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.

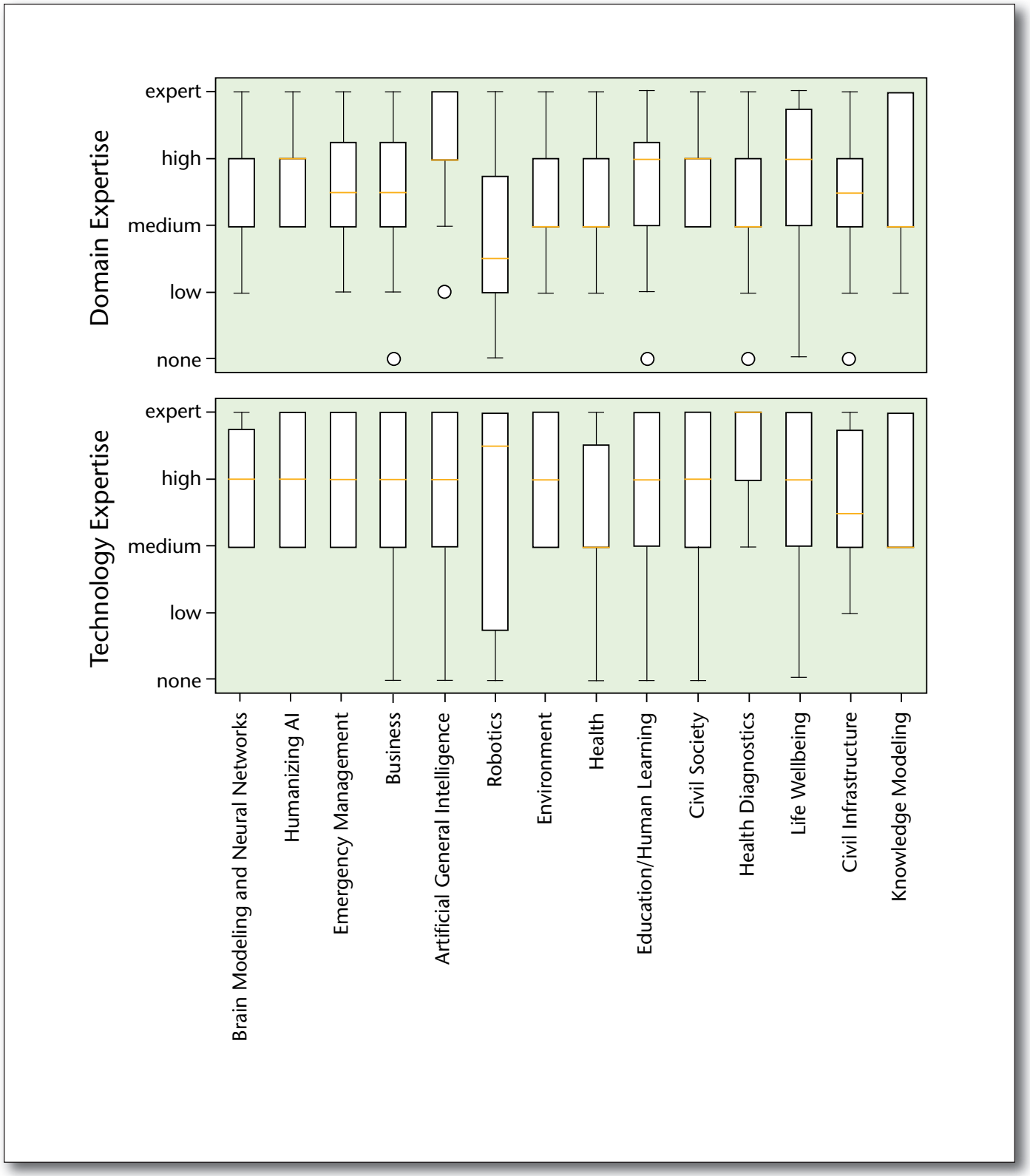


Figure 1. Box and Whisker Plots.

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.

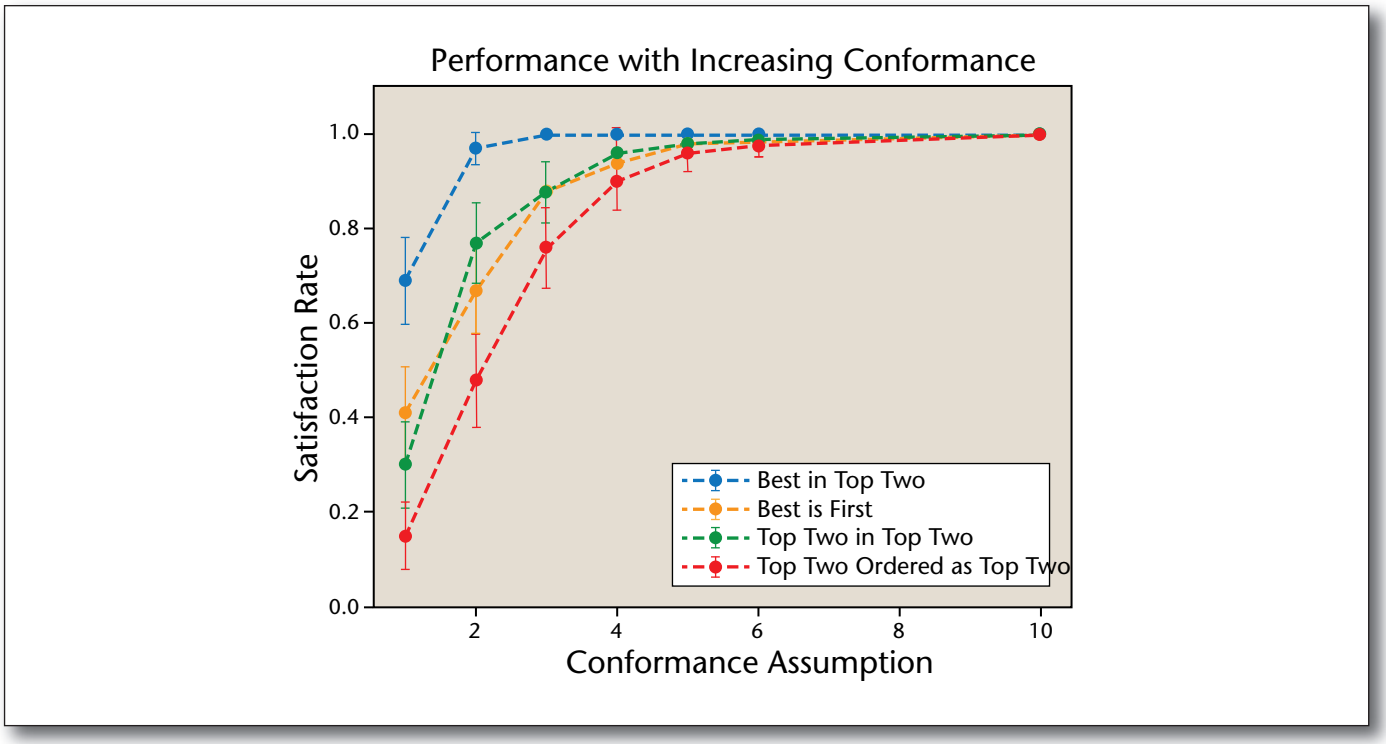


Figure 2. Performance Validation of the Weighted Metric via Monte Carlo Trials.

