to organize appendices and making available extra space in publications to accommodate the needed documentation.

For AI research to become open and more reproducible, the research community and publishers have to establish suitable practices. Authors need to adopt these practices, disseminate them to colleagues and students, and help develop mechanisms and technology to make it easier for others to adopt them.

Our objective with this article is to highlight the benefits of reproducible science and to propose initial, modest changes that can increase the reproducibility of AI research results. There are many additional actions that could and should be taken, and we look forward to further dialogue with the AI research community on how to increase the reproducibility and scientific value of AI publications.

Acknowledgements

This research was funded in part by the National Science Foundation under grant ICER-1440323. This work has in part been carried out at the Telenor-NTNU AI Lab, Norwegian University of Science and Technology, Trondheim, Norway. The recommendations proposed are based on the Geoscience Paper of the Future and the Scientific Paper of the Future best practices developed under that award. Thanks to Sigbjørn Kjensmo for all the effort put into surveying the state of the art of reproducibility of AI.

Notes

1. rd-alliance.org/outcomes.
2. wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines.
3. www.datacite.org.
4. www.dcc.ac.uk/resources/how-guides/cite-datasets#sthash.MJQjNn3i.dpuf.
5. www.copdess.org/statement-of-commitment.
6. zenodo.org.
7. figshare.com.
8. dataverse.org.
9. creativecommons.org.
10. www.w3id.org.

References

Altman, M., and King, G. 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine* 13 (3/4).

Baker, M. 2016. Is There a Reproducibility Crisis? *Nature* 533(7604): 452–54.

Begley, C. G., and Ellis, L. M. 2012. Drug Development: Raise Standards for Preclinical Cancer Research. *Nature* 483(7391): 531–33. doi.org/10.1038/483531a.

Bouquet, P.; Serafini, L.; Zanobini, S.; and Benerecetti, M. 2003. An Algorithm for Semantic Coordination. Paper presented at the Second International Semantic Integration Workshop. Sanibel Island, FL, October 20–23.

Braun, M. L., and Ong, C. S. 2014. Open Science in

Machine Learning. In *Implementing Reproducible Research,* edited by V. Stodden, F. Leish, and R. D. Peng, 343. Boca Raton, FL: CRC Press.

Data Citation Synthesis Group. 2014. Joint Declaration of Data Citation Principles, edited by M. Martone. San Diego, CA: FORCE11. doi.org/10.25490/a97f-egyk.

DeRisi, S.; Kennison, R.; and Twyman, N. 2003. The What and Whys of DOIs. *PLoS Biology* 1(2): 133–34, e57. doi.org/10.1371/journal.pbio.0000057.

De Weerdt, M. M.; Gerding, E. H.; Stein, S.; Robu, V.; and Jennings, N. R. 2013. Intention-Aware Routing to Minimise Delays at Electric Vehicle Charging Stations. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence,* 83–89. Palo Alto, CA: AAAI Press.

Downs, R. R.; Duerr, R.; Hills, D. J.; and Ramapriyan, H. K. 2015. Data Stewardship in the Earth Sciences. *D-Lib Magazine* 21 (7/8). doi.org/10.1045/july2015-downs.

Fokkens, A.; Erp M. V.; Postma, M.; Pedersen, M.; Vossen, P.; and Freire, N. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1691–701. Stroudsburg, PA: Association for Computational Linguistics.

Garijo, D.; Kinnings, S.; Xie, L.; Xie, L.; Zhang, Y.; Bourne, P. E.; and Gil, Y. 2013. Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. *PLos ONE* 8(11): e80278. doi.org/10.1371/journal.pone.0080278.

Gil, Y. 2017. Thoughtful Artificial Intelligence: Forging A New Partnership for Data Science and Scientific Discovery. *Data Science* 1(1–2): 119–29. doi.org/ 10.3233/DS-170011.

Gil, Y.; David, C. H.; Demir, I.; Essawy, B. T.; Fulweiler, R. W.; Goodall, J. L.; Karlstrom, L.; Lee, H.; Mills, H. J.; Oh, J.; Pierce, S. A; Pope, A.; Tzeng, M. W.; Villamizar, S. R.; and Yu, X. 2016. Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance. *Earth and Space Science* 3(10): 388–415. doi.org/10.1002/2015EA000136.

Gil, Y.; Garijo, D.; Ratnakar, V.; Mayani, M. R.; Adusumilli, R.; Boyce, H.; Srivastava, A.; and Mallick, P. 2017. Towards Continuous Scientific Data Analysis and Hypothesis Evolution. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4406–14. Palo Alto, CA: AAAI Press.

Goodman, A.; Pepe, A.; Blocker, A. W.; Borgman, C. L.; Cranmer, K.; Crosas, M.; Stefano, R. D.; Gil, Y.; Groth, P.; Hedstrom, M.; Hogg, D. W.; Kashyap, V.; Mahabal, A.; Siemiginowska, A.; and Slavkovic, A. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology* 10(4): e1003542. doi.org/10.1371/journal.pcbi.1003542.

Gundersen, O. E., and Kjensmo, S. 2018. State of the Art: Reproducibility in Artificial Intelligence. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 1644–51. Palo Alto, CA: AAAI Press.

Hanson, B.; Lehnert, K.; and Cutcher-Gershenfeld, J. 2015. Committing to Publishing Data in the Earth and Space Sciences. *Eos: Earth and Space Science News* 96. doi.org/doi:10.1029/2015EO022207.

He, K.; Zhang, X., Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning that Matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 3207–14. Palo Alto, CA: AAAI Press.

Hunold, S. 2015. A Survey on Reproducibility in Parallel Computing. arXiv preprint arXiv:1511.04217 [cs.DC]. Ithaca, NY: Cornell University Library.

Hunold, S., and Träff, J. S. 2013. On the State and Importance of Reproducible Experimental Research in Parallel Computing. arXiv preprint arXiv:1308.3648 [cs.DC]. Ithaca, NY: Cornell University Library.

Ioannidis, J. P. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2(8): e124. doi.org/10.1371/journal.pmed.0020124.

Joly, Y.; Dove, E. S.; Kennedy, K. L.; Bobrow, M.; Ouellette, B. F. F.; Dyke, S. O. M; Kato, K.; and Knoppers, B. M. 2012. Open Science and Community Norms: Data Retention and Publication Moratoria Policies in Genomics Projects. *Medical Law International* 12(2): 92–120. doi.org/10.1177%2F0968533212458431.

Klein, M.; de Sompel, H. V.; Sanderson, R.; Shankar, H.; Balakireva, L.; Zhou,K.; and Tobin, R. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS One* 9(12): e115253. doi.org/10.1371/journal.pone.0115253.

Lithgow, G. J.; Driscoll, M.; and Phillips, P. 2017. A Long Journey to Reproducible Results. *Nature* 548(7668): 387–88.

Mooney, H., and Newton, M. P. 2012. The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication* 1(1): eP1035. doi.org/10.7710/2162-3309.1035.

National Research Council. 2012. *For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press. doi.org/10.17226/13564.

Nosek, B. A.; Alter, G.; Banks, G. C.; Borsboom, D.; Bowman, S. D.; Breckler, S. J.; Buck, S.; Chambers, C. D.; Chin, G.; Christensen, G.; Contestabile, M.; Dafoe, A.; Eich, E.; Freese, J.; Glennerster, R.; Goroff, D.; Green, D. P.; Hesse, B.; Humphreys, M.: Ishiyama, J.; Karlan, D.; Kraut, A.; Lupia, A.; Mabry, P.; Madon, T.; Malhotra, N.; Mayo-Wilson, E.; McNutt, M.; Miguel, E.; Levy Paluck E.; Simonsohn, U.; Soderberg, C.; Spellman, B. A.; Turitto, J.; VandenBos, G.; Vazire, S.; Wagenmakers, E. J.; Wilson, R.; and Yarkoni, T. 2015. Promoting an Open Research Culture. *Science* 348(6242): 1422-25. doi.org/10.1126/science.aab2374.

Piwowar, H. A.; Day, R. S.; and Fridsma, D. B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi.org/10.1371/journal.pone.0000308.

Starr, J.; Castro, E.; Crosas, M.; Dumontier, M.; Downs, R. R.; Duerr, R.; Haak, L. L.; Haendel, M.; Herman, I.; Hodson, S.; Hourclé, J.; Kratz, J. E.; Lin, J.; Nielsen, L. H.; Nurnberger, A.; Proell, S.; Rauber, A.; Sacchi, S.; Smith, A.; Taylor, M.; and Clark, T. 2015. Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications. *PeerJ Computer Science* 1(1): e1. doi.org/10.7717/peerj-cs.1.

Stodden, V.; McNutt, M.; Bailey, D. H.; Deelman, E.; Gil, Y.; Hanson, B.; Heroux, M. A.; Ioannidis, J. P. A.; and Taufer, M. 2016. Enhancing Reproducibility for Computational Methods. *Science* 354(6317): 1240–41. doi.org/10.1126/ science.aah6168.

Task Group on Data Citation Standards and Practices. 2013.

Out of Cite, Out of Mind: Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal 12*:CIDCR1–7 doi.org/10.2481/dsj.OSOM13-043.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: article 160018. doi.org/10.1038/sdata.2016.18.

Odd Erik Gundersen (PhD, Norwegian University of Science and Technology) is an adjunct associate professor at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, where he teaches courses and supervises master students in AI. Gundersen has applied AI in the industry, mostly for startups, since 2006. Currently, he investigates how AI can be applied in the renewable energy sector and for driver training.

Yolanda Gil (PhD, Carnegie Mellon University) is director of knowledge technologies and an associate division director at the Information Sciences Institute of the University of Southern California, and a research professor in computer science and in spatial sciences. Her research is on intelligent interfaces for knowledge capture and discovery, semantic workflows and metadata capture, social knowledge collection, computer-mediated collaboration, and automated discovery. Gil has served on the Advisory Committee of the Computer Science and Engineering Directorate of the National Science Foundation. She initiated and chaired the W3C Provenance Group that led to a community standard in this area. Gil is also a Fellow of AAAI and was elected as its 24th president in 2016.

David W. Aha (PhD, University of California, Irvine) leads a section within NRL's Navy Center for Applied Research in AI, in Washington, DC. In addition to encouraging reproducible research, his interests include mixed-initiative intelligent agents, deliberative autonomy, explainable AI, case-based reasoning, and machine learning. He has co-organized many events on these topics, launched the UCI Repository for ML Databases, served as an AAAI councilor, cocreated AAAI's AI Video Competition, and currently leads the evaluation team for DARPA's Explainable AI program.

Integrating Artificial and Human Intelligence in Complex, Sensitive Problem Domains: Experiences from Mental Health

Munmun De Choudhury, Emre Kiciman

■ *This article presents a position highlighting the importance of combining artificial intelligence approaches with human intelligence, in other words, the involvement of humans. To do so, we specifically focus on problems of societal significance, stemming from complex, sensitive domains. We first discuss our prior work across a series of projects surrounding social media and mental health, and identify major themes for which augmentation of AI systems and techniques with human feedback has been and can be fruitful and meaningful. We then conclude by noting the implications, in terms of opportunities as well as challenges, that can be drawn from our position, both relating to the specific domain of mental health and for AI researchers and practitioners.*

Artificial intelligence methods are becoming a critical tool for impacting a variety of domains of broad societal significance (Boyd and Crawford 2012), from economic development (Jean et al. 2016) and education (He et al. 2015) to the environment (Dietterich 2009) and agriculture (Vasisht et al. 2017). A significant strength of AI in domains such as these is its ability to turn new sources of data into signals relevant to a domain. These new data sources allow, for example, AI to expand our ability to more easily reach and help vulnerable populations, to more quickly detect people at risk of poor outcomes, to identify customized or personalized solutions, and to enable early interventions.

Many of these societally significant domains are complex — understanding the mechanisms, dynamics, and interactions at work is challenging, and because the issues involve personal information and artifacts about individuals, they require careful, responsible attention among researchers and stakeholders. Further, not only do these problems involve using AI to derive insights from data, but they also require determining if those insights are practical and can be used to help relevant domain experts and stakeholders. Consequently, AI alone provides only a partial perspective when the goal is to interpret and translate the methods and findings to real-world settings. To validate and complement a particular AI analysis, we must go beyond a particular dataset or regime and bring in external domain knowledge of assumptions and plausible mechanisms. Judeah Pearl notes the limitations of approaches informed only by "naked data" and argues that one needs knowledge from outside the data (Pearl 2018):

> Data science is only as much of a science as it facilitates the interpretation of data — a two-body problem, connecting data to reality. Data alone are hardly a science, regardless how big they get and how skillfully they are manipulated.

In our experiences applying artificial intelligence methods to the analysis of new data sources to better understand the complex and sensitive domain of mental health, we have often drawn on human intelligence for prior knowledge, oversight, and analysis to augment pure AI methods. Four of the broad issues we have come across that have required such augmentation with human intelligence: ensuring construct validity, assumptions on unobserved factors, understanding data biases, and navigating sensitivities.

Ensuring Construct Validity

First, when using AI to extract core measurements, we are concerned with the construct validity of these measures. That is, are we actually measuring what we think we are measuring? For example, if we are trying to measure mood from the language people use on social media, are the words they use reflective of the moods they are actually experiencing? While this may sometimes be the case, self-presentation bias, cultural norms, word ambiguities, and even song lyrics can complicate the association between people's experienced moods and their expression on social media. If not recognized and corrected, these false associations can entirely threaten the validity of our measurements and, through them, any conclusions we might wish to draw from the data.

Assumptions on Unobserved Factors

Second, when we are attempting to understand a phenomenon through its representation in data, we must be aware that our observations may be significantly influenced by unobserved factors. When using AI methods to model people's behaviors and their reactions to an event or treatment within the data, for example, we must take into account that people will also be affected by external cultural factors, social influence, seasonal dynamics, larger trends, and other events not captured within the data. How these factors manifest can vary as well: each may vary across individuals in a dataset, or, alternatively, affect all individuals simultaneously. These unobserved factors can confound our understanding of the situation, causing us to misunderstand the underlying mechanisms and draw the wrong conclusions about the severity of a situation or about the recommendations for action to improve a situation.

Understanding Data Biases

Third, when using AI for data-driven learning about a complex domain, we must have an understanding of the biases within the data being studying. Due to limitations in the data, it is possible that our learnings are only valid under certain situations or for a certain group of people. In the context of mental health, for example, the complexities of the domain mean that conclusions drawn based on a limited subpopulation might be very different than conclusions drawn for another subpopulation or for the population as a whole. To generalize what we are learning, we must have validation that the people and the specific situations we are studying through a dataset are representative of the broader phenomenon we care about.

Navigating Sensitivities

Finally, many of the societally relevant domains where AI frameworks and tools have been found to be promising also tend to be areas where decision-making is high stake and high cost, meaning that mistakes and errors can have serious implications for human life, both figuratively and literally, as well as for human cost. In other words, if AI is employed to make decisions in an automated fashion, errors are unacceptable, although building 100 percent fault-free AI systems is far from reality today. To realize the potential of AI in these critical and important domains, the involvement of humans and experts is paramount, to ensure that there are adequate mechanisms to circumvent the mistakes made by the AI systems, to ensure that adequate risk mitigation protocols are in place when inappropriate or dangerous decisions are made by AI, and also to ensure that decision-making is arrived at in some collaborative fashion between the AI system and humans.

These issues are not concerns only when using sophisticated AI methods, of course. Construct validity, unobserved factors, data biases, and domain sensitivities are challenges faced by any quantitative analysis. We argue, however, that these issues are particularly critical threats to the validity of AI-driven analysis because of the challenges of interpreting and understanding the limitations of many black-box AI methods. Moreover, the challenges of interpretability and understanding are exacerbated and the cost of mistakes are magnified in a complex and sensitive domain.

