AI RFI Responses, October 26, 2018

Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan RFI Responses

DISCLAIMER: The RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government, National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (AI), NSTC Subcommittee on Machine Learning and AI, NSTC Subcommittee on Networking and Information Technology Research and Development (NITRD), NITRD National Coordination Office, and/or any other Federal agencies and/or government entities.
We bear no responsibility for the accuracy, legality or content of all external links included in this document.

Response to RFI on National AI R&D Strategic Plan

Contacts: John R. Smith, IBM Research AI, Mark O'Riley, IBM Government and Regulatory Affairs

The Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO), National Science Foundation, has requested input, including from those directly performing Artificial Intelligence (AI) research and development (R&D) and directly affected by such R&D, on whether the *National Artificial Intelligence Research and Development Strategic Plan*, October 2016, should be revised and, if so, the ways in which it may be improved. This document is the response from IBM Research AI to this Request for Information (RFI). The sections below indicate important challenges in AI R&D where strategic aims in the *National Artificial Intelligence Research and Development Strategic Plan* should be added or modified. The sections below are aligned with the seven Strategy areas in the current R&D plan.

Strategy 1: Making Long-Term Investments in AI Research

Making AI work in practice – learning from less data

Recent progress in AI R&D, notably in the area of Deep Learning, has produced dramatic gains in performance in accuracy of computers to perform tasks. This includes advances in computer vision, natural language processing, speech recognition and other machine capabilities for analysis and synthesis of unstructured and perceptual data. However, these advances are largely obtained within the limited scope of "Narrow AI". Given a well-defined machine learning task and large enough labeled data resources, Deep Learning can produce accurate Neural Network models. Examples include face recognition models that are trained from billions of face images or a language translation system that is trained from a large corpora of parallel language data. Narrow AI works well today for a small number of large-data tasks. However, the challenge facing industry for practical deployment of AI in production environments is that Narrow AI is not enough. There is an urgent need to broaden AI for industry and enterprise applications, which require support for a large number of small-data tasks. In some applications it may only be possible to obtain a handful of training examples such as in the case of detecting a rare manufacturing defect using a visual inspection system. Practical applications of AI in industry require increased long-term investments in AI R&D to ensure the continued development of methods that are able to effectively learn from much less data than is required for Deep Learning based systems today.

Applying AI for Decision Support

Industry applications of AI require systems that are not only accurate but fair, secure and explainable. These applications of AI typically influence human decisions. Consider the case of using AI to analyze skin lesion images for presence of melanoma. The trained classification

model can output a probability score for disease. However, this alone is not be enough information for the clinician or patient, especially given the high stakes nature of disease diagnosis. Beyond a prediction and probability score, the AI system needs to explain how and why the AI model reached its output. However, with the advances of Deep Learning, the Neural Network models themselves are becoming much more complex with hundreds of layers and hundreds of millions of weights. As a result, they are becoming harder to interpret. This is producing a widening gap of explainability, which impacts the core viability of AI to impact industry applications requiring decision support. Similarly, Deep learning models can have hidden characteristics that greatly compromise their performance in practice. The data-driven nature of Deep Learning means that models directly reflect the nature of its training data. If there are biases in the training data, the model will be biased, which can have undesired implications. The obvious example of this is the appearance of facial recognition systems in commercial use today that do not perform equivalently across the full diversity of human faces, when taking into account age, gender and race. Similarly, Deep Learning models can be poisoned by accidental or intentional introduction of samples of wrong training data. Addressing these challenges is important for moving from Narrow AI to the broader application of AI for industry and requires long-term investments in AI R&D. This sustained R&D is needed to advance trust in AI systems to meet industry's needs and adequately address requirements for robustness, fairness, explainability and security.

Strategy 2: Develop effective methods for human-AI collaboration

Human-Computer Interaction in Automated Machine Learning

We are seeing a recent rise in automation of AI models, from data augmentation to model creation as well as the use of AI for overall synthesis of application code. "AI for AI" brings new challenges in enabling trust between the automation and the data scientist, developer, or researcher responsible for the model. When should the automation defer to the human? What metrics must be exposed? What overrides and levers should be presented? What safeguards should be put in place? The automation also needs to be able to know when to bring in the expert to enhance results and when it can do better on its own. In essence there is a spectrum from being fully manual to being an assisted process to being a fully autonomous automation in the creation of AI capabilities. R&D advances are needed, not just on the generation of these models, but also on how the automation and the human effectively collaborate for the best outcomes. A goal of industry is to enable AI to learn more naturally and effectively. It is necessary to bootstrap existing systems to generate or speed up the creation of models. This includes the need to leverage the 'human in the loop'. This role of machine teaching is critical, and we must make advances in enabling humans to more effectively interact and teach machines. In Strategy 7, we further advocate for AI teachers as an important and emerging skillset and role.