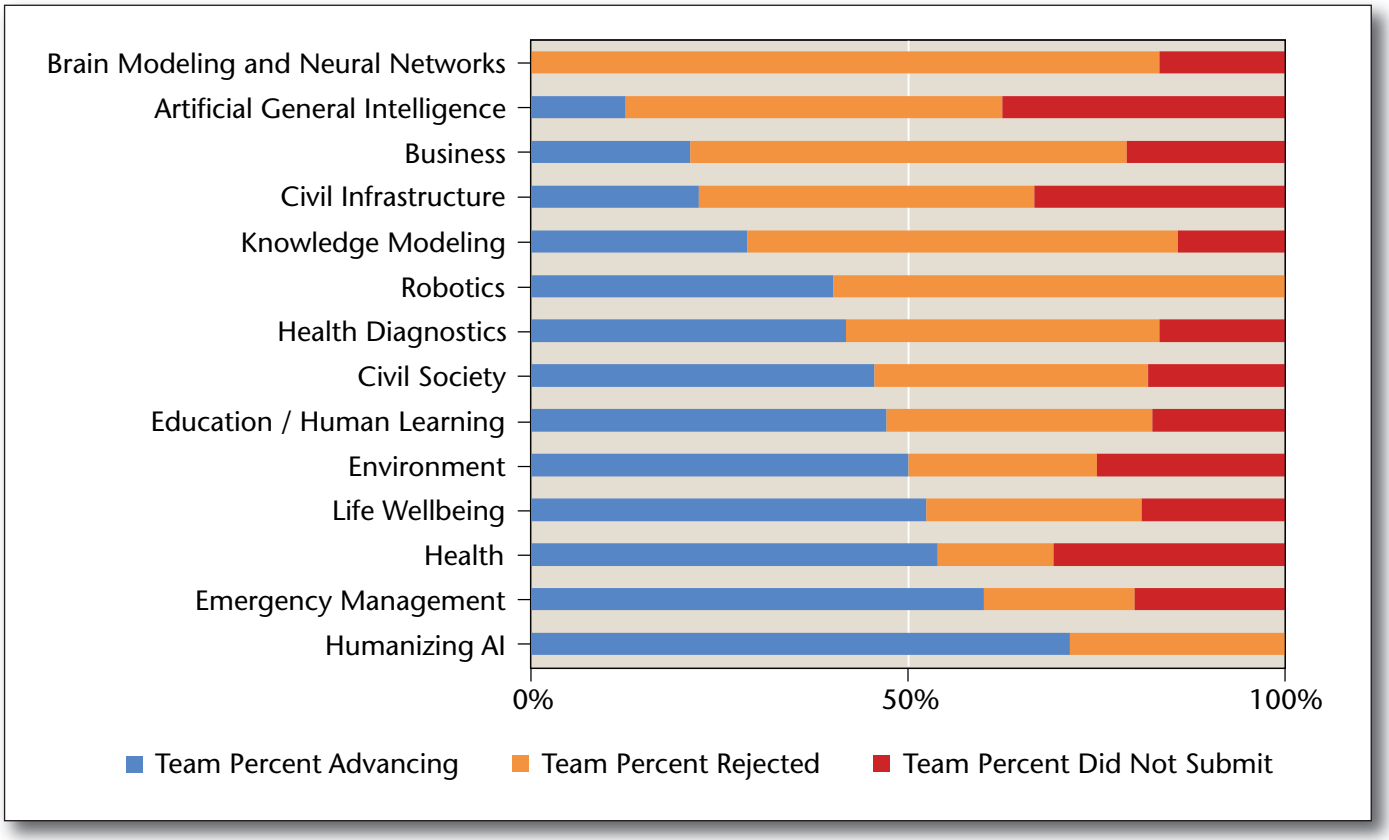


Figure 3. Team Advancement Bar Chart (Percentages).

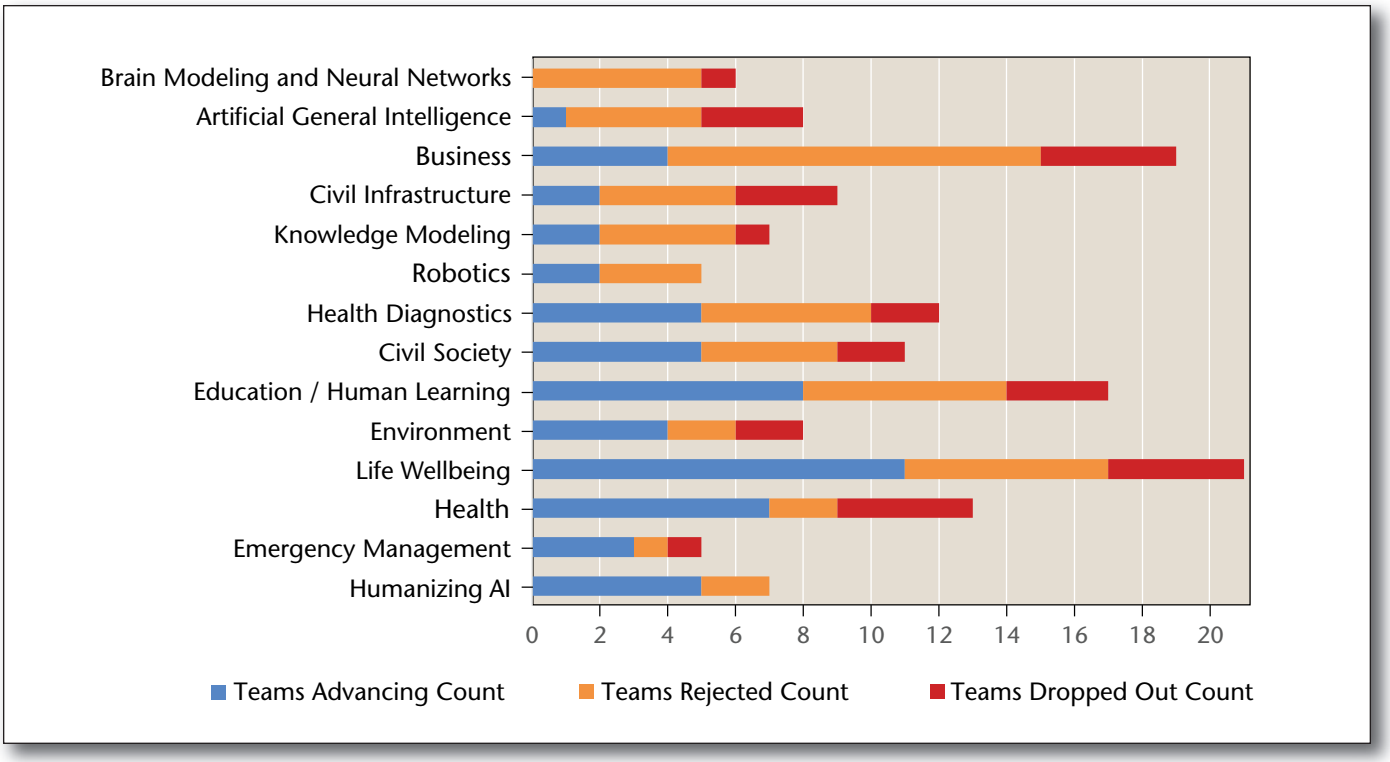


Figure 4. Team Advancement Bar Chart (Team Count).

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).

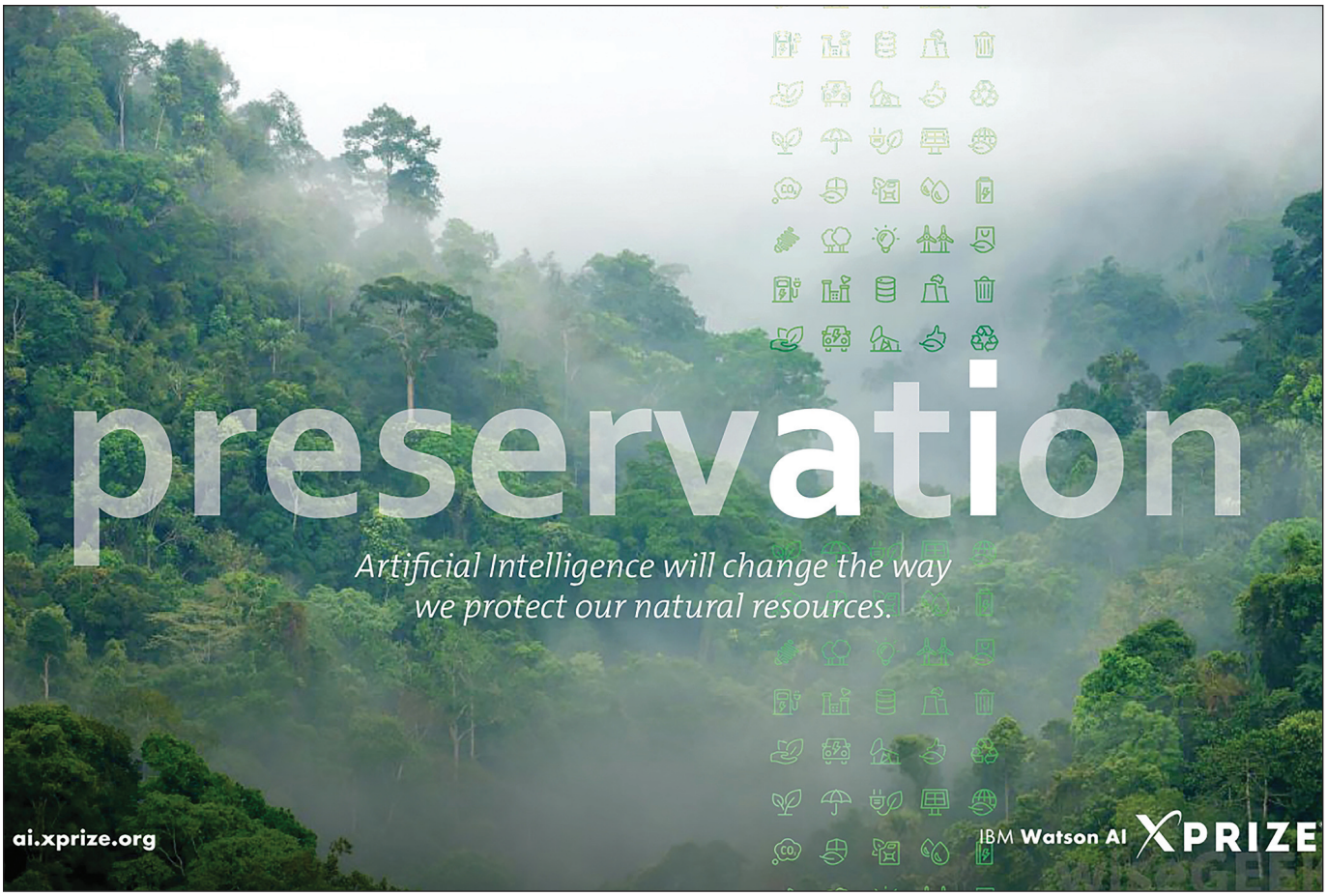
4. Such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (standards.ieee.org/news/2017/ieee_global_initiative.html).

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.