Artificial Intelligence Measurement and Evaluation Workshop Summary

Introduction

The Artificial Intelligence Measurement and Evaluation Workshop was a three-day virtual workshop that took place June 15-17, 2021. The workshop brought together stakeholders and experts to identify the most pressing needs for AI measurement and evaluation.
NIST is assigned responsibility by statute to advance underlying research for measuring and assessing AI technologies. That includes the development of AI data standards and best practices, as well as AI evaluation and testing methodologies and standards. NIST is working collaboratively with the private and public sectors to help prioritize and work on its AI activities. The full workshop agenda as well as all of the available video recordings of the panels are available on the Artificial Intelligence Measurement and Evaluation Workshop page.

Background

Each day of the workshop started with a keynote address, followed by 4 panel discussions. Each panel discussion had a moderator and three to six panelists. Each panel began with the panelists providing some introductory remarks on the topic of the panel followed by a discussion led by the moderator. Throughout the workshop, a slack workspace was open to all workshop participants, where they could ask questions and provide their own comments and thoughts on the topics under discussion. Appendix A: References from Panelists and Participants contains a list of suggested references and links that were provided in the slack channel or by panelists during each of the panels.

Panel Summaries

This section contains brief summaries of the discussions for each of the panels from both the panelists and the slack discussions.

Panel 1 - Measuring with Purpose

The panelists discussed the unique challenges of machine learning (ML) and artificial intelligence (AI) systems when it comes to risks, such as bias, robustness, and explainability. There are risks in AI systems where existing regulations apply, such as safety components of regulated products, health systems, and areas of law enforcement. The panelists suggested AI systems could focus on efficacy measures such as: what customers want and what their concerns are. The panelists suggested focusing on measures that matter and not what is easy to measure. The evolution of metrics is key to maintaining efficacy of those measures. A hybrid of various risk-based approaches with respect to characteristics that are desirable in the system was suggested as a possible best option.

Tess DeBlanc-Knowles of White House Office of Science and Technology Policy moderated and introduced the panel, highlighted the importance of AI technologies to the public and private sectors, and emphasized the need for trustworthy AI. The ability to measure and evaluate the performance of AI throughout the lifecycle of an AI system was stressed as a key component of establishing trust. Various needs to address existing gaps in the common understanding and evaluation of metrics and benchmarks for reliability, and the description and impacts of failure modes, among other issues were mentioned. The moderator also expressed a need for shared infrastructure for the development and testing of AI systems.

Michael Hind of IBM Research began by introducing a comparison of ML work with traditional software evaluations. Both traditional software and ML evaluations involve verifying that the actual output matches the expected output for a given input, but there is a difference in testing corner cases, which could be well-defined in traditional software, but are often not well-defined or are unknown for ML. Hind identified a non-exhaustive list of risks which included bias, adversarial robustness, explainability, uncertainty, privacy, generalizability, and data quality. When performing evaluations of AI systems, it is important to know who the consumer of the evaluation will be and what is the desired outcome for the tasks the system will be performing.

Salvatore Scalzo of the European Commission discussed the European Commission's proposed framework on AI released on the 21st of April, 2021. Key regulatory concepts included: internal market legislation, a layered risk-based approach, and a level playing field for EU and non-EU players. The framework takes a broad definition of AI in order to cover as many techniques and approaches as possible with an enumeration of those techniques and approaches which are covered as an updateable annex to provide legal certainty. The risk-based approach intends to allow low or minimal risk applications to go ahead unhindered and to layer additional responsibilities and obligations as the risk increases from minimal risk, to transparency risk, to high risk, to unacceptable risk. Transparency risk (e.g., impersonation/bots) is where there is an obligation to inform or be transparent about the use of AI in situations where the AI will be used. High risk applications (e.g., medical devices, recruitment) will be subject to compliance and conformity assessment requirements. Unacceptable risk applications (e.g., social credit scoring system) are prohibited. The framework mandates a number of obligations for providers including quality management, technical documentation, conformity assessment, system registration, and post-market monitoring.

Jane Pinelis of the US Department of Defense (DoD) Joint Artificial Intelligence Center (JAIC) began by pointing out that a robust test process is one which verifies that the systems under test will work across a full system of conditions. The DoD has a challenging set of operational requirements and unique mission sets further complicating comprehensive testings. Testing is a concern and in order to deploy at the speed of relevance, the US will be challenged by systems' lack of performance guarantees that outweigh any benefits, if behaviors are not known, understood, or recognized. AI brings with it new failure modes, inherent dependence on training data and methods which directly tie to data quality and representativeness. The JAIC has created a test and evaluation framework for AI that has improved operational testing within the DoD. Some principles include using Design of Experiments methods and "Survival Analysis". Four pillars of testing at the DoD are: 1) testing the algorithms on reserve test data, 2) system integration, 3) human systems integration, and 4) operational tests. The JAIC publishes best-practice guides and provides shared access to technology including cloud-native test harnesses.