Visualization in Explainability and Interpretability of AI Models

In the presence of models with low algorithmic interpretability, it is necessary to understand the mechanics of their behavior. Visualization advances, with a deep understanding of the underlying algorithms, are one way in which these models can be understood, improved and debugged. Recent advances include extensible dashboard tools that display core metrics all the way to determining what parts of a neural network are responsible for what feature or being able to improve models through an interactive visual experience, and finally being able to do 'what-if' scenarios. An example of the innovation and the thrust behind is highlighted in the first VISxAI workshop http://visxai.io.

Advanced Interaction Beyond Language

Machines not only have to understand language and intent but also support new modalities of interaction including voice, dialog, gestures and gaze. Innovation is needed in the ability for the machine to 'understand' the human and for the machine to respond in ways that are natural and comprehensible to the human. Thus, this requires dynamic behavior from the machine and advances in dynamic composition of AI skills and models. In addition, advances are needed in to ensure humans understand the machine's responses and drive interactions to successful outcomes. This requires an understanding of how humans perceive machine responses. Finally, the move to language-based interaction advantages users that speak the languages the AI systems are learning. This applies to other modalities such as gesture that can be culture dependent. Innovation is required to address these developments and ensure that there is balanced access and interaction regardless of language or culture factors.

Strategy 3: Understand and address the ethical, legal, and societal implications of AI

AI ethics for human-AI teams

AI systems are increasingly making decisions, but even more they are helping and supporting humans in making more informed and better decisions. The best results are obtained when AI augments and complements humans' capabilities, rather than trying to replace them. In these scenarios, humans are the decision makers, while AI systems support them in making more grounded decisions. Trust is essential in such humans-machine teams, otherwise AI will not be fully adopted, with the risk of not exploiting all its benefits. AI systems working with humans need to be designed to follow the laws and ethical norms that are suitable for a specific task, context, and culture where the AI system is going to be deployed. We therefore need to identify and define ways to model values and ethical priorities, to embed them into AI systems, and to use them also to improve the ethical stance of the human-machine team.

Value alignment

Data-driven approaches such as deep learning allow AI systems to be creative when choosing the path to solving a problem. Such freedom can be helpful since it can find solutions that humans would not have found. However, they can also allow undesired paths that the

examples fail to eliminate. When trying to embed values into an AI system, this could mean that unethical behaviors could possibly result. It is a challenge to understand how to safely and compactly model behavioral frameworks where AI's freedom and creativity, so successful to achieve great accuracy and performance, is guided to act within suitable ethical guidelines.

AI ethics and fairness by design

A technology that is not able to make fair decisions cannot be trusted. We need to understand how to identify the relevant fairness definition for the task at hand, as well as how to define bias detection and mitigation techniques. Designers and developers need to be helped in their everyday job, so that they avoid introducing unintended bias in the data or the models. Incentives to create diverse and inclusive developers' teams should be put in place, and such teams should be made aware of the possible sources of bias in data or models. Widely adopted mechanisms can facilitate explainability about the design choices that developers make about bias detection and mitigation, as well as other ethical properties of the AI system they develop.

Causal information for explainability and intervention

Even in AI systems where undesired bias is eliminated and suitable values are embedded, trust can only be achieved if such properties are usefully conveyed to the humans using the AI system or that are impacted by its decisions. Contextual AI explainability is therefore key. To achieve it, we need to identify how to combine data-driven AI systems, that are so successful in finding correlations in data, with capabilities to derive causal information. The capability to infer causal information will also allow us to go from prediction to intervention, so to be able to shape the future rather than just to predict it.

Multi-disciplinary and multi-stakeholder approach

Understanding how to embed values and ethical norms into an AI system needs a multi-disciplinary and multi-stakeholder approach, where also the impacted communities may have a voice. The creation of multi-stakeholder communities should be facilitated to improve and share a common understanding of bias and ethical issues in AI, supported by open initiatives, that include the sharing of open-source code for bias detection and mitigation algorithms, and unbiased training datasets.

Strategy 4: Ensure the Safety and Security of AI Systems

Use cases, threats, and attacks

The fields of AI and security intersect in three distinct and meaningful ways: the use of AI for improving the safety and security of systems; ensuring the safety, trustworthiness, and security of AI systems in production; and protection against AI used as a tool for harm. The advancement of AI for security is critically hampered by the unavailability of enough labeled data for training and evaluation. The security of AI systems requires mode investment in metrics and benchmarks for evaluation, as well as investment in the privacy and confidentiality of AI data and models, and corresponding use cases, such as federated learning. The

provenance of data and models, including tests and evaluations, builds accountability and mitigates some adversarial attacks, such as poisoning. The security of the runtime of AI models, at both training and inference time, is also important to ensure the models have not been manipulated, including new deployment paradigms such as AI at the edge.

Using AI for Security

The use of AI for bettering security can be accelerated by the increased availability of representative labeled datasets and benchmarks that can be used to train and evaluate models. Most security-relevant data is unlabeled, limiting the ability to evaluate accuracy and performance. Insufficient data also makes models less robust, provably making them more susceptible to adversarial attacks, overfitting (bias), and memorization (privacy threats). Security applications tend to be less tolerant to false positives than other domains, such as visual recognition and ASR (automatic speech recognition). Due to the high volume of data, even models with extremely low false positive rates result in an unsustainable number of alerts that must be adjudicated by an analyst, or it reduces the availability and integrity of the system it is designed to protect. Further, natural human behavior tends to have high variance and variability, making it easy for an adversary to blend into normal behavior or users and system processes (lest the model produces more false positives).

New Security Threats against AI

The privacy of AI data and models is more than an ethical consideration, it is also a security risk. An AI model can leak information about training data, such as membership inference (through overfitting, memorization, and confidence) or model inversion (inferring sensitive values from a target input given a known output). The models themselves can be stolen, either by direct reconstruction of model weights using known input-output pairs or by treating the target model as an oracle to label unknown samples and training a surrogate model, possibly with a different architecture. Investment is also required in metrics, benchmarks, and standards for evaluation the robustness of AI models against many attacks (adversarial samples, poisoning, inversion, etc.). Current best practice involves subjecting an AI model against the current state-of-the-art attacks in both white- and black-box settings and evaluating attack success given an Lp norm bound. First, this practice only ensures robustness against known attacks, and recently experience has illustrated this is a flawed technique resulting in defenses being broken quickly, often before formal publication of the defenses. Second, the reliance on Lp norms is insufficient. First, Lp norms don't correlate well with human perception. Second, Lp norms have little meaning in many domains, such as malware. Finally, there is an tradeoff between model accuracy and robustness that must be better understood to understand the limits an AI model can be deployed in practice.

Protecting Against AI-powered Attacks

Adversaries are beginning to leverage AI as a new tool in their arsenal, turning it into a weapon that must be defended against. It has been argued that AI will be as important as nuclear power for national security, where "activities that currently require lots of high-skill labor, such as

Advanced Persistent Threat operations, may in the future be largely automated and easily available on the black market." The use of AI will make AI-powered attacks: more evasive by hardening attacks against deployed defenses, cautious and stealthy not to trigger alerts and fly under the radar; more pervasive through automation and scalability and removing human bottlenecks; and more adaptive, being creating, and identifying new vulnerabilities and weaknesses to exploit. The first glimpse of this possible capability was illustrated in the GCG.

Strategy 5 - Develop shared public datasets and environments for AI training and testing

Addressing Critical Blockers to Effective Data Set Sharing and Use

As the National AI R&D Strategic Plan currently states, data sharing is essential for AI. The sharing of datasets is mandatory yet is still massively hindered by opacity in terms and conditions as well as data use policies and regulations. We need ways to safely share data and unlock their value, for both corporations as well as research institutions. Innovation is needed on methodologies and license terms and conditions to enable this sharing in a clear and frictionless manner across different contexts, as well as standards on meta-data specifications for describing these datasets, including the provenance information.

The creation of simpler and more standardized data sharing and licensing models will create greater certainty for large and small entities and reward developers who carefully play by the rules. For example, the Linux Foundation has made available model Community Data License Agreements (https://cdla.io/) to provide a licensing framework to support collaborative communities built around curating and sharing "open" data. Efforts to create datasets for which permitted use in the field of AI R&D is both clear and broad, including with respect to the requirements of privacy law, will pay great dividends in reducing transaction costs currently associated with legal uncertainty. The solutions lie in better licensing terms and methodologies for safe creation and responsible use of shared data that spurs scientific progress.

Recommendations:

Institutional and policy innovation is required to ensure that a) the data gathering practice produces data with integrity and availability in mind and that b) new standards are created for data usage terms and conditions and meta-data specifications for describing the data sets and their provenance.

Several emerging technical areas could be the key to innovation in this space. Privacy protecting federated learning is a promising area that can perhaps balance the needs for protecting privacy and the needs for more data sharing for better AI by enabling collaborative learning without giving away one's data. The increasingly popular blockchain technology has great potential to ensure scalable and secure data governance.

Developing public shared AI models and environment to share them

In addition to sharing datasets, the sharing of models should be equally encouraged. Model 'zoos' are emerging for different communities with no standard format or structure. Additionally, innovation is needed to standardize on model quality metrics across dimensions beyond accuracy: the lineage of a model, its license and usage terms taking into account the terms of the data on which it was trained, metrics on robustness and fairness. There has been some movement in this direction such as in IBM's push for AI FactSheets.

Advocating for Open Government Data and Models

In alignment with the National AI R&D Strategic Plan goal to make a wide variety of datasets accessible for AI, we advocate for open government data, which is critically important for AI. Currently, only a fraction of existing government datasets is available in full, free, and usable formats. We support government efforts to open more of its datasets in accessible and usable formats with significant positive implications through integration into AI systems. Government data reflects transparent collection methods, particularly when provided with data provenance since users obtain data directly from the source. As such, when these data feed into AI systems, they add transparency to the systems. AI systems using government data as input are more transparent since the source of data is clearly known. Open government data also open a vast resource of high quantity and quality of datasets. Currently the reserves of open data are a very limited pool of input, with which we have still made remarkable gains in AI technology. With more data, we can have more AI applications and more accurate AI systems. Furthermore, the quality of AI solutions relies on the quality of data input. Thanks to agency collection standards developed decades ago, government data offer long and consistent records of information. The longevity and consistency of data records provide robust datasets that can contribute to robust AI systems. Perhaps even more importantly, open government data are a valuable tool in the AI battle on bias. They provide a vast number of diverse datasets from different regions, economic classes, and sectors. The more diversity reflected in data translates to more diverse AI systems and outcomes. Furthermore, government data also reduce the digital divide because they represent all parts of the population, not just the people with digital access, which most data sources only represent. Finally, in many cases the greater number of datasets available for input allows for a larger sample size within which to compare, detect, and eliminate a biased dataset.

Open Platforms and Reproducibility

We are encouraged by the emergence of cloud training and testing environments, as well as the movement for AI conferences to encourage reproducibility in papers through repeatable experiments and code. The start of benchmarks such as DAWNbench are also helpful in advancing and leveling the playing field. Open source has proven to be a very strong factor in AI, which is also very promising in terms of the future. Open APIs have increased innovation that scales aggressively. Therefore, we recommend an increased focus and to continue to encourage and incentivize these efforts.

Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks

Developing benchmarks for explainability, fairness and security

Standards and metrics have a critical role in advancing the development and application of AI technology. Standard benchmarks, data sets and open challenge tasks such as ImageNet (www.image-net.org) have been essential for propelling AI R&D forward in areas such as computer vision, natural language processing and speech recognition. While standards and benchmarks have been vital for obtaining more accurate AI models, standards and benchmarks need to address new requirements for broadening AI for industry and enterprise applications. AI systems to be explainable, fair, secure and robust. New standards and benchmarks are needed to advance the relevant AI R&D and assess and validate the performance on AI systems in practice on these dimensions. A good example of this is the assessment of fairness of facial recognition systems. Industry applications of facial recognition need systems to be not only accurate but also provide the same level of accuracy regardless of human diversity across age, gender and race. While a greater number of datasets and benchmarks have been created for measuring the accuracy of facial recognition, there is an outstanding need to ensure and assess fairness of these systems. Consider for example the NIST Face Recognition Vendor Test (FRVT) (https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing). The NIST evaluation provides a benchmark for assessing the accuracy of facial recognition in terms of false non-match rate (FNMR) across several face datasets. However, this benchmark does not adequately address fairness and support the full diversity of human faces. Organizations like NIST can increase support for standards and benchmarks for explainability, fairness, security and robustness.

Strategy 7: Better understand the national AI R&D workforce needs

Developing a workforce of AI teachers

The development of AI systems today is human-labor intensive and requires significant expertise not only in technology areas such as Deep Learning but also at the level of translating application and industry domain requirements into repeatable data science processes. While the development and availability of AI platforms, tools and software has improved, a gap remains in terms of National AI R&D workforce to support the advancement of AI and deployment for industry. Additionally, given the fast pace of development of the AI field, a sustained workforce is needed to conduct basic and applied research to ensure national competitiveness in AI. At the level of application and industry domain requirements, a workforce of **AI teachers** is needed. Given the data-driven nature of development of AI systems, these AI teachers need to be skilled at obtaining, curating and labeling appropriate data resources for training AI models. The AI teachers are responsible for the developing the curricula for training AI models. As a result, the ability to scale the development of AI systems is gated by the number of AI teachers. Thus, there is a need to better understand needs at a

national level for the number and domain expertise of AI teachers. Similarly, at the technical level of AI model training, data scientists have a critical role. The data scientists make fundamental decisions on Neural Network architecture, which requires significant expertise. The data scientists are also increasingly burdened with maintaining operational AI systems that are subject to continuously changing workloads and drifts in data distributions. As AI is deployed more extensively in industry this maintenance will require a far greater numbers of AI teachers and data scientists.