In the remainder of this article, we highlight our experiences using artificial and natural intelligence together in complement, and discuss open challenges and opportunities for future research. The specific domain we focus on for our position concerns mental health.

Integrating Artificial and Human Intelligence

In this section, we discuss some key methods for augmenting AI approaches with the help of natural intelligence, specifically, human involvement. We draw from a variety of projects in our prior research that surround the complex domain of mental health.

Source of Gold Standard Information

One of the common places where researchers tend to

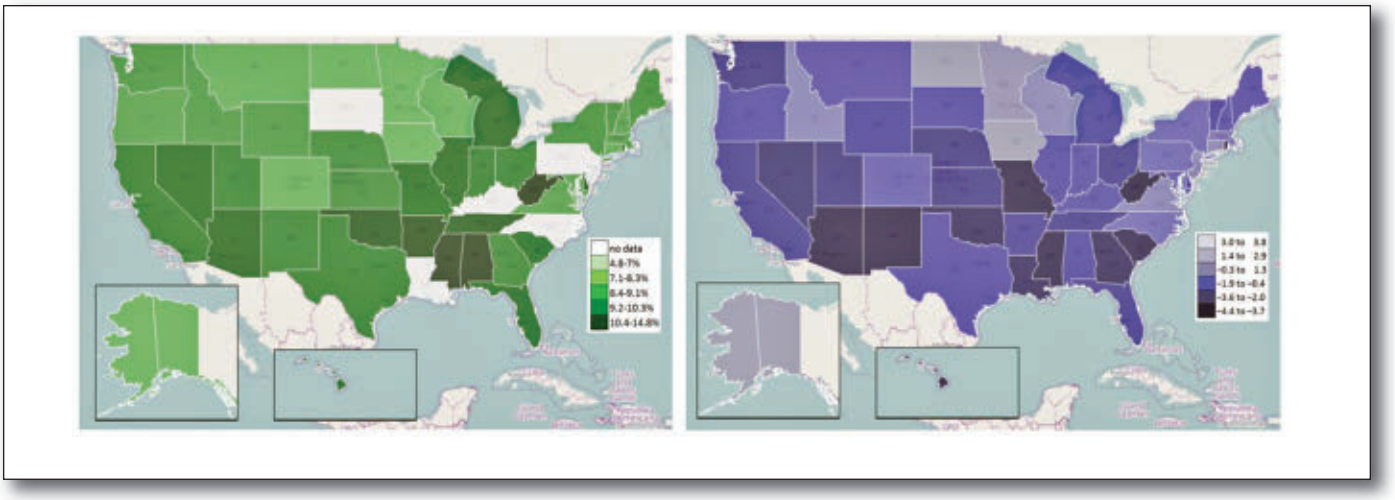


Figure 1. Social Media Index of Depression Compared to Self-Reported Survey Data.

Our prior work showing (on the left) heatmap rendering of actual CDC data and (on the right) Twitter-predicted depression in various US states (De Choudhury, Counts, and Horvitz 2013). Note that in both figures, higher intensity colors imply greater depression. A linear regression fit between the actual and predicted rates shows positive correlation of 0.51.

leverage human intelligence in their social media data modeling and analyses lies in gathering gold standard information that can later be employed in supervised learning models. This gold standard information often also acts as a test of the construct validity of the underlying measures. In the domain of mental health, this approach translates to compiling ground truth information about the true mental health states of individuals, communities, and populations that is independently assessed beyond what the AI techniques may provide.

In our prior work, we have extensively utilized this form of human feedback. For instance, we used crowdsourcing, particularly through the Amazon Mechanical Turk platform, to collect (gold standard) assessments from several hundred (nearly 400) Twitter users who reported that they have been diagnosed with clinical depression, using the CES-D (Center for Epidemiologic Studies Depression Scale)(Eaton et al. 2004) screening test (De Choudhury et al. 2013). Based on this cohort for whom we had offline assessments of depression, we developed several affective, behavioral, cognitive, linguistic, and domain-specific measures and used them to develop AI techniques that quantify an individual's social media behavior for a year in advance of their reported onset of depression (as assessed from their offline psychometric data). Then we leveraged these multiple types of signals from these measures to build a depression classifier that distinguished an at-risk cohort from a control group, and was able to predict, ahead of onset, whether an individual is vulnerable to depression. Our models show promise in predicting outcomes with an accuracy of 70 percent and precision of 0.74.

Further, we evaluated this model by comparing it with gold standard offline statistics of prevalence of depression in the United States (De Choudhury, Counts, and Horvitz 2013). As shown in figure 1, we found our social media index of depression compared well with the rates, obtained via self-reported survey data, as given by the Centers for Disease Control and Prevention. Similar approaches were used in other work from our team. This work includes research that developed AI techniques to leverage Facebook data and self-reported information to predict risk of postpartum depression in new mothers (De Choudhury et al. 2014), shown in figure 2, and that employed expert-generated clinical appraisals from clinical psychologists and psychiatrists to assess and curate the quality of online data related to schizophrenia and psychosis (Birnbaum et al. 2017). In the latter, in particular, expert feedback and ground truth on psychosis allowed us to situate the trends and patterns derived from individuals' social media data into what is known about the illness, its diagnosis, and its trajectory over time. We found that compared to a control group, the psychosis cohort exhibited marked linguistic changes on Twitter in the period following their self-disclosure of their illness on Twitter. After verifying these changes with expert annotations, we found the post-disclosure period to be characterized by lowered stereotypy such as word repetitiveness (–24 percent) and linguistic complexity (–63 percent) and by increased readability (+47 percent) and topical coherence (+81 percent). Figure 3 illustrates these findings.

In a similar vein, in a different work (Chancellor et al. 2016), we employed feedback from clinical psychologists as gold standard information to develop an inference model for mental illness severity (MIS) in pro-eating disorder posts on Instagram. Instead of

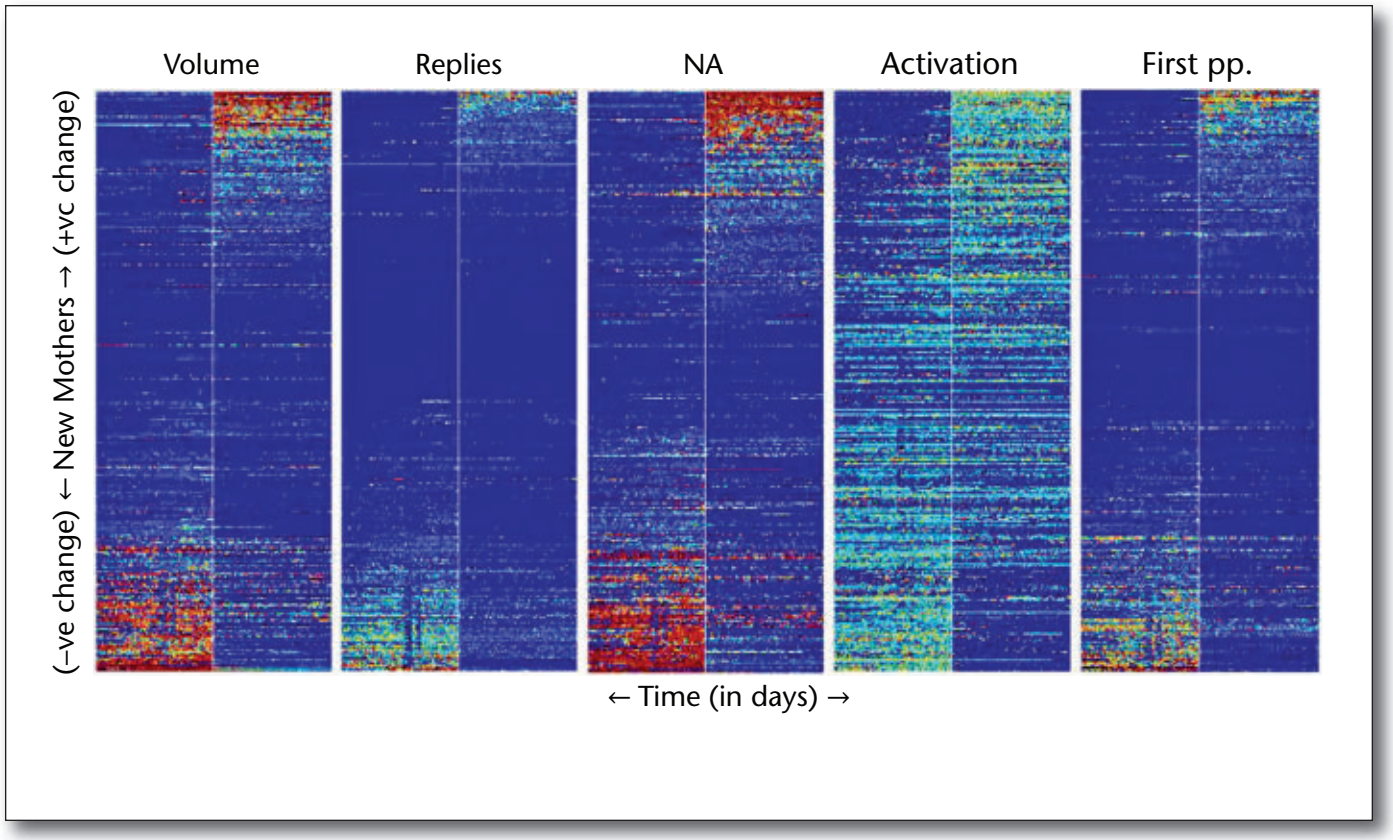


Figure 2. Facebook Data and Self-Reported Information to Predict Risk of Postpartum Depression.

Our prior work examining social media–based postpartum changes in activity, socialization, affect, and interpersonal attention of new mothers (De Choudhury et al. 2014). The heatmaps show individual-level changes in the postnatal period, compared to the prenatal phase. For 15 percent of mothers, these changes (for example, increase in NA and activation) are considerably higher following childbirth.

getting expert annotations on posts directly, a method that does not scale well to large datasets, we obtained them on outcomes of topic models. This strategy allowed us to scale our inference framework to a large corpus of Instagram posts, where we developed a semisupervised approach to map the labels on the topics to posts from users. Examples of high MIS content spans from expression of negative self-perceptions to disordered thoughts about eating to graphic illustration of acts that could lead to physical and emotional harm or death. This coincorporation of human feedback as gold standard information and of analytical AI-based data enabled deep explorations into the manifestation of MIS on the Instagram platform. We found that users who share pro-eating disorder content on Instagram exhibit a trend of increasing MIS in their content over time.

Interpreting Large-Scale Analysis

As we noted, AI approaches can be complemented with human feedback for interpreting the outcomes of an analysis or a computational model. Another way to combine AI methods with human intelligence is to have experts contextualize the AI findings in existing theory or theoretical/conceptual frameworks. By integrating knowledge from existing theories and frameworks, we can test our understanding of underlying mechanisms and our assumptions on unobserved factors that might be affecting our conclusions.

In prior joint work (De Choudhury et al. 2016), the authors developed a causal inference framework (Pearl 2009) to assess the likelihood that an individual will transition to discussions of suicidal ideation, given a history of mental health discourse on social media. This framework was developed on a large dataset of 880 users who shared more than 12K posts and 100K comments on the social media site Reddit. The output of the framework included words and phrases that indicated the likelihood of future suicidal ideation, given their usage in a post. Specifically, we applied a high-dimensional stratified propensity score method (Rosenbaum and Rubin 1983). This approach attempts to isolate the effects of a particular treatment from the effects of covariates by dividing the treatment (those who use a particular

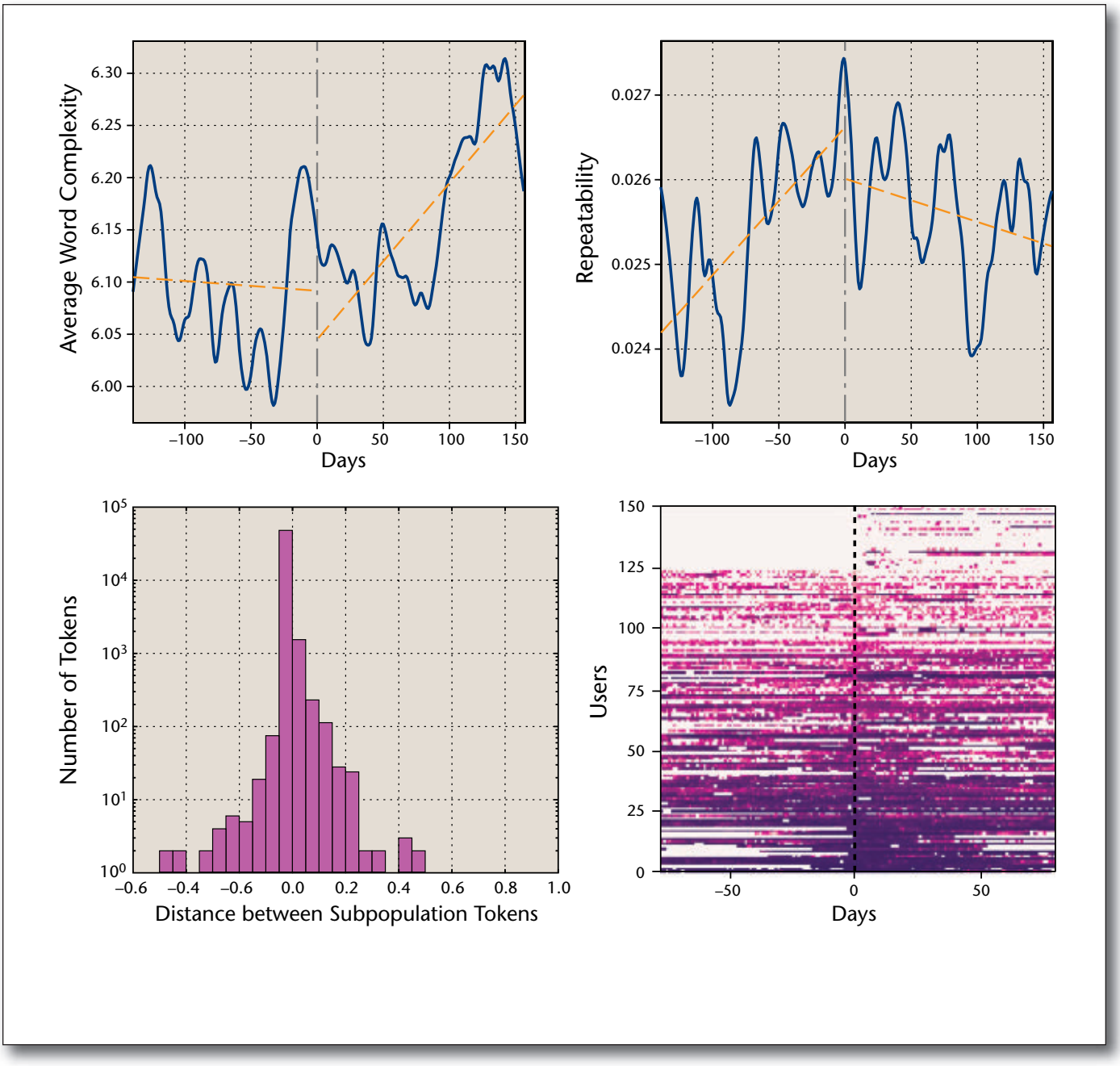


Figure 3. Post-Disclosure Period Changes.

Our work (Ernala et al. 2017) showing notable changes in linguistic organization in a clinically appraised psychotic population following diagnosis disclosures on Twitter (day 0 is day of disclosure).

word/phrase) and control groups (those who do not use the same word or phrase) into strata where the covariates of the treatment subgroup within a strata are statistically identical to the covariates of the control subgroup. Each strata is thus, in essence, artificially approximating a randomized controlled trial where the "assignment" of a treatment is statistically uncorrelated with covariates, allowing us to better distinguish the possible causal effects of a treatment on the outcome, in this case, being whether or not a specific Reddit user posted about suicidal ideation.

However, we noted that the linguistic cues, given by the aforementioned the causal framework (see table 1), did not allow us to examine how specific types of risk markers were associated with suicidal ideation, as illustrated in clinical psychology theories. To enable such comparison, we clustered these linguistic cues via spectral clustering to identify, via

expert annotations, what themes led to increases or decreases in suicidal ideation. Then, we qualitatively, using reviews from the same experts, interpreted these themes with the sociocognitive model of suicide (Rudd 1990), to understand what risk markers of suicide are manifested in social media, and to what extent the linguistic cue clusters align with what is known from existing theories in the psychology domain to exacerbate or alleviate the risk of suicidal ideation.

For instance, we found themes containing words or phrases like "have nothing," "no real," "kill myself," "abandoned," and "die" that experts noted to relate to signals of hopelessness among individuals. The cognitive psychological integrative model of suicide (Dieserud et al. 2001) has identified hopelessness as a mediating variable between mental illness and suicidal ideation and there is ample evidence of the decisive role of hopelessness as an indicator both of current suicide intent and as a predictor of future suicidal behavior (Kashden et al. 1993; Glanz, Haas, and Sweeney 1995):

> But I want to di*e*. I feel so *abandoned*. I must be *an idiot*. I hope for some random event to kill me so that nobody has to be guilty. My loved ones would mourn me but they would move on. At least easier than if I actively killed myself.

We also observed manifestation of impulsive tones in a different theme given by the spectral clustering approach and labeled by the experts. The cognitive suicide model also suggests that impulsivity resulting from cognitive deficits (for example, cognitive rigidity, dichotomous thinking, inability to generate or act on alternative solutions) are prominent markers of suicide ideation (Beck 1979; Kashden et al. 1993):

> Theres a terrible feeling through my whole body every waking moment I have and theres only 2 ways to *ending it*. It hasnt been getting better only worse, I am *freaking out*. The only thing stopping me is I dont know about/have access to anything that would make it quick and clean

Next, the cognitive suicide model has further found lowered self-esteem and self-efficacy to be important attributes among those who are prone to suicide ideation (Schwarzer and Fuchs 1995). Feelings of social isolation and loneliness, conceptualized as a part of the cognitive vulnerability, have consistently been shown to be related to suicidal ideation, attempts, and completions (Bonner and Rich 1988). We found that tokens of one of the extracted themes contained a tone of decreased self-esteem, including that of guilt, self-loathing, and regret:

> I am too ugly to even make friends. I *hate it*. People do not want to be associated with me because of my image. I have tried talking to girls and they've all told me to go away and to just give up. So here I am, *giving up* and ending everything.

Together, these findings demonstrate that when human involvement is sought in interpreting the outcomes of large-scale AI approaches, we obtain a much richer, grounded understanding of the specific problem context. Moreover, interpreting results within the context of existing theories also provides a way to test the ability of our conclusions to generalize beyond our specific analysis, to generalize beyond specific datasets, or to focus on specific social media sites.

Improving Computational Models

In this subsection describing the coutilization of artificial and human intelligence for mental health, we describe joint work of the authors in which human insights were incorporated to revise the outcomes of a computational AI-based framework. Such approaches can be a way to fill in the gaps left behind by use of AI techniques alone, especially those gaps that are attributed to the limited "view" on human behaviors and mental health allowed by AI approaches.

We briefly summarize such an approach from our prior work. Utilizing comments received on posts shared in Reddit mental health communities as a proxy for social support, in recent research (De Choudhury and De 2014; De Choudhury and Kiciman 2017), we developed a human-machine hybrid statistical methodology that modeled and quantified the effects of the language of these comments in individuals who do and do not post on a suicide support community in Reddit. Applying stratified propensity score matching (Caliendo and Kopeinig 2008) in a iterative fashion, similar to the approach previously described, we first identified linguistic features (words/phrases) in comments that showed significant effects. We realized that, while the comparability of posts is conventionally judged through purely statistical measures, in our domain these statistics over low-level textual features may miss higher-level semantics of the text. Our contribution lay in realizing that because treatment assignment (that is, the provision of social support) is performed by human commenters who are replying to posts, we can augment our statistical analysis of balance with expert human assessments of balance.

Thus, we obtained expert assessments on the presence of suicidal ideation risk markers in posts associated with these features, for which the rater relied on their offline understanding and knowledge of the risk markers of suicide. Across different propensity strata, the raters specifically assessed "balance." That is, if their expert assessment of risk markers of suicidal ideation aligned on pairs of posts in the same propensity strata, then we would infer the treatment and control groups for that particular linguistic feature and strata to be balanced. If not, we would assume that the groups in that strata are not comparable to each other and that our propensity score matching analysis needs further tuning to identify more accurately balanced treatment and control

Treatment Token	Count	Coverage	Treatment Effect	z	χ^2
Increased Change					
depression	318	0.901	0.3	8.04	7.78
useless	53	0.801	0.51	7.05	6.53
suicide	143	1	0.32	6.66	5.03
anxiety	216	1	0.24	6.56	4.11
suicidal	111	0.9	0.34	6.56	5.37
i_almost	40	0.901	0.52	6.44	4.22
and_an	45	0.7	0.51	6.4	6.15
medicine	41	0.8	0.52	6.38	4.86
unless_i	38	0.9	0.53	6.36	4.47
hug	42	0.8	0.52	6.36	4.9
money_i	35	0.801	0.52	5.89	3.96
out_as	34	0.701	0.53	5.89	4.76
this_happened	35	0.901	0.51	5.89	3.72
this_world	37	0.8	0.5	5.88	4.17
over_i	35	0.901	0.51	5.86	3.58
still_a	36	0.7	0.51	5.85	4.68
off_a	35	0.801	0.51	5.85	4.24
loneliness	37	0.8	0.5	5.84	3.99
class_and	34	0.901	0.52	5.84	3.39
alone_i	77	1	0.34	5.84	3.91
Decreased Change					
captain	11	0.4	-0.6	-4	4.24
differences	16	0.601	-0.57	-4.47	3.56
the_trip	11	0.601	-0.57	-3.76	3.2
intimate	11	0.501	-0.57	-3.73	2.93
to_in	20	0.701	-0.56	-4.92	4.1
too hard	16	0.601	-0.56	-4.4	3.56
suspect	16	0.701	-0.56	-4.4	3.04
always a	14	0.601	-0.56	-4.15	3.29
be_working	14	0.601	-0.56	-4.12	2.73
keep your	12	0.601	-0.56	-3.82	2.46
straight up	12	0.601	-0.56	-3.82	2.38
preferred	11	0.601	-0.56	-3.71	2.43
awesome_i	11	0.501	-0.56	-3.68	2.86
s_at	21	0.801	-0.55	-4.83	3.33
stated	20	0.801	-0.55	-4.8	3.66
slight	18	0.701	-0.55	-4.61	3.3
and_enjoy	17	0.601	-0.55	-4.44	3.48
gotten_to	16	0.7	-0.55	-4.35	2.77
it_work	15	0.501	-0.55	-4.22	4.17
came_from	15	0.701	-0.55	-4.21	2.76

Table 1. Linguistic Cues.

Statistically significant treatment tokens obtained via propensity score matching that contribute to increased as well as decreased change in likelihood of posting in SW.

Token	Strata	Treatment Post	Control Post
High Propensity Strata			
not easy	6	a reason behind my depression is how small by body frame is. i've never cared much about muscle but it's obviously one of the reasons i've been alone (friendships and relationships) for my whole life.	i'm aware there's no way to avoid pain 100%, which is why i'm attempting to go for the least painful way. we've talked in detail about exactly why our issues are troubling for each of us, so he knows that already
advice but	6	i don't even know what all i feel. ashamed, angry, at myself and at the family that never did a thing to support me before. i'm seriously thinking about just pulling out i'm tired of trying, and failing, over and over again.	feeling like shit but noone to talk to, just need a friend who can cheer me up. noones online on facebook that i can talk to so just alone right now ...
Low Propensity Strata			
seek	2	i realize that i'm having depression. i have not showered for a week now, unable to sleep and always thinking negative about myself	i noticed during the livestream, even though that he wasn't using their (i'm assuming) condenser microphone, i felt that his volume and the tones of his voice sounded much more "comfortable" with the headset.
slow down	1	an american football fan but i am intrigued by the world cup. i remember watching 4 years ago and was fascinated. does anyone know of a quality app i can get on my phone that i can use to keep up with it?	greetings people, greetings people, i am a worthless nobody. i guess i want to take more of your time in the vain hopes that you'll somehow be able to make me feel better.

Table 2. Qualitatively Assessed Post Pairs and Associated Comment Tokens.

Post pairs and associated comment tokens qualitatively assessed to correspond to balanced and imbalanced treatment and control groups. Text has been slightly paraphrased to protect the identities of the users.

groups. Once the unbalanced strata were identified by the human raters, then we filtered the posts to the corresponding comparable subpopulations. We then modified our method to compute a local average treatment effect only over the strata deemed to be balanced by human assessments, so as to assess the effects of specific linguistic features of comments in future risk to suicidal ideation

With the help of these human assessments and as shown in table 2, we found that the effects of getting a token in a comment may not be homogeneous. Certain users may see little effect of getting a token (low-propensity strata), while others see a large effect (higher-propensity strata). By employing human raters in this task, we showed how the outputs of causal inference methods can be amalgamated with expert feedback to improve results.

The fact that we obtain better results by incorporating human feedback in an AI task like the one previously described is further clarified while investigating the context of use of specific linguistic tokens in comments, and situating those tokens in theoretical framework of social support. In the users who show reduced likelihood of suicidal ideation in our dataset, we found comments on their posts to contain greater expression of esteem (31 percent) and network support (23 percent), followed by emotional support (16 percent). Informational support (9 percent) and acknowledgments (5 percent) were relatively lower for comments containing tokens that decrease the likelihood of posting about suicidal thoughts. Overall, this distribution indicates the positive impact of esteem and network support in reducing one's future risk of suicidal ideation expression, a result which can be accurately inferred by filtering the outcomes of causal inference based on human feedback.

Implications

Our research shows that, while AI approaches have made and continue to make significant strides into domains like mental health, the involvement of natural intelligence in the form of human feedback is critical to the success of these efforts. Given the complexities and sensitivities of this domain, human insights and the integration of domain knowledge can situate the efforts in existing research, theory, and what is needed for further validation of insights with carefully designed experiments and empirical study designs. Importantly, for the same reasons of domain complexity and sensitivity, we caution against automatic deployment of the described AI approaches and emphasize that human involvement will help translate their potential benefits to real-world mental health context — similar arguments have been made earlier (Amershi et al. 2014) as well as more recently in domains outside of mental health. In the paragraphs that follow, we discuss some of these mixed-initiative, human-machine partnered implications of this research.

Clinical and Self-Reflection Interventions

The approaches that we discuss in this article can have widespread implications for the mental health clinician community. Currently, there is limited ability to aid chronic mental illness management (Simon and Ludman 2009). Patient-reported experiences in the form of clinical interviews and questionnaires have played a central role in management of these conditions for more than a century (Liberman 1988). These approaches do not include evidence-based assessments — behavioral, emotional, or cognitive symptoms must be recalled from a patient's memory — a method prone to retrospective recall bias. Time and budgetary constraints further limit psychiatrists from conducting more thorough and frequent in-person evaluations. These constraints preclude time-sensitive and objective monitoring of symptoms, and an ability to detect subtle and burgeoning changes that may not surface in patients' self-reports. With the human-machine mixed initiative approaches we presented here, technologies can be developed that allow clinicians to monitor patients' symptoms and identify patterns that may be harbingers of adverse health events in the future. This way, clinicians will be able to engage in evidence-based decision-making, beyond what is possible within the realms of in-person therapeutic settings. To reiterate, the involvement of humans, in this case stakeholders like the patient and the clinicians, is critical to ensure that the technologies function in a way that is accountable, interpretable, actionable, and transparent.

Interventions may also be designed, based on the human-machine approaches previously discussed, that promote self-reflection of one's (for example, a patient's) activity and behavior around mental health on various social media platforms. Our methods might be employed for the self-assessment of behavior, cognition, and affect, or might serve as an early warning mechanism for individuals struggling with mental health concerns. Reflective interventions, guided by an expert, such as a support network member or a clinician, could also be designed to reveal longitudinal trends relating to specific mental health markers, such as that of suicide ideation. Such an intervention might be used for instance, to identify time periods of anomalous patterns, which are known to be otherwise difficult for individuals to keep track of. The logging of these longitudinal trends can also serve as a diary-style data source to help caregivers or other trained professionals and clinicians gain a deeper understanding of an individual's risk for dangerous behaviors in the future.

Social Media Interventions

Individuals whose posted content contains phrases and other linguistic constructs relating to mental health risk, as revealed by our AI-based methods, may be flagged in the interfaces of moderators and other clinical experts for help and support, thereby involving a human "in the loop." Community moderators, support volunteers, and the social media platform creators and owners themselves may also be allowed to maintain a "risk list" in their interfaces that would include individuals forecasted by our AI methods to exhibit signs of increased risk in the future. Such a list may sort or rank individuals based on their forecasted risk score. This approach would allow improved preparedness on the part of the moderators, platform owners, and experts to bring timely and appropriate help to those in need. Further, on being informed that an individual could be prone to risk in the future, moderators and experts may make provisions to connect them with appropriate mental health resources (for example, a web-based hotline like Crisis Text Line, or a community like 7 Cups of Tea2), encouraging peers or trusted friends and family, or field private messages with relevant information on help seeking or therapy.

Finally, our work also includes implications for volunteers intending to provide social support to vulnerable groups on social media. Applications could be developed, leveraging the human-machine collaborative techniques we discussed, that continually educate such volunteers to be self-aware and learn about what kind of information is perceived to be beneficial to social media users seeking help and support around mental health challenges.

AI Implications

Our work also raises questions about the challenges that AI as a field faces in realizing these domain implications. Such questions largely pertain to how the AI tools are used and in what ways human intelligence can alleviate some of the challenges

that arise in real-world deployment of these tools and approaches. We discuss two such aspects.

Guarding Against Errors and Negative Outcomes

Many AI approaches, including some of the ones described in this article, have been considered as inscrutable black boxes of decision-making (Horvitz 2017). Improvements in computational power coupled with the availability of large volumes of training data (such as from social media) — data used to train AI models that infer mental health state — are driving advancements in machine learning, which are reinforcing the black-box phenomenon. In fact, in the context of mental health, neural networks are becoming an increasingly popular way of making predictions of a variety of symptoms and attributes (Chancellor et al. 2017; Manikonda and De Choudhury 2017; Reece and Danforth 2017). Neural networks, or largely, deep learning, are so opaque that it is practically impossible to understand what they deduce from training data and how they reach their conclusions — making it hard to judge their correctness in a domain where accuracy is critical to human life. Thus, these advances in machine learning techniques are enabling the creation of black-box AI approaches that, although they have better predictive power, are significantly more complex, especially to a layperson like a patient or a clinician, and are also less interpretable or explainable, again to the same layperson, who is likely to benefit the most from their outcomes.

Black boxes are also vulnerable to risks, such as accidental or intentional biases, errors, and frauds, thus raising the question of how to "trust" these systems and tools that make important and sensitive inferences about an individual's mental health state. Incorrect interpretation of the output of these systems (for example, what does mental health risk really mean), inappropriate use of the output (for example, using them directly in diagnosis or treatment), and disregard of the underlying assumptions (for example, that every individual is different, and so is their social media use and mental health state) can have drastic consequences. Involving humans can help correct some of these biases, and a "human" face to AI systems that make predictions about a person's behavior and mental health is likely to be more trustworthy to stakeholders.

Privacy and Ethics

Our work also raises important questions relating to privacy and ethics, questions that pose vexing complexities to the variety of stakeholders who are likely to be impacted by this research. Again, the involvement of experts and other individuals "in the loop" or toward the general functioning of AI systems will be helpful in tackling and addressing some of these challenges.

Mental Health Counselors and Clinicians. While our work provides new opportunities for mental health clinicians and counselors to learn what factors and attributes might precipitate risk for states such as depression and suicidal ideation, it also raises important ethical obligations. In a typical therapeutic setting, a clinician has control of the information that is sought, gathered, and used. The inferences and assessments about a patient's mental health are also made by the clinician themselves, by incorporating their understanding of a patient's state as well as other types of relevant collateral information. However, when these inferences are made by an algorithm, what should be the clinician's response, how should they act on this information? How can they navigate the therapeutic relationship with a patient, in the face of information delivered through an AI tool, while respecting the patient's privacy needs and therapeutic expectations?

In essence, AI-based technology provides an unprecedented opportunity to engage people both outside traditional mental healthcare settings and far earlier in the course of illness (Baumel et al. 2018). Capitalizing on this opportunity, however, requires stakeholders like clinicians and counselors to challenge underlying assumptions about traditional pathways to mental health treatment and care. Further, AI approaches to identifying illness and tracking symptoms will need human feedback with respect to redefining existing clinical rules and regulations. Although the potential beneficial impact of AI technology integration could be transformative, new critical questions regarding clinical expectations and responsibilities will require resolution.

Social Media Platform Owners, Designers, Moderators, and Participants. The techniques we presented could allow moderators, support volunteers, and owners and designers of social media platforms to make improved decisions and choices based on forecasted likelihood of risk. However, when inferences and assessments are made by an AI system instead of solely a human, what are the obligations for the moderators, the volunteers, the platform creators, or the community as a whole when they discover an individual to be at a higher likelihood of risk for behavior such as suicidal ideation? How can social media sites reap the benefits of our method and gain from the design opportunities outlined, while at the same time protect their ethical obligation to act upon situations that may need an intervention? We also envision ethical questions regarding revealing to social media users the implications of the use of certain type of language or certain patterns of activity, or surfacing to them inferred risk measures. To this end, interventions would require careful consideration because there is a delicate line between overintrusiveness and concern (in AI terms, balancing false positive and false negative rates). As noted in our prior work, further research is needed to better define the trajectory between online activity and making first clinical contact to explore opportunities for digital intervention (Birnbaum et al. 2017). In fact, eth-

ical challenges go beyond the uses of the outcomes of AI technology. Without appropriate human involvement, due to the Hawthrone effect (McCarney et al. 2007), stigma, or other self-censorship reasons, over time individuals may eventually refrain from offering cues that might reveal their risk. How can social media sites, then, continue to be platforms of authentic expression and a means that enable disclosure of deep-seated mental health concerns? How can AI tools leverage human feedback in ways that ameliorate these self-censorship challenges?

One way to tackle these challenges could be to thoroughly assess the acceptability of our method or the technologies it enables to different stakeholders, thus incorporating human feedback into the design and functioning of the AI systems. This strategy constitutes a promising direction for future research. Collaborations between AI researchers, mental health experts, community moderators, designers, developers, social media companies, and ethicists can also help develop protocols and guidelines that facilitate the use of our work in practical contexts in the future.

Conclusion

In this article, we presented a discussion of the role of human involvement in deriving meaningful value out of AI techniques and approaches. We highlighted work from several threads of our prior research to describe this agenda, particularly focusing on the domain of mental health.

To conclude, the underlying impetus for investigating these types of problems of societal significance with AI is, of course, the desire to help people improve their (here, mental health) outcomes, whether through early identification of people at risk, better personalization of treatments, or discovery of new treatment strategies. Bridging the gap between insights derived from AI approaches and real-world action will require combining the outcomes of the approaches with human feedback, interventions, and simultaneous human/empirical observations to provide strong validations of benefits. The challenges posed in moving from AI outcomes to intervention in social media platforms are particularly exacerbated in sensitive domains — for example, how to get informed consent from very large populations when it comes to mental health assessments or how to ensure interventions that avoid real-world harm while respecting privacy of individuals online. It will be a significant challenge to develop new protocols that safely translate insights from observational studies of AI methods/tools, to active experimentation involving expert feedback, and then to large-scale deployments involving real people, while simultaneously respecting principles of individual autonomy, minimizing risk of harm, and ensuring that benefits and risks are distributed across all parties who are directly or indirectly, positively or less beneficially, affected by the underlying AI.

References

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35(4): 105–20. doi.org/10.1609/aimag.v35i4.2513.

Baumel, A.; Baker, J.; Birnbaum, M. L.; Christensen, H.; De Choudhury, M.; Mohr, D. C.; Muench, F.; Schlosser, D.; Titov, N.; and Kane, J. M. 2018. Summary of Key Issues Raised in the Technology for Early Awareness of Addiction and Mental Illness (TEAAM-I) Meeting. *Psychiatric Services* 69(5): 590-92. doi.org/10.1176/appi.ps.201700270.

Beck, A. T. 1979. *Cognitive Therapy of Depression*. New York: Guilford Press.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research* 19(8): e289. doi.org/10.2196/jmir.7956.

Bonner, R. L., and Rich, A. 1988. Negative Life Stress, Social Problem-Solving Self-Appraisal, and Hopelessness: Implications for Suicide Research. *Cognitive Therapy and Research* 12(6): 549–56. doi.org/10.1007/BF01205009.

Boyd, D., and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, and Society* 15(5): 662–679. doi.org/10.1080/1369118X.2012.678878.

Caliendo, M., and Kopeinig, S. 2008. Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* 22(1): 31–72. doi.org/10.1111/j.1467-6419.2007.00527.x.

Chancellor, S.; Kalantidis, Y.; Pater, J. A.; De Choudhury, M.; and Shamma, D. A. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 3213–3226. New York: Association for Computing Machinery. doi.org/10.1145/3025453.3025985.

Chancellor, S.; Lin, Z. J.; Goodman, E.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work, and Social Computing*, 626–38. New York: Association for Computing Machinery. doi.org/10.1145/2818048.2819973.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the Fifth Annual ACM Web Science Conference,* 47–56. New York: Association for Computing Machinery. doi.org/10.1145/2464464.2464480.

De Choudhury, M.; Counts, S.; Horvitz, E.; and Hoff, A. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing,* 626-38. New York: Association for Computing Machinery. doi.org/10.1145/2531602.2531675.

De Choudhury, M., and De, S. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *Proceedings of the Eighth International Confer-*

ence on Weblogs and Social Media, 71–80. Palo Alto, CA: AAAI Press.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression via Social Media. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*, 128–37. Palo Alto, CA: AAAI Press.

De Choudhury, M., and Kiciman, E. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk. In *Proceedings of the 11th International Conference on Web and Social Media*, 32–41. Palo Alto, CA: AAAI Press.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 2098–110. New York: Association for Computing Machinery. doi.org/10.1145/2858036.2858207.

Dieserud, G.; Røysamb, E.; Ekeberg, Ø.; and Kraft, P. 2001. Toward an Integrative Model of Suicide Attempt: A Cognitive Psychological Approach. *Suicide and Life-Threatening Behavior* 31(2): 153–68. doi.org/10.1521/suli.31.2.153.21511.

Dietterich, T. G. 2009. Machine Learning in Ecosystem Informatics and Sustainability. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 8–13. Palo Alto, CA: AAAI Press.

Eaton, W. W.; Smith, C.; Ybarra, M.; Muntaner, C.; and Tien, A. 2004. Center for Epidemiologic Studies Depression Scale: Review and Revision (CESD and CESD-R). In *Instruments for Adults*. Vol. 3 of *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment,* 3rd ed., edited by M. E. Maruish, 363–77. New York: Routledge.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. In *Proceedings of the ACM Human Computer Interaction* 1(2): article 41. New York: Association for Computing Machinery. doi.org/10.1145/3134678.

Glanz, L. M.; Haas, G. L.; and Sweeney, J. A. 1995. Assessment of Hopelessness in Suicidal Patients. *Clinical Psychology Review* 15(1): 49–64. doi.org/10.1016/0272-7358(94)00040-9.

He, J.; Bailey, J.; Rubinstein, B. I.; and Zhang, R. 2015. Identifying At-Risk Students in Massive Open Online Courses. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence,* 1749–55. Palo Alto, CA: AAAI Press.

Horvitz, E. 2017. AI, People, and Society. *Science* 357(6346): 7. doi.org/10.1126/science.aao2466.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science* 353(6301): 790–94. doi.org/10.1126/science.aaf7894.

Kashden, J.; Fremouw, W. J.; Callahan, T. S.; and Franzen, M. D. 1993. Impulsivity in Suicidal and Nonsuicidal Adolescents. *Journal of Abnormal Child Psychology* 21(3): 339–53. doi.org/10.1007/BF00917538.

Liberman, R. P. 1988. *Psychiatric Rehabilitation of Chronic Mental Patients*. Washington, DC: American Psychiatric Press. doi.org/10.1176/ps.39.8.893.

Manikonda, L., and De Choudhury, M. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 170–81. New York: Association for Computing Machinery. doi.org/10.1145/3025453.3025932.

McCarney, R.; Warner, J.; Iliffe, S.; Van Haselen, R.; Griffin, M.; and Fisher, P. 2007. The Hawthorne Effect: A Randomised, Controlled Trial. *BMC Medical Research Methodology* 7(1): 30. doi.org/10.1186/1471-2288-7-30.

Pearl, J. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3: 96–146. doi.org/10.1214/09-SS057.

Pearl, J. 2018. Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution. arXiv preprint arXiv:1801.04016 [cs.LG]. Ithaca, NY: Cornell University Press.

Reece, A. G., and Danforth, C. M. 2017. Instagram Photos Reveal Predictive Markers of Depression. *EPJ Data Science* 6(1): 15. doi.org/10.1140/epjds/s13688-017-0110-z.

Rosenbaum, P. R., and Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1): 41–55. doi.org/10.1093/biomet/70.1.41.

Rudd, M. D. 1990. An Integrative Model of Suicidal Ideation. *Suicide and Life-Threatening Behavior* 20(1): 16–30.

Schwarzer, R., and Fuchs, R. 1995. Changing Risk Behaviors and Adopting Health Behaviors: The Role of Self-Efficacy Beliefs. In *Self-Efficacy in Changing Societies,* edited by A. Bandura, 259–88. Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511527692.

Simon, G. E., and Ludman, E. J. 2009. It's Time for Disruptive Innovation in Psychotherapy. *The Lancet* 374(9690): 594–95. doi.org/10.1016/S0140-6736(09)61415-X.

Vasisht, D.; Kapetanovic, Z.; Won, J.; Jin, X.; Chandra, R.; Sinha, S. N.; Kapoor, A.; Sudarshan, M.; and Stratman, S. 2017. Farmbeats: An IOT Platform for Data-Driven Agriculture. In *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation,* 515–29. Berkeley, CA: Advanced Computing Systems Association.

Munmun De Choudhury is an assistant professor in the School of Interactive Computing at the Georgia Institute of Technology where she directs the Social Dynamics and Wellbeing Lab. Prior to joining Georgia Tech, De Choudhury was a faculty associate with the Berkman Klein Center for Internet and Society at Harvard and a postdoc at Microsoft Research, following obtaining her PhD in computer science from Arizona State University. De Choudhury's research interests are in computational social science. With her students and collaborators, De Choudhury focuses on developing computational methods to assess, understand, and improve personal and societal mental health from online social interactions.

Emre Kiciman is a principal researcher at Microsoft Research. Kiciman's research interests include social computing and computational social science, causal inference methods, and information retrieval. His current work focuses on causal analysis of large-scale social media timelines, using social data to support individuals and policymakers across a variety of domains, and more broadly on the implications of AI on people and society. Kiciman's past research includes entity-linking methods for social media and the web, deployed in the Bing search engine; and foundational work on applying machine learning to fault management in large-scale internet services.

AAAI News

Fall News from the Association for the Advancement of Artificial Intelligence

AAAI Honors High School Students at Intel ISEF

AAAI is pleased to announce the winners of the recent AAAI Special Awards at the Intel International Science and Engineering Fair, held May 13-18, 2018 in Pittsburgh, Pennsylvania. The winners of the AAAI honors were as follows:

First Prize

Nikita Zozoulenko (Linköping, Sweden) for Dense Face Detection and Improving Temporal Convolutional Networks for Automatic Image Captioning.

Second Prize

Matthew Dong and Pratham Soni (Troy, Michigan, USA) for Context Aware Medical Image Super Resolution Using Convolutional Neural Networks.

Third Prize

Han Qi (Tianjin, China) for Changing the Ratio of an Image Intelligently According to Its Contents: An Image Processing Tool Based on Pixel Weight and Face Detection.

Honorable Mention

Honorable Mention went to the following students:

Alice Martynova (Los Gatos, California, USA) for An Affordable, Autonomous, AI-Enhanced Microscope to Enable Efficient Diagnosis of Parasitic Infection in Developing Countries.

Kavya Kopparapu (Alexandria, Virginia, USA) for GlioVision: A Platform for the Automatic Assessment of Glioblastoma Tumor Features, Molecular Identity, and Gene Methylation from Histopathological Images Using Deep Learning.

David Lyons (Gates Mills, Ohio, USA) for Automatic Contouring Methods for Adaptive Radiotherapy in Cancer Patients Using Artificial Intelligence and a Virtual Mobile Robotic Assistant.

Savitha Srinivasan (Bellevue, Washington, USA) for Development of Semi-Supervised Machine Learning Models to Predict Enhancer Regions in Polygenic Developmental Diseases.

Mihir Patel (Alexandria, Virginia, USA) for Optimizing Reinforcement Learning Through Dynamic Environment Manipulation.

AAAI thanks Stephen Smith (head judge), Laura Barbulescu, Jean Oh, and Zachary Rubinstein, all from Carnegie Mellon University, who served as AAAI's judges at the Intel event, for their generous donation of time and effort.

The Intel International Science and Engineering Fair, a program of Society for Science and the Public, is the world's largest precollege science competition, and includes more than 1,800 high school students from more than 75 countries, regions and territories. Society for Science and the Public, a nonprofit organization dedicated to public engagement in scientific research and education, owns and has administered the International Science and Engineering Fair since its inception in 1950. In 1958, the competition became international for the first time when Japan, Canada and Germany joined.

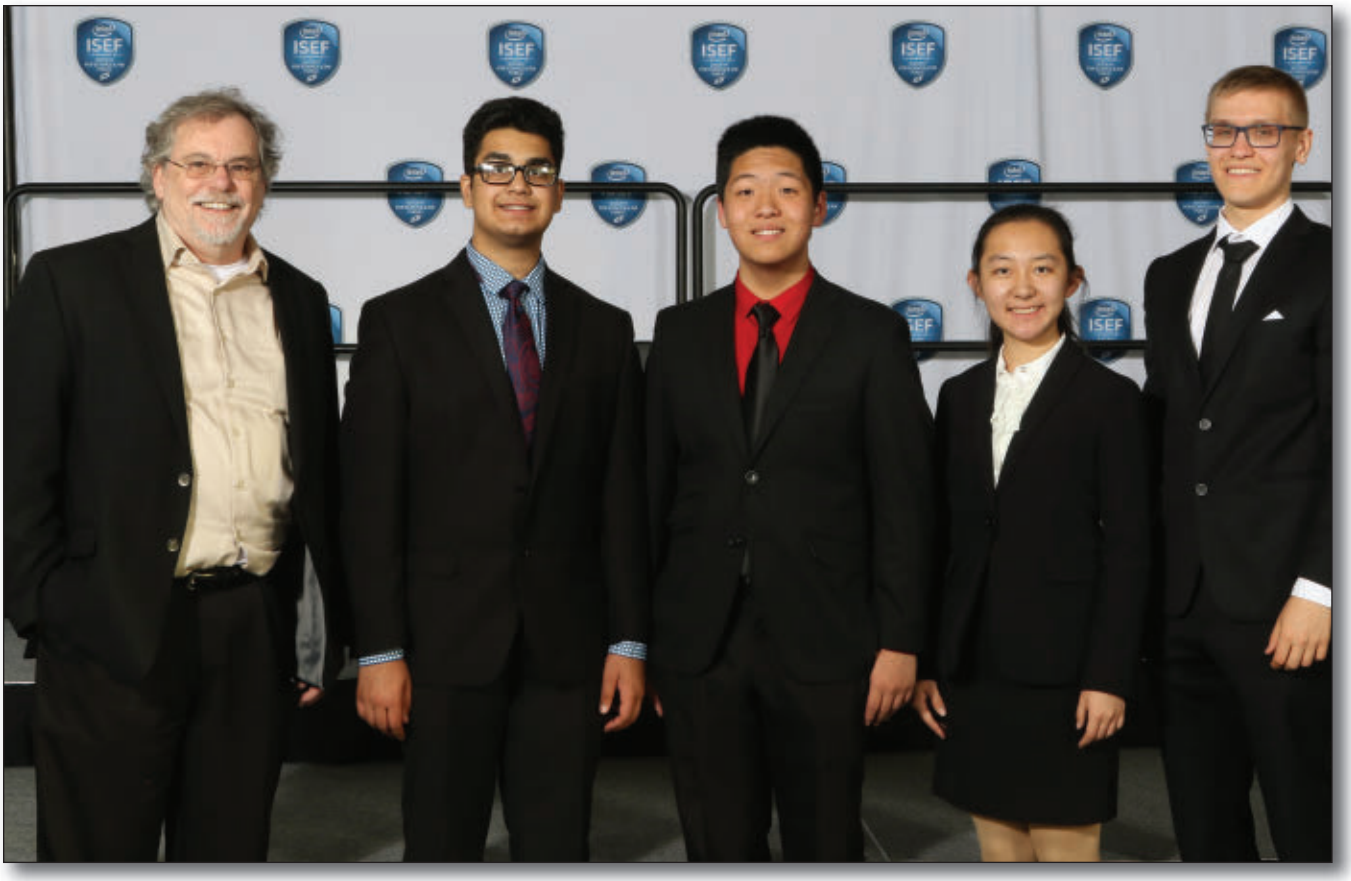
The Intel International Science and Engineering Fair 2018 is funded jointly by Intel and the Intel Foundation with additional awards and support from dozens of corporate, academic, governmental and science-focused organizations. To learn more about Society for Science and the Public, visit www.societyforscience.org.

AAAI and ACM to Cosponsor Second Conference on AI, Ethics, and Society

AAAI is pleased to announce the continuation of its collaboration with ACM in cosponsoring the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society. The second AAAI/ACM Confernce on AI, Ethics, and Society (AIES 2019) will be held January 27–28, 2019, in Honolulu, Hawaii, USA

As AI is becoming more pervasive in our lives, its impact on society is becoming more significant and concerns and issues are arising regarding aspects such as value alignment, data handling and bias, regulations, and workforce displacement. Only a multidisciplinary and multistakeholder effort can find the best ways to address these concerns, including experts in various disciplines, such as ethics, philosophy, economics, sociology, psychology, law, history, and politics. This conference is designed to address these issues in a scientific context.

The first meeting of this conference in 2018 attracted over 250 partici-



AAAI 2018 Special Awards Recipients at the Intel International Science and Engineering Fair.

Left to right: Stephen Smith (Judge), Pratham Soni, Matthew Dong, Han Qi, and Nikita Zozoulenko.

pants. The program of the conference will include peer-reviewed paper presentations, invited talks, panels, and working sessions. Papers submitted to the conference should address questions related to any of the topics listed with a rigorous scientific approach. We expect papers submitted by researchers of all the disciplines involved. For complete information and the full call for papers, please see www.aies-conference.com. Abstract submission is required by November 2, followed by full paper submission by November 5.

AAAI-19 Student Abstract and Poster Program

The goal of the AAAI-19 Student Abstract and Poster program is to provide a forum in which students can present and discuss their work during its early stages, meet some of their peers who have related interests, and introduce themselves to more senior members of the field. The program is open to all students at the Undergraduate, Masters, and Doctoral levels. Abstracts are due September 21, 2018. See aaai.org/Conferences/AAAI-19/aaai19studentcall.

AAAI-19 Demonstrations Program

The AAAI-19 Demonstrations Program is intended to foster discussion and exchange of ideas among researchers and practitioners from academe and industry by presenting software and hardware systems and research prototypes of such systems, including their capabilities and workings. Accepted demonstrations will be allocated one time slot during one of the main conference evening poster programs, and will have a short paper included in the proceedings. Submissions from everyone, including authors of paper submissions to AAAI, IAAI, and AAAI-19 workshops, are encouraged.

Short papers and video or slides are due September 21, 2018.

AAAI-19 Workshop Program

The AAAI-19 workshop program includes the following 16 workshops covering a wide range of topics in artificial intelligence:

Affective Content Analysis

Agile Robotics for Industrial Automation Competition

Artificial Intelligence for Cyber Security

Artificial Intelligence Safety

Dialog System Technology Challenge

Engineering Dependable and Secure Machine Learning Systems

Games and Simulations for Artificial Intelligence

Health Intelligence

Knowledge Extraction from Games

Network Interpretability for Deep Learning

Plan, Activity, and Intent Recognition

Reasoning and Learning for Human-Machine Dialogues

Reasoning for Complex Question Answering

Recommender Systems Meet Natural Language Processing

Reinforcement Learning in Games

Reproducibility in AI

The AAAI-19 Workshop Call for Participation is now available at aaai.org/Conferences/AAAI-19/ws19 call. The recommended date for workshop submissions is November 5, unless otherwise noted at the individual workshop websites. Submission requirements vary for each workshop. Please consult the individual workshop description for complete information about where to submit your paper and a link to the workshop supplementary website, where more detailed information will be available.

Be an AAAI Sponsor, Exhibitor, and Job Fair Participant!

AAAI invites you to participate as a sponsor, exhibitor, and job fair employer at AAAI-19. Your participation will give you instant visibility to this diverse group of AI professionals, representing a host of research areas such as search, planning, knowledge representation, reasoning, natural language processing, robotics and perception, multiagent systems, statistical learning, and deep learning, as well as applications in diverse domains such as healthcare, sustainability, transportation, and commerce.

In 2019, AAAI will highlight research in the emerging area of artificial intelligence for social impact. Sponsors and exhibitors enjoy a host of benefits, including complimentary technical registrations.

The AAAI/ACM SIGAI job fair is a

Photo courtesy iStock

Join Us in Honolulu for AAAI-19 / IAAI-19 / EAAI-19!

The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), the Thirty-First Conference on Innovative Applications of Artificial Intelligence (IAAI-19), and the Ninth Symposium on Educational Advances in Artificial Intelligence (EAAI-19) will be held January 27 — February 1 at the Hilton Hawaiian Village in Honolulu, Hawaii, USA.

Registration information will be available in early November on the AAAI website. The cut-off date for hotel reservations is December 25, 2018, but we encourage you to secure your room early. For more information about the AAAI-19 block of rooms at the Hilton Hawaiian Village, please see

www.aaai. org/Conferences/AAAI-19/hotel-and-travel.

The conference venue is located on the island of Oahu directly on Waikiki. Oahu offers a wealth of things to see and do. Explore the rich history of the Islands at one of the many museums or historical sites, take a hike in a lush tropical forest, or admire the unparalleled views from the conference site. Hilton Hawaiian Village offers a host of conveniences right on property, but is also close to a full range of shops and restaurants, stretching to Diamond Head at the other end of Waikiki. For complete information about options in Honolulu, please see the Hawaiian Islands visitors page at www.gohawaii. com.

Photo courtesy iStock

HCOMP-19 to Be Held at Skamania Lodge in Stevenson, Washington, USA

Please join us for the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019), to be held October 28–30 at the Skamania Lodge, located in the Columbia River Gorge National Scenic Area in the town of Stevenson, Washington, 45 minutes from Portland, Oregon. The resort offers a distinctly Pacific Northwestern experience, featuring a three-story stone fireplace and sweeping river views.

HCOMP is the premier venue for disseminating the latest research findings on crowdsourcing and human computation. While artificial intelligence (AI) and human-computer interaction (HCI) represent traditional mainstays of the conference, HCOMP believes strongly in inviting, fostering, and promoting broad, interdisciplinary research. This field is particularly unique in the diversity of disciplines it draws upon, and contributes to, ranging from human-centered qualitative studies and HCI design, to computer science and artificial intelligence, economics and the social sciences, all the way to digital humanities, policy, and ethics. We promote the exchange of advances in human computation and crowdsourcing not only among researchers, but also engineers and practitioners, to encourage dialogue across disciplines and communities of practice.

Please visit www.humancomputation.com/2019 for more details as they become available.

place for students and professionals looking for internships or jobs to meet with representatives from companies and academia in an informal meet-and-greet atmosphere. Past fairs have attracted more than 20 companies and hundreds of interested job seekers!

For complete details about all of these programs, please visit aaai.org/Conferences/AAAI-19 or write to AAAI at aaai19@aaai.org.

The preferred deadline for notification of intent to participate is October 15, 2018.

Program Chairs for AAAI-19, IAAI-19, and EAAI-19

Pascal Van Hentenryck (Georgia Institute of Technology, USA), and Zhi-Hua Zhou (Nanjing University, China) are the program chairs for AAAI-19.

Michael Wollowski (Rose-Hulman Institute of Technology, USA) and Nate Derbinsky (Northeastern University, USA) are the EAAI-19 Symposium Cochairs.

Karen Myers (SRI International, USA) is the IAAI-19 chair.

We hope to see you in Honolulu in January!

AAAI-19 Student Scholar and Volunteer Program

AAAI is pleased to announce the continuation of its Student Scholarship Program for 2019, which is cosponsored by AAAI and the *AI Journal*. The Student Scholar Program provides partial travel support for students who are full-time undergraduate or graduate students at colleges and universities; are members of AAAI; submit papers to the conference program or letters of recommendation from their faculty advisor; and submit scholarship applications to AAAI by November 15, 2018.

In addition, repeat scholarship applicants must have fulfilled the volunteer and reporting requirements for previous awards. In the event that scholarship applications exceed available funds, preference will be given to students who have an accepted technical paper, and then to students who are actively participating in the conference in some way. However, all eligible students are encouraged to apply.

The Student Volunteer Program is an essential part of the conference and student participation is a valuable contribution. Volunteers will support AAAI organizers in Honolulu. In 2019, a limited number of complimentary technical program registrations will be available for students who volunteer during the conference. Preference will be given to participating students for the volunteer positions. Local students or students not requiring travel assistance can apply for the Volunteer Program if openings are available. AAAI membership is required for eligibility. The deadline for volunteer applications is November 15, 2018.

For further information about the Scholarship Program or the Volunteer Program, please contact AAAI at scholars19@aaai.org. The application is available at aaaiforms.wufoo.com/forms/z1abmgdf0qjxv60/.

2018 AAAI Fall Symposium Series

The Association for the Advancement of Artificial Intelligence's 2018 Fall Symposium Series will be held Thursday through Saturday, October 18-20 at the Westin Arlington Gateway, Arlington Virginia, adjacent to Washington, DC. The titles of the six symposia are:

- Adversary-Aware Learning Techniques and Trends in Cybersecurity
- Artificial Intelligence for Synthetic Biology
- Artificial Intelligence in Government and Public Sector
- A Common Model of Cognition
- Gathering for Artificial Intelligence and Natural System
- Integrating Planning, Diagnosis and Causal Reasoning
- Interactive Learning in Artificial Intelligence for Human-Robot Interaction
- Reasoning and Learning in Real-World Systems for Long-Term Autonomy

This year, the symposium will include two special events:

Growing Federal Support for AI Research

Henry Kautz, Division Director, CISE/IIS, National Science Foundation will present a short talk at the beginning of the Friday evening plenary session, entitled Growing Federal Support for AI Research. He will describe current efforts to expand and coordinate support by government and industry in AI research. At the National Science Foundation, support of AI is now an agency-wide priority. In addition to traditional support from CISE/IIS, support for AI is coming from a number of recent cross-division and interagency programs, such as the Future of Work at the Human-Technology Frontier.

NSF is growing funding for research on the social impacts of AI.All major federal agencies that fund AI R&D, including NSF, Department of Defense, Health and Human Services, NASA, and many others, have begun to coordinate their efforts through a new AI working group at NITRD. The Community Computing Consortium is leading a study with support from NSF to create a roadmap that is expected to influence government support for AI research over the next decade.

AI for K-12

AAAI and the Computer Science Teachers Association (CSTA) recently formed a joint working group to develop national guidelines for teaching K-12 students about artificial intelligence. Inspired by CSTA's national standards for K-12 computing education, the "AI for K-12" guidelines will define what students in each grade band should know about artificial intelligence, machine learning, and robotics.

The working group will also create an online resource directory where teachers can find AI-related videos, demo software, and activity descriptions they can incorporate into their lesson plans. The first AI4K12 Symposium will be colocated with the Fall Symposium Series on Saturday, October 20. For more information, please see the AI4k12 website at github.com/ touretzkyds/ai4k12/wiki.

Reception and Plenary Session

The Fall Symposium Series will feature an informal reception on Thursday, October 18. A general plenary session, in which the highlights of each symposium will be presented, will be held on Friday, October 19. Participation is open to active participants as well as other interested individuals on a first-come, first-served basis. Each participant will be expected to attend a single symposium.

Registration

The final deadline for registration is September 21, 2018. For registration information, please contact AAAI at fss18@aaai.org or visit AAAI's web site at www.aaai.org/Symposia/Fall/fss18.php.

A hotel room block has been reserved at the Westin. The cut-off date for reservations is September 26, 2018 at 5:00 PM ET. Please call +1-888-627-7076 (reference AAAI) for reservations, or reserve a room online via the URL previously noted.

Join Us for AIIDE-18 In Alberta, Canada!

The Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-18) will be held at the University of Alberta, in Edmonton, Alberta, from November 13-17, 2018.

AIIDE is the definitive point of interaction between entertainment software developers interested in AI and academic and industrial AI researchers. Sponsored by AAAI, the conference is targeted at both the research and commercial communities, promoting AI research and practice in the context of interactive digital entertainment systems with an emphasis on commercial computer and video games. This year's conference features a special topic of "Situated Entertainment," and will include speakers, panels, and paper sessions that focus on a broad range of complementary areas of interactive digital entertainment.

AIIDE-18 will include three invited speakers: Ana Paiva (University of Lisbon and INESC-ID); Darius Kazemi (Feel Train); and Theresa Duringer (Temples Gates Games). The program will also include technical paper presentations, a poster and demonstration session, a doctoral consortium, and the annual Starcraft AI Competition. Plans are also underway for an industry event, connecting conference participants with the local AI and Games industry in Alberta. The Playable Experiences track will continue this year with five systems with articulable innovation in the use of AI directly affecting the user's experience, including "Escape Plan" and "At the Bar": Dynamic Characters Driven by Spirit AI Character Engine; Vox Populi: The Ustradian Games; Project Hastur: An Evolutionary Tower Defense Game; Camelot: An Interactive Narrative Sandbox Environment; Bottery; and PASS: A Game for Social Skills Training.

Finally, the main conference program will be preceded by two workshop days on November 13 and 14. The workshops are as follows:

Artificial Intelligence for Strategy Games
(W1) *November 13*
skatgame.net/mburo/aiide18ws

Experimental AI in Games Workshop
(W2) *November 13-14*
www.exag.org

Joint Intelligent Narrative Technologies / Intelligent Cinematography and Editing Workshop
(W3) *November 13-14*
sites.google.com/ncsu.edu/intwiced18

Learning to Play: Multi-Agent Reinforcement Learning in MalmÖ Competition
(W4) *November 14*
marlo-ai.github.io

The full conference program and registration information is available at aiide.org. The late registration deadline is October 12. Onsite rates will be in effect after that date. Preregistration is strongly encouraged. The online registration form is available at www.regonline. zcom/aiide18, and will be open through the conference period. Onsite registration will be held in the foyer of the Centennial Centre for Interdisciplinary Science (CCIS) Building, University of Alberta, North Campus. For more information about registration or hotels in the area, please consult www.aiide.org, or write to aiide18@aaai.org.

ICWSM To Be Held in Germany in 2019

Please join us for the Thirteenth International AAAI Conference on Weblogs and Social Media, to be held at the Conference Center Kolpinghaus Munich Central GmbH, in Munich, Germany, June 11–14, 2019. This interdisciplinary conference is a forum for researchers in computer science and social science to come together to share knowledge, discuss ideas, exchange information, and learn about cutting-edge research in diverse fields with the common theme of online social media. This overall theme includes research in new perspectives in social theories, as well as computational algorithms for analyzing social media. ICWSM is a singularly fitting venue for research that blends social science and computational approaches to answer important and challenging questions about human social behavior through social media while advancing computational tools for vast and unstructured data. Full conference details will be posted at www.icwsm.org/2019 as they become available.

AAAI Fellows Nominations Solicited

The 2019 Fellows Selection Committee is currently accepting nominations for AAAI Fellow. The AAAI Fellows program is designed to recognize people who have made significant, sustained contributions to the field of artificial intelligence over at least a ten-year period. All regular members in good standing are encouraged to consider nominating a candidate. At least two references must accompany nominations. The nominator or one of the references must be a AAAI Fellow who is a current member of AAAI. For further information about the Fellows Program, please contact AAAI at fellows19@aaai.org. Nomination materials are available on the AAAI web site at www.aaai.org/Awards/fellows.php. The deadline for nominations is September 28, 2018.

2019 AAAI Special Award Nominations

AAAI is pleased to announce the continuation of several special awards in 2019, and is currently seeking nominations for the 2019 Feigenbaum Prize, the 2019 AAAI Classic Paper Award, the AAAI Distinguished Service Award, and the AAAI/EAAI Outstanding Educator Award. The AAAI Feigenbaum Prize is awarded biennially to recognize and encourage outstanding Artificial Intelligence research advances that are made by using experimental methods of computer science. The "laboratories" for the experimental work are real-world domains, and the power of the research results are demonstrated in those domains. The 2019 AAAI Classic Paper Award will be given to the author of the most influential paper(s) from the Eighteenth National Conference on Artificial Intelligence, held in 2002 in Edmonton, Alberta, Canada. The 2019 AAAI Distinguished Service Award will recognize one individual for extraordinary service to the AI community. The AAAI/EAAI Outstand-

ing Educator Award honors a person (or group of people) who has made major contributions to AI education that provide long-lasting benefits to the AI community. Awards will be presented at AAAI-19 in Honolulu, Hawaii, USA. Complete nomination information, including nomination forms, is available at aaai.org/Awards/awards.php. The deadline for nominations is September 28, 2018. For additional inquiries, please contact Carol Hamilton at awards19@aaai.org.

AAAI Senior Member Grade of Membership

AAAI is now taking applications from regular members for the AAAI Senior Member grade of membership. This status is designed to recognize members who have achieved significant accomplishments within the field of Artificial Intelligence. To be eligible for nomination for Senior Member, candidates must be consecutive members of AAAI for at least five years and have been active in the professional arena for at least ten years. Applications should include information that details the candidate's scholarship, leadership, and/or professional service. At least two references, one of which must be written by a AAAI Fellow or a current AAAI Senior Member must accompany the senior member application. References should be submitted by colleagues who know the candidate, and are familiar with their work and accomplishments. Each year a maximum of 25 members will be elected to the Senior status. All applications and references must conform to the requirements listed on the form, and must be received by September 28, 2018. For complete details and the application form, please see www.aaai.org/Awards/senior.php, or contact Carol Hamilton at seniormember19@aaai.org.

AAAI Elects New President-Elect and Executive Councilors

AAAI is pleased to announce and congratulate the new slate of officers and councilors for the AAAI Executive Council.

President
Yolanda Gil (USC Information Sciences Institute, USA)

Past President
Subbarao Kambhampati (Arizona State University, USA)

President-Elect
Bart Selman (Cornell University, USA)

Secretary-Treasurer
David E. Smith

Incoming Councilors (through 2021)

Cristina Conati (University of British Columbia, Canada)

Eric Eaton (University of Pennsylvania, USA)

Ayanna Howard (Georgia Institute of Technology, USA)

Ariel Procaccia (Carnegie Mellon University, USA)

In addition to these individuals, eight councilors elected in 2016 and 2017 will continue their terms of service during the coming year. For a complete list, please refer to www.aaai.org/Organization/officers.php.

AAAI also thanks the four retiring Councilors, for their dedicated service and generous donations of time: Charles Isbell (Georgia Institute of Technology, USA), Diane Litman (University of Pittsburgh, USA), Jennifer Neville (Purdue University, USA), and Kiri L. Wagstaff (Jet Propulsion Laboratory, USA).

AAAI Executive Council Meeting Minutes

The AAAI Executive Council Meeting was held on February 3, 2018 in New Orleans, Louisiana, USA

Attending: Subbarao Kambhampati, Yolanda Gil, Tom Dietterich, Ted Senator, David Smith, Blai Bonet, Gene Freuder, Ashok Goel, Charles Isbell, David Leake, Diane Litman, Steve Smith, Matthijs Spaan, Peter Stone, Kiri Wagstaff, Shlomo Zilberstein, Sheila McIlraith, Kilian Weinberger, Carol Hamilton, Mike Hamilton.

Not attending: Jennifer Neville, Mausam, Michela Milano, Qiang Yang, Claire Monteleoni, Cynthia Rudin

AAAI president Rao Kambhampati commenced the meeting at 9:00 AM

First Call for Nominations for 2019 Executive Council Election

The 2019 Nominating Committee is seeking nominations from the AAAI membership for the position of Executive Councilor. In 2019, AAAI members will elect four new councilors to serve three-year terms on the AAAI Executive Council. All elected councilors are required to attend all council meetings each year (usually 1-2 in person and 2-3 via telecon), and actively participate in AAAI activities. Nominees must be current members of AAAI. The Nominating Committee encourages all regular AAAI members in good standing to place an individual's name before them for consideration. (Student and institutional members are not eligible to submit candidates' names.) The Nominating Committee, in turn, will nominate eight candidates for councilor in early spring. In addition to members' recommendations, the committee will actively recruit individuals in order to provide a balanced slate of candidates. AAAI regular members will vote in late spring, and the new members of the Executive Council will be installed in the summer of 2019.

To submit a candidate's name for consideration, please send the following information to Carol Hamilton, Executive Director, AAAI, 2275 East Bayshore Road, Suite 160, Palo Alto, CA 94303; by fax to 650/321-4457; or by email to hamilton@aaai.org:

Name

Affiliation

City, State or Province, Country

Email address

URL

Year of membership in AAAI

Approximate number of AAAI publications

At least two sentences describing the candidate and why he or she would be a good candidate

Please include any additional information or recommendations that would be helpful to the Nominating Committee. Nominators should contact candidates prior to submitting their names to verify that they are willing to serve, should they be elected. The deadline for nominations is March 1, 2019.

AAAI 2019 Spring Symposium Series

AAAI presents the 2019 Spring Symposium Series, to be held Monday – Wednesday, March 25–27, 2019, at Stanford University. The topics of the nine symposia will be:

Artificial Intelligence (AI), Autonomous Machines and Human Awareness: User Interventions, Intuition and Mutually Constructed Context
Ranjeev Mittu and Don Sofge (Naval Research Laboratory), W.F. Lawless (Paine College)

Beyond Curve Fitting — Causation, Counterfactuals and Imagination-Based AI
Elias Bareinboim (Purdue University), Sridhar Mahadevan (Adobe & UMass), Prasad Tadepalli (Oregon State), Csaba Szepesvari (DeepMind & University of Alberta), Judea Pearl (University of California, Los Angeles)

Combining Machine Learning with Knowledge Engineering
Andreas Martin and Knut Hinkelmann (FHNW University of Applied Sciences and Arts Northwestern Switzerland)

Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness
Takashi Kido (Preferred Networks, Inc., Japan) and Keiki Takadama (The University of Electro-Communications, Japan)

Privacy-Enhancing Artificial Intelligence and Language Technologies
Shomir Wilson (Pennsylvania State University), Sepideh Ghanavati (University of Maine), Kambiz Ghazinour (Kent State University), Norman Sadeh (Carnegie Mellon University)

Story-Enabled Intelligence
Dylan Holmes (MIT), Leilani H. Gilpin (MIT), Jamie C. Macbeth (Smith College)

Towards Artificial Intelligence for Collaborative, Open Scientific Discovery
Evan Patterson (Stanford University), Ioana Baldini (IBM Research AI), Peter Bull (DrivenData)

Conscious AI Systems
Antonio Chella (University of Palermo and ICAR-CNR, Italy), David Gamez (Middlesex University, UK), Patrick Lincoln (SRI International), Riccardo Manzotti (IULM University, Italy) Jonathan Pfautz (DARPA)

Verification of Neural Networks
Clark Barrett (Stanford University) and Alessio Lomuscio (Imperial College London, UK)

For additional information, and links to the supplementary websites for each symposium, please see www.aaai.org/Symposia/Spring/sss19.php.

Submissions for the symposia are due to organizers on November 2, 2018. Notification of acceptance will be given by December 3, 2018. Registration information will be available by December 15, 2018.

Please contact AAAI at sss19@aaai.org with any questions.

and welcomed everyone. He noted there were two very important issues on the day's agenda, including the Publications Transition Committee Report and the Code of Conduct. Participants introduced themselves, as this was the first in-person meeting for the new councilors.

Minutes

The Council approved the minutes of the November 30, 2017 meeting.

Updates from the President

Kambhampati noted that the current AAAI/ACM Conference on AI, Ethics, and Society is one of the first examples of this kind of cooperation between the two societies. Francesca Rossi worked tirelessly to make the conference a reality, and it has been very successful, with over 250 participants. Several companies have provided sponsorship and the conference overall has garnered a lot of interest from industry and the press. One last-minute addition was a graphic recording of the conference.

The AAAI-18 conference is live-streaming the invited talks this year, which will give access to the talks in a

AAAI Member News

Wolfgang Wahlster Honored

AAAI congratulates AAAI Fellow Wolfgang Wahlster for the unique distinction of being made an honorary citizen by his hometown of Saarbruecken, Germany. Wahlster is being honored for establishing Saarbruecken as an internationally recognized place for research excellence in computer science and AI during the last 35 years. Only three others have been appointed as honorary citizens of Saarbrueckenin during the last 15 years, including Willi Graf, Tavi Avni, and Max Braun. Honorary citizenship is the highest honor a German community can award.

Wolfgang Wahlster is the Director and CEO of the German Research Center for Artificial Intelligence (DFKI) and a professor of computer science at Saarland University. He has published more than 200 technical papers and 12 books on user modeling, spoken dialog systems, mobile and multimodal user interfaces, the semantic web, as well as the internet of things and services. He is a Fellow of AAAI, ECCAI, and GI. In 2001, the president of Germany presented the German Future Prize to Wahlster for his work on intelligent user interfaces, the highest personal scientific award in Germany. He was elected Foreign Member of the Royal Swedish Nobel Prize Academy of Sciences in Stockholm and Full Member of the German National Academy of Sciences Leopoldina that was founded in 1652.

He has been awarded the Federal Cross of Merit, First Class of Germany. He holds honorary doctorates from the universities of Darmstadt, Linkoeping and Maastricht. He serves on the Executive Boards of the International Computer Science Institute at UC Berkeley and EIT Digital. He is the editor of Springer's Lecture Notes in Artificial Intelligence series and serves on the editorial board of various top international computer-science journals. In 2013, Wolfgang Wahlster received the IJCAI Donald E. Walker award for his substantial contributions, as well as his extensive service to the field of AI throughout his career.

more timely manner than in the past. Kambhampati will be tweeting the livestreaming link.

AAAI has a seat on the Partnership for AI (PAI), which Kambhampati currently fills. The group established seven thematic pillars last year and is now forming working groups around those pillars. PAI is now asking partnership members if they want to serve on those working groups, and Kambhampati will be following up with specific people on that request. He will nominate interested individuals and the PAI will select the final participants. PAI will be establishing its headquarters in San Francisco, California USA.

AAAI has hired a Media Relations firm to spearhead media interest. The firm has vetted interest from the press, and has organized a press conference during the conference.

The Community Meeting will be held on Tuesday, February 6 at 5:00 PM. This meeting is open to all conference registrants and AAAI members. Ted Senator noted that this meeting also satisfies the requirement for the annual AAAI business meeting.

Yolanda Gil noted that AAAI signed a memorandum of understanding with Iridescent Learning, who do outreach for K-12 learners. A tutorial was held to teach AI researchers how to develop devices that can be used in outreach, and a follow-up event was held in a local school. One hundred and thirty families signed up to participate in a hands-on building project requiring AI engineering. The AAAI event kicked off a large campaign for Iridescent, during which they will hold similar events in 100 countries, featuring the Curiosity Machine. Gil hopes this effort will grow at future AAAI conferences.

Standing Committee Reports

Awards and Fellows Committees

Tom Dietterich, chair of these two committees, reported that the awards process overall had a larger set of nominations this year, which better represented the diversity of the field along several different lines, including international affiliation, research area, gender, as well as others. He noted that we do not get enough nominees for the Distinguished Service Award, and he encouraged everyone to consider who appropriate candidates might be. In addition, there is still considerable room for growth in the Senior Members program. Dietterich also mentioned that there is an outstanding proposal for a Dissertation Award, which would be given jointly by AAAI and ACM/SIGAI. The proposal is currently being reviewed by ACM. It is hoped that the first round of nominations will be accepted in Fall 2018. The winner will be given a speaking slot in a parallel session at the AAAI conference. One requirement currently being reviewed is whether one must be a pro-

fessor to serve on the selection committee, or is it possible for someone from industry to serve on the committee. Dietterich noted that his main concern is that people on the committee would have been recently involved in reading dissertations. Both David Smith and Kiri Wagstaff noted that they serve on similar awards committees and are actively involved in these processes, so perhaps membership does not have to be limited to professors. Charles Isbell asked for clarification about whether both supervising and serving on committees should be required, or just serving on committees. Dietterich agreed that recent service on a dissertation committee should be sufficient.

Conference

Sheila McIlraith and Kilian Weinberger, AAAI-18 program cochairs, presented a report on the conference submission and review process. McIlraith noted that the technical submissions and final content of the conference is the largest in 32 years. Weinberger reviewed statistics for attendance, submissions, and acceptances, noting that all were significantly higher than in 2017. Although the cochairs had recruited a large set of reviewers in order to reduce reviewer load, they ended up keeping the same reviewer load but being well prepared for the tremendous spike in submissions. The acceptance rate remained at the same level as 2017, netting a much larger number of technical papers for the conference. In examining the origin of papers by country, they discovered that there was a 60 percent increase from China, with the next largest increase coming from the US. The US also had the largest rate of accepted papers. China's acceptance rate remained the same as in 2017, despite the larger number of submissions. The largest research area for accepted papers was machine learning, followed by vision and NLP and machine learning. Vision experienced a 257 percent increase in submissions, along with a 285 percent increase in acceptances. This caused the biggest strain on the program committee. McIlraith and Weinberger noted that many of the papers in other areas use machine learning techniques. The cochairs made an effort to attract more machine learning work by placing the submission deadlines after the Neural Information Processing Systems conference (NIPS) notification date, and explicitly invited NIPS resubmissions in the call for papers. McIlraith noted that it was difficult to know how many papers were submitted as a result of the resubmission option in the call for papers as authors were not required to state if this was the case. McIlraith suggested that the Conference Committee should consider the submission and attendance statistics carefully for future planning in order to maintain the high quality of the AAAI conference, and to continue to serve the traditional base of researchers who regularly attend AAAI. While many machine-learning students attended the conference to present their work, it was more difficult to find session chairs for the machine learning sessions. This may suggest that machine learning researchers do not consider AAAI their home conference. McIlraith thanked all the members of the program committee who stepped up to accept greater loads due to the dramatic increase in submissions.

Ted Senator inquired about the difference between applications or machine learning application papers and IAAI papers. McIlraith and Weinberger noted that the keywords have been used for several years, and that authors are allowed to self-identify the appropriate category for their paper. They did not classify papers. It might be time to review the standard keywords to be sure they are still appropriate for classifying papers as accurately as possible.

Weinberger noted that acceptance rates for the largest areas of submission (machine learning [ML], vision [VIS] and natural language [NL]) were fairly consistent. Most areas hovered at the overall 25 percent acceptance rate, with a few areas below that and several 5–10 percent higher than that. Gil noted that the elimination of the two special tracks on computational sustainability and AI and cognitive systems did not seem to affect the submission and acceptance on papers in those areas, so it appears that the two special tracks of the past several years were affective.

McIlraith and Weinberger reminded the group that they enforced a new rule, strictly limiting the number of submissions to 10 per author. While the vast majority of authors only had one submission, over 150 still had more than 5 each. IJCAI is now enforcing a similar rule. In addition, AAAI does not allow author names to be added after submission. There was a short discussion about how these requirements were tracked at the final submission stage and whether there might be any way for original submission information to be imported into the digital library to reduce the need for manual entry and checking. Mike Hamilton noted that it is important for the author to submit the final data because of publication requirements, and assured the group that all papers are thoroughly checked for submission and formatting requirements. Finally, the cochairs noted that the number of allowed pages was increased to seven, plus one for references only. Weinberger remarked that allowing an indefinite number of pages for references is becoming more commonplace, and that the conference committee might consider this option.

While AAAI-18 did not have special tracks as in the past, it did have three areas that were closely overseen by special tracks chairs — human AI collaboration, cognitive systems, and computational sustainability. Human AI collaboration was identified as a special emerging area, and was given a dedicated track in the conference, including four 30-minute invited presentations. Rather than separate program committees and separate reviewing tracks, these areas were folded into the main reviewing system to allow for more consistent reviewing and also to acknowledge that they may cross over several areas in their content. The track chairs suggested program committee members to cover these areas and served as second area chairs on all submitted papers in their respective areas. Authors were given the opportunity to self-identify to be eligible to participate in the special review process for these special tracks. As with all papers, the area chairs, special track chairs, and senior program committee (SPC) members were given the opportunity to

assign one known and trusted reviewer to each of their assigned papers.

Shlomo Zilberstein complimented McIlraith and Weinberger on the process that they used and asked them if they thought it worked well and should continue in the future. They reported that it did work well, but would like to have had larger participation in the reviewer assignment process by eligible people. For 2018 there was only a small 2-day window during which this could be done, and they would like to see the practice built into the system earlier. A short discussion ensued about the tradeoffs in not having special tracks, especially for the cognitive systems area. Zilberstein expressed his hope that AAAI settles on an approach to all of these issues that can be consistent for several years. Kambhampati noted that it will be important to loop the 2019 cochairs into the discussion.

Weinberger and McIlraith explained the reviewing process. They decided to use the Conference Management Toolkit (CMT) system, which could be integrated with the Toronto Paper Matching System. Neither ConfMaster nor EasyChair were able to provide this option. They also added another level to the program committee structure with 63 area, special, and emerging topics chairs. Unfortunately, CMT was not able to easily adapt to this extra level, but the cochairs created a workaround to differentiate area chairs from the senior program committee. Finally, they created a workflow chair, who worked with Weinberger on developing scripts to process the TPMS data and bids.

Weinberger and McIlraith pointed out two things that they would like 2019 cochairs to consider. First, the International Conference on Machine Learning (ICML) deadline at the end of the AAAI-18 week may have affected attendance by machine learning researchers at AAAI-18. Second, AAAI-18 still has a very small sponsorship rate compared to other conferences, such as IJCAI, NIPS, or ICML. Third, future chairs should further address the issue of inferior reviews up front to clarify the expectation for an acceptable review.

Another area that they did not have time to pursue was a mechanism to help improve the quality of oral presentations. Some ideas that were considered for this effort included creating an AAAI YouTube channel to archive talks or invited talks; requiring authors to do a trial run of their presentation for an area chair or SPC member (NIPS requires this); or requiring authors to submit their slides ahead of time for review. Peter Stone noted that the NIPS program also fostered student-faculty connections. McIlraith also mentioned the possibility of creating an online form for feedback to authors about their talks during the conference (opt-in only). Gil suggested having a speaker helpdesk where students could go for help with their slides and talk. She also suggested a webinar or conference session for reviewers to learn how to review better.

Kambhampati thanked Weinberger and McIlraith for the huge effort they put in to deal with the enormous upsurge in submissions and reviewing, as well as to create a memorable invited speaker program and the new emerging topic program. He noted that they worked until the last possible moment, arranging for a special debate to be held that evening. The Council also thanked them for all their work.

Zilberstein thanked the program chairs and also thanked the Executive Council for their support during his tenure as Conference Committee chair. He will be stepping down, and will be replaced by Peter Stone. Kambhampati thanked Zilberstein for his years of service, and also thanked Stone for assuming this responsibility. Finally, Weinberger and McIlraith thanked Carol Hamilton and the rest of the AAAI staff for all their support during the planning and execution of the conference.

Finance

Ted Senator reviewed the financial status of the organization, giving a brief history dating back to 2002. In particular, he noted the percentage that can be withdrawn each year from the investments. The standard practice for nonprofits and universities is to remain in the 3–5 percent range. In recent years, we have withdrawn less than that. Senator also noted that our financial policy states that we try to keep no more than 65 percent in equities or stocks, and 35 percent in bonds. At the end of December 2017, we rebalanced the investment portfolio to better align our accounts with the policy.

Recently, Senator did an analysis to see if our international investments align with the international make-up of our members. Interestingly, the investment distributions by region, compared to memberships in those regions, were close. Going forward this alignment may be important should our expenses not always be based in US dollars.

Senator thanked people for reviewing the tax returns last fall, and noted that this tax return requirement ensures transparency. He also reported that the budget was approved prior to the commencement of the fiscal year 2018. Gil asked about the pending bylaws issue. Hamilton explained that she will be following up on this with the AAAI lawyer to identify all sections of the bylaws that are currently out of sync with current practice. Senator thanked the Council for his 15 years as Secretary-Treasurer, and Kambhampati thanked Senator for his extraordinary service. Kambhampati nominated David Smith as Senator's successor, and asked for approval of this appointment. The motion carried unanimously.

Publications

David Leake reviewed the Publications Transition Plan, put together by members of the Publications Committee and a few other people with experience in related areas, such as social media and open source publications. The committee identified the types of publishing activities that AAAI should support moving forward, and then what would be the most effective way to support those activities. Currently, AAAI publishes its own conference proceedings as well as those for several other conferences, technical reports for workshop and symposium papers, and the *AI Magazine*. In addition, there are plans for an expanded online version of *AI Magazine*. In examining publications, the committee evaluated the benefits of each publication to AAAI, AAAI members, and the AI community as a whole, balanced against financial considerations, as well as trends in

content generation and access, and available resources. The committee's work resulted in three proposals: (1) the future scope of AAAI publications, (2) volunteer and staff structure for AAAI publications, and (3) a set of immediate action items. The committee agreed that AAAI should continue to produce the *AI Magazine* in archival hard and electronic (PDF) copy as well as in the proposed expanded online version now in development, emphasizing that *AI Magazine* is the flagship publication for AAAI.

The committee also recommended the continuation of proceedings publication in electronic format, but to discontinue the hard-copy version. The committee recommended discontinuing the production of technical reports for workshops and symposia, leaving open the option for workshop and symposium organizers to publish the work, which is usually very preliminary, at other sites. IJCAI has followed this process for its workshops.

For non-AAAI conferences, proposals will be reviewed on an individual basis by the Publications Committee, taking into account benefits, staff time available, and financial considerations.

The committee also outlined the structure of the publications operation, continuing with the executive director as the publisher, who will hire and supervise staff, and manage expenses according to AAAI policies. They recommended hiring a full-time publications manager, who would eventually replace the current Publications Director, Mike Hamilton. This manager would serve on the Publications Committee, working directly with authors and volunteers. The position will require an experienced and skilled candidate due to the complex set of responsibilities involved in the publications process. The *AI Magazine*, led by the editor in chief, will fall under the purview of the Publications Committee, and will be guided by an advisory board, including, for instance, the previous editor of the magazine.

The editor in chief will nominate a series of associate editors, who will be assisted by the *AI Magazine* Editorial Board. Each associate editor will assume responsibility for a specific issue of the magazine, and will be overseen by the editor in chief. All appointments will be approved by the Publications Committee chair, in consultation with the Publications Committee, as needed. The editor in chief would also be approved by the president. All positions are for three-year terms to encourage participation. Renewal of terms could be considered.

The action items needing immediate approval were as follows: (1) The executive director should hire a publications manager as soon as possible, with input from the president, Publications Committee chair, and editor in chief; (2) The editor in chief should recruit the members of the Advisory Board and associate editors; (3) The editor in chief should develop a phased proposal for the web version of the magazine, including a list of requirements, which will be evaluated by the executive director and Executive Council, should there be significant financial impact for them. The executive director will keep the Executive Committee or Executive Council apprised of potential needs for additional resources and staff in the publications area; and (4) The committee proposed a regular annual evaluation of the *AI Magazine* in all formats with the aim of finding ways to continually improve the magazine. The review would be conducted by the editor in chief, in conjunction with the Publications Committee chair. The results would be referred to the Publications Committee with a plan for implementation of recommendations.

Diane Litman inquired whether the Publications Committee has ever considered the publication of a hybrid technical paper-journal paper model, often seen in the database community. Leake deferred the inquiry to the Conference Committee. Zilberstein noted that the journal paper track has been used with some success by other conferences. While AAAI cannot publish previously published work, AAAI may encourage authors to present their journal work at the conference in the future. The publication of an extended abstract might be possible.

Zilberstein asked about the possibility of making AAAI papers more routinely included in indexing databases. Hamilton responded that the assignment of DOIs to all proceedings papers is scheduled for completion in the next 18 months. This will make indexing of these papers quite routine.

Gil asked if the Publications Committee considered moving away from a central publications process to a more online, distributed function. Leake noted that the committee did consider this model, but decided that for the flagship publications it was important to maintain very high quality that could only be achieved by having an in-house function.

Isbell asked if the symposium and workshop organizers will be given explicit instructions about how they can use the AAAI name and logo when seeking publication at another site. Leake agreed that these guidelines should be distributed to organizers before any action is taken. He will review the current guidelines and revise them as needed.

Gene Freuder asked if the migration to an electronic-only publication of the proceedings might provide an opportunity to simplify the camera-ready submission process. Leake will follow up on this with the Publications Committee in conjunction with the Conference Committee.

Blai Bonet pointed out that other conferences will require stability moving forward, and Leake agreed, noting that they will also need sufficient lead time to make other arrangements if necessary. Goel responded to David Smith about whether the online version of the *AI Magazine* will be quarterly or more dynamic. He explained that the print and PDF version of the magazine will continue to be quarterly, but that the online version will have several dynamic features. It will feature more than just articles, including things like columns, podcasts, blogs, and a member-only area for comments about specific articles. Freuder inquired about whether the manager would be an expert in online systems; Goel responded that his upcoming proposal requests an additional hire for someone who will be an expert in such things as website maintenance, user experience, and topical layout, for instance.

Dietterich noted that we need to consider whether we want to hire an in-house webmaster. Hamilton noted

that we currently have outside assistance with the development and maintenance of the website, and this resource may be sufficient for the early stages of the online *AI Magazine* delivery. Gil asked about a timeline, and Leake noted that they hope to hire a manager as soon as possible. Freuder asked about whether the elimination of some of the publications will result in less of a time commitment for the publications manager. Leake and staff responded that the concentration on the main proceedings (and possibly other proceedings) and the *AI Magazine* will make the job manageable as opposed to what it is now. In addition, the publications manager will rely on trained freelancers or other services to complete many of the routine tasks.

Leake thanked Mike Hamilton for all of his work on publications, and Kambhampati thanked the Publications Transition Committee for its work on putting together such a detailed proposal.

Goel gave a status report on the *AI Magazine*, including recent and planned future issues. He noted that getting guest editors for each issue is quite challenging. Goel has moved away from special issues on specific topic areas, instead featuring articles from award winners, invited speakers, tutorial speakers, or AAAI Fellows. He is hoping that articles derived from invited talks at other conferences will help build bridges to those communities. A side effect might be to increase interest and membership in AAAI. He also gave a brief overview of the online magazine plans, which he hopes will be very dynamic and multi-modal. The schedule could be as frequent as daily, and it will be very accessible and searchable. The magazine will have both member and nonmember areas. The member area will include an area where they can publish things accessible by members only. He reiterated that the traditional quarterly format of the magazine will feature mainly expository article in the future. He envisions the addition of many columns — perhaps about 20 — often linking AI to a specific topic area, such as education or computer vision, with the goal of building further bridges. Special forums on topical issues, such as AI and Ethics, will also be added. He may work with other publications on exchanging certain articles. The online content will be curated.

Goel presented a timeline for rollout of the online magazine, noting that the earliest version would not be available for another 6–8 months. Goel has recruited five associate editors, who were former members of the *AI Magazine* Editorial Board. Goel's vision for the Editorial Board is to increase the size and ongoing participation of the members by making each person responsible for a specific column in the magazine.

Several current members of the Editorial Board opted to move to the newly formed Advisory Board instead. Goel requested that an additional full-time staff member be hired for a preliminary period of three years whose primary responsibility will be to design, curate, and maintain the *AI Magazine* website. He also requested a half-time administrative assistant at AAAI for an initial period of one year to help with the management of and communication with the newly expanded *AI Magazine* structure, including the five new associate editors and 20 column editors.

Carol Hamilton noted that some of the tasks that need to be accomplished may be covered by the new publications manager, but it is too soon to tell at this point. Kambhampati agreed that we need to get a little further into the proposed publications transition before all of the needs can be assessed. He also shared a concern brought up earlier by Tom Dietterich about the impact on the membership dues of this large increase in overhead expenses.

Gil wondered about the possibility of having someone at Goel's university assist with the management of editorial correspondence. Goel was open to all possibilities, but felt strongly about the need for some support in this area in order not to lose interest from the associate editors that are already on board. Dietterich also asked Goel to consider recalibrating his expectations initially for the online content so that the need for support might not be quite as high, at least in the initial period.

Gene Freuder suggested that it might be possible to garner corporate sponsors for some of the infrastructure needed to support the magazine. No decision was made on the two requests for additional staff, deferring until after the completion of the publications transition discussed earlier.

Goel will provide a more detailed list of responsibilities for the proposed staff at the next Executive Council meeting, and will also draft a letter to his home institution regarding support from that side. The associate editors will concentrate on the hard copy version of the magazine in the near future. Kambhampati thanked Goel for the time and energy he has already spent on future planning for the magazine.

Ethics

At the last Executive Council meeting, the Ethics Committee presented a first draft of a proposed code of ethics and professional conduct, and asked for feedback. Since that time, they incorporated suggestions into the current draft now up for review by the Council. Overall, the Council had asked for a higher-level document, with less specific language. Kambhampati asked the Council to make final suggested changes to the draft so that the final wording could be approved at the meeting. This will enable the Council to move to the other pending tasks, such as formulating a more detailed harassment policy for AAAI conferences. The Council reviewed the current document line by line and made changes, as needed, discussing in depth issues raised as they progressed.

Peter Stone asked for a bit of background about the statement, and Kambhampati explained that it had been in the works for a couple of years. Francesca Rossi and Sonia Chernova worked on the initial draft, and researched similar statements by other organizations. The existence of this kind of policy enables an organization to institute other operational policies that can point to this as a standard. The purpose of this particular statement is to set a standard for behavior, and is not intended to justify punitive measures for noncompliance. Zilberstein noted that this could be seen as part of our vision statement, making known AAAI values. The current mis-

sion statement for the organization focuses mainly on the charge for the organization and not on professional behavior.

Kambhampati explained that this code of conduct might guide the Council in making a decision about a specific incident or issue being discussed at a future meeting. The Council agreed on a final draft, and asked to have it reviewed by legal counsel. Once reviewed again by the Council, the code of conduct will be posted online for member comments. Ted noted, however, that the Council can approve the code of conduct without member approval. Gil also noted that it is important to have this finalized so that the Council can draft a harassment policy.

Finally, Kambhampati asked for volunteers to serve on the Ethics Committee, as some members have rotated off.

Membership

Gene Freuder noted that the Membership Committee had two major items on their agenda last fall, including the review of the distinguished speaker applications and several pending chapter applications. The distinguished speaker application decisions were made, and a process for implementing the program is currently under development. The committee is awaiting additional information before final decisions are made on the chapter applications. The committee requested that the application form be updated to reflect all the requirements set forth in the chapter program description. Freuder explained that the current applications comprise a wide variety of requests from countries, such as the Dominican Republic and Romania, as well as US-based student chapters. The committee has also been looking at the possibility of adding additional membership categories, particularly in the industry area. The committee suggested that we have a AAAI pin made that could be distributed to new members (and current members at the conference or by request). Gil also suggested that the committee consider stickers instead, as these seem to be more popular.

Tom Dietterich suggested developing a line of t-shirts for people to choose or purchase. New members could select something as a bonus. The committee's idea is to provide a way to promote pride in being a AAAI member. Hamilton will propose a few items for the committee's review.

Government Relations

Steve Smith reported that AAAI joined in a statement with several other organizations, including ACM, IEEE, and CRA, protesting the tax-cut bill, which proposed to remove the waiver for graduate student tuition. The clause was removed from the bill.

Education

Wagstaff noted that the ongoing charge of the committee is to assist in answering mostly student queries that come through the website with general questions about AI or requests for interviews. There are also plans in the works to contribute to a special issue of *AI Magazine* on AI and ethics. Sven Koenig and Meinolf Sellmann will serve as the special editors. The committee is not only focused on the education of students, but also on professional education. Wagstaff suggested the committee might be able to contribute to an effort to improve technical reviews and talks. Gil suggested reaching out to the EAAI community for their input and involvement. Dietterich noted that there was strong interest in professional development opportunities from the membership questionnaire collected a few years ago. Kambhampati mentioned that he would like to have a series of webinars, featuring the Distinguished Speakers. This structure for the distinguished speakers program might prove to be far more effective than sending speakers to specific venues.

Dietterich suggested a follow-up questionnaire to assess interest in specific types of education, such as short tutorials, continuing education credits, and week-long courses. Wagstaff agreed that this would be useful. Gil added that, based on her experience, presenting webinars that are very accessible and get the word out about AI are in high demand and would be very useful to the general public. Freuder noted that efforts to increase engagement from different segments of the community seem to be dispersed among several committees, and that it may be necessary to coordinate these efforts and create a better mechanism to channel these efforts. Hamilton confirmed that the mechanism to survey members is in place.

New Business

Data Liberation Initiative

Tom Kalil, former head of the Office of Science and Technology Policy, is working with the Eric Schmidt Foundation, which is willing to pay to liberate data (mostly from federal agencies). This requires anonymizing the data so that it can be used for various research purposes. Kalil is seeking ideas about what kind of data should be liberated, and wondered if AAAI would like to run a workshop to explore this issue.

Kambhampati is investigating the possibility of a special meeting at the spring symposium or a fall symposium, with follow-up at AAAI-19, and asked for volunteers to help organize a special track. Gil also suggested that we coordinate with ChaLearn, an organization that might be able to better identify whether to begin to identify appropriate data.

AAAI Industry

Engagement Committee

Kambhampati noted that there is currently a large amount of interest in AI from industry and Gene Freuder has been looking at ways to leverage that interest to support outreach activities. Freuder suggested the Council needs to consider how various kinds of industry engagement can shape the organization. Freuder asked for additional members of the committee to help examine the issues. In addition to Council members, he is seeking members who have worked in or closely with industry. Gil nominated Karen Myers, who will chair the IAAI-19 conference. Freuder circulated a list of suggestions for engaging industry, and Kambhampati directed the Council's attention to the topic of running meetings that would be targeted at industry participation rather than technical meetings.

A related issue involves whether we want to participate in industry-based meetings by lending our name and brand and possibly providing an invited speaker. There have been a range of requests for our participation in these

meetings, and it is unclear what level of real participation some of these meetings want, if any. The Council felt strongly that there should be some mechanism for AAAI to showcase recent advances in AI if we are going to lend our name to the event. On the other hand, some meetings may not be ones that we want to lend our name to at all. Kambhampati noted that having clearer policies in place is necessary to deal with the current requests. Hamilton will provide a recap of these requests so that the committee can better assess the need and direction to go.

The Council also discussed whether AAAI should sponsor and organize its own industry event. Senator noted, and others agreed, that we do not have expertise in this kind of conference organization, and it would be difficult to compete in this arena. Instead, we should try to leverage the current interest in AI in order to showcase IAAI deployed applications or other noteworthy research trends on AI.

Gil suggested connecting more with research in industry, developing outreach efforts to involve those individuals and labs more in our activities. Concentrating our efforts on industry conferences that are focused on markets and forecasting would be quite removed from AAAI's research focus. Creating a channel for academics to connect with industry would be quite valuable for members and conference attendees. We could sponsor a forum where people could come to find out about a wide spectrum of research in AI — not just machine learning or data mining — but a whole range of technologies in AI.

Senator mentioned that KDD has developed an invited track that features senior researchers in industry who employ data science, and have use applied data science to achieve a research goal. The sessions tend to attract large audiences, but do not compete with the research conference.

Gil noted that the AI in Practice session at AAAI-17 followed this model. The AI in Practice event was not held in 2018 for a variety of reasons, including competing programs and the availability of local industry representatives to speak. Kambhampati noted that moving it to the afternoon of the last day might avoid the problems encountered with competing tutorials in 2017. IJCAI did this in 2016 and it was quite successful.

AI Hub

Tom Dietterich reported that they have not reached the fundraising goal yet, so they are investigating other vehicles for funding, such as funding agencies or the Partnership for AI.

AI Topics

Kambhampati reported that *AI Topics* is seeking a renewal of their contract for two years at $20,000 per year. He asked the Council to consider the various models currently being considered, including *AI Topics,* AI Hub and the expansion of *AI Magazine*. Tom Dietterich explained that AI Hub will feed off of *AI Topics,* as AI Hub is just an aggregator for these kinds of services. Hamilton noted that both AI Hub and the *AI Magazine* expansion are in the rudimentary stages, and encouraged the Council to continue *AI Topics* at this time. Currently, there is $10,000 in the 2018 budget. The request was referred to the Executive Committee for a final recommendation.

Social Media Volunteers

Kambhampati is seeking volunteers to help with social media alerts on Facebook and Twitter. Tom Dietterich volunteered to help, and Jennifer Neville is also helping with this effort. Hamilton noted that a staff member might be able to help with routine posts.

Conflict of Interest

Ted Senator presented an overview of the current concept for the AAAI Conflict of Interest statement, which is intended to elaborate the responsibilities of the Executive Council. It further defines the duty of loyalty, one of the three primary duties of Executive Council members. Often conflict of interest statements refer directly to financial dealings — a subject also covered in the AAAI Bylaws — but this iteration is intended to be a bit broader in scope.

Due to the many hats that councilors wear in their professional and volunteer commitments, this document is meant to address issues that might arise with regard to split loyalties. In general, it is the responsibility of the individual to disclose a conflict of interest, in conjunction with independent determination, meaning that the individual cannot be the sole determiner of a conflict of interest. Finally, individuals must recuse themselves from discussion of and final decisions regarding an issue identified as a conflict of interest.

While there are well-established principles for dealing with conflicts of interest, Senator asked the group to consider some of the ways that council members might encounter conflicts of interest.

Isbell asked if AAAI currently has a process in place for independent determination, and Senator confirmed that there is no process in place yet. He also noted that in addition to determining the process, the Council also has to consider how it will be applied in the future. Having a conflict of interest statement and process in place is an additional requirement of the IRS, and appears as a question on the tax form. AAAI councilors will be required to acknowledge this policy in writing.

David Smith reported that he and Ted Senator have been reviewing two possible versions, and will draft a final version. Kambhampati recommended that the legal version required by the IRS should be completed as soon as possible, and reviewed by legal counsel. A second accompanying document will be added later to encompass issues that are unique to AAAI councilors and program chairs. Isbell offered to assist with the drafting of these documents.

Ajournment

Kambhampati adjourned the meeting at 3:45 PM.

AAAI Conferences Calendar

This page includes forthcoming AAAI sponsored conferences, conferences presented by AAAI Affiliates, and conferences held in cooperation with AAAI. AI Magazine *also maintains a calendar listing that includes nonaffiliated conferences at www.aaai.org/Magazine/calendar.php.*

AAAI Sponsored Conferences

AAAI Fall Symposium Series. The AAAI 2018 Fall Symposium Series will be held 18–20 October 2018, in Arlington, Virginia adjacent to Washington, DC. USA.

URL: www.aaai.org/Symposia/Fall/fss18.php

Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. AIIDE-18 will be held 13–17 November 2018 in Edmonton, Alberta, Canada.

URL: aiide.org

The Thirty-Third AAAI Conference on Artificial Intelligence. AAAI-19 will be held 27 January – 1 February 2019 at the Hilton Hawaiian Village in Honolulu, Hawaii USA.

URL: www.aaai.org/aaai19

Thirty-First Innovative Applications of Artificial Intelligence Conference. The IAAI-19 Conference will be held 29–31 January 2019 at the Hilton Hawaiian Village in Honolulu, Hawaii USA.

URL: www.aaai.org/aaai19

AAAI Spring Symposium Series. The AAAI 2019 Spring Symposium Series will be held 25–27 March 2019, at Stanford University adjacent to Palo Alto, CA USA.

URL: www.aaai.org/Symposia/Spring/sss19.php

The Fourteenth International AAAI Conference on Web and Social Media. ICWSM-19 will be held 11–14 June in Munich, Germany.

URL: www.icwsm.org/2019

Seventh AAAI Conference on Human Computation and Crowdsourcing. The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019) will be held 28–30 October at the Skamania Lodge, in Stevenson Washington USA.

URL: www.humancomputation.com/2019

Conferences Held by AAAI Affiliates

16th International Conference on Principles of Knowledge Representation and Reasoning (KR 2018). KR 2018 will be held October 30 – November 2, 2018, in Tempe, Arizona, USA.

URL: www.kr.org

Second AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019). AIES 2019 will be held 27–28 January 2019 in Honolulu, Hawaii, USA.

URL: www.aies-conference.com

The Thirty-Second International FLAIRS Conference. FLAIRS-19 will be held 19–22 May in Sarasota, Florida USA.

URL: sites.google.com/view/flairs-32homepage/home

The Twenty-Ninth International Conference on Automated Planning and Scheduling. ICAPS-19 will be held in July 2019 in Berkeley, California, USA.

URL: icaps19.icaps-conference.org

Conferences Held in Cooperation with AAAI

Tenth International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. IC3K-2018 will be held 18–20 September in Seville, Spain.

URL: www.ic3k.org

Seventeenth International Joint Conference on Computational Intelligence. IJCCI-2018 will be held 18–20 September in Seville, Spain.

URL: www.ijcci.orgg

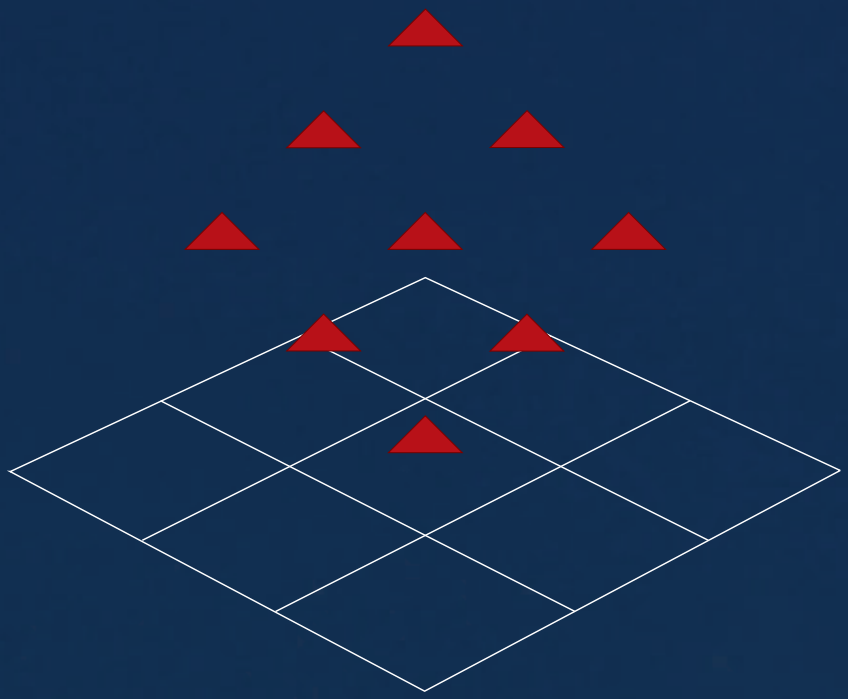
Tenth International Conference on Artificial Intelligence and Law. ICAIL-2019 will be held 17-21 June in Montréal, Québec, Canada.

URL: www.icail2019-cyberjustice.com

Thirty-Second International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems. IEA/AIE-2019 will be held 9–11 July 2019 in Graz, Austria.

URL: ieaaie2019.ist.tugraz.at

 ISSN 0738-4602



2018 AAAI Fall Symposium Series

October 18–20, 2018

Arlington, Virginia

www.aaai.org/fall

AAAI-19 / IAAI-19

Waikiki Beach

Hilton Hawaiian Village
Honolulu, Hawaii, USA

January 27 – February 1, 2019
www.aaai.org/aaai19

Aloha!

Pascal Van Hentenryck (Georgia Institute of Technology)
Zhi-Hua Zhou (Nanjing University)
Program Chairs

Photo courtesy 400tmax, iStock