Bill Scherlis of Defense Advanced Research Projects Agency (DARPA) described four areas of focus. The first area is AI evaluation and trustworthy AI in the context of mission applications. Adversarial AI is a well-established research area; AI models can be spoofed, and this drives the need for secure and trustworthy AI. Modern ML is not trustworthy and is not likely to become so to a sufficient extent for mission autonomy. DARPA is trying to develop systems with a certain amount of autonomy with the requirement of trustworthiness. DARPA has a program called "Assured Autonomy" and another called Guaranteeing AI Robustness Against Deception (GARD) which focuses on modeling various kinds of attacks and malicious inputs and how to achieve resilience. The second area concerns data, how to protect, how it is used, and how to do ML with much less data. "Learning with less labeling" and self-supervised transfer learning both aim to reduce the extent of labeled data required. The third area is overcoming some of the intrinsic challenges of pure ML, including opaqueness. Explainable AI is an attempt to make models more transparent, but there is a danger that the natural language explanations can provide an appearance of robustness that may not be valid. DARPA is looking at hybrid systems which blend statistical approaches with symbolic AI approaches. The fourth area DARPA is interested in is human interactions with systems. While progress is being made, the problem is getting more difficult to solve as tasks become more complex and systems become more capable and their roles become more intertwined with human team members.

Jack Clark of Anthropic introduced his work on "Frontier Models" and performing test and evaluation for large scale computationally intensive model developments. There is a need to perform more than one set of tests against a model. Anthropic has found in their work that it is better to have many somewhat reliable tests than to have a small number of very reliable ones. When a single system is tested fifty different ways there is an opportunity to characterize it. In addition to using existing tests, Clark believes new tests need to be developed and it is best for those coming up with the tests if they are co-located with the technical developers of the model. Thoughts on future trends included how to decide whether a system is giving an accurate explanation of an output instead of something we wanted, investigating large models which appear to display the "few shot learning" trait (a large model and with minimal new class data has good capacity for identifying the new class).

Chuck Howell of the MITRE Corporation (MITRE) started by inviting everyone to review the report for the National Security Commission on AI, where he served as a member of the research and analysis staff for the committee. Howell then introduced the idea of tool qualification and how this is an important concept in avionics, where the tools used to develop, assess, and test the avionics box are themselves competent and appropriate for use. For example, when looking at a particular recommendation or classification of an image and attempting to investigate why the models arrived at a particular output, when the tools used to perform this analysis are themselves subject to scrutiny, they often come up short. MITRE is looking for "rhyming challenges" and opportunities where problems and solutions in some other domain could potentially be applied in the AI domain.

Panel Discussion: The moderator asked, "Where can federal policy play a role in advancing AI evaluation?" The discussion began on the topic of difficulty for companies to build the appropriate data sets for test and evaluation, especially when looking at fairness and bias. Government could help with constructing these data sets and providing shared computational resources to help in performing the test and evaluations. The point was raised that there are often unclear problem descriptions and that it would be helpful to develop clear requirements that would assist engineers and scientists in knowing what to look for. The ability for 3rd parties to perform "glass-box" testing[1] was also raised and setting up some capacity for performing this

[1] Glass-box testing is defined by ISO/IEC/IEEE 29119-1:2013 as, "dynamic testing in which the tests are derived from an examination of the structure of the test item."

type of testing. Transparency was also mentioned as an important consideration and the need for systems to have hooks and other capabilities to allow for ease in test and evaluation.

The next question presented dealt with the definition of data quality in data sets. There was general agreement that data quality is important and that mechanisms for documenting the data sets and the constraints and limitations of the data sets was important.

Finally, a conversation around whether to emphasize robustness above performance in testing benchmarks was worthwhile had some disagreement from the panelists who felt that a risk based approach may be ideal in that what factors to focus on in test and evaluation would be driven by the risks presented by the system.

Slack Discussion: One point raised in the slack discussion that was not mentioned by the participants was the idea that data sets must be dynamic, with the example given that a corpus of text from the 1920's used to parse common speech in 2020 would not be an ideal data set to start from. Additionally, there was some conversation about how to measure the quality of a data set.

Panel 2 - Overview of Past & Current Evaluations

This panel discussed a key challenge to current and past evaluations: building sufficiently large test collections is unfeasible. An approach suggested by the panelists was to look for the easiest solution to understand, verify that solution, and maintain it. The panelists discussed the desire to build evaluations from impactful effects on the system and looking to provide feedback that is intuitive and actionable. The panel suggested there is a desire for a Risk-Based Evaluation Framework which could be easily implemented. The panelists highlighted that accuracy does not always imply understanding.

Mark Przybocki of NIST moderated the panel. In his opening remarks, he described the NIST Evaluation Research Paradigm as a cycle with stages: planning, data, evaluation, workshop, and (back to) planning. Mark then described several evaluation models that are used by NIST depending on the maturity of the technology and the goals of the evaluation. These evaluation models include: challenge problems, honor system approach, system delivery, progress tests, ongoing leaderboard, and human in the loop or assessor evaluations.

Ellen Voorhees of NIST described the work of the Text REtrieval Conference (TREC), a long-running set of natural language processing (NLP) evaluations. Each evaluation of offline retrieval used a test collection, which is a benchmark task that consisted of documents, queries, and relevance judgements. TREC pioneered the use of pooling to build large test collections and has gone on to build hundreds of collections for dozens of different tasks. TREC has shown that an emphasis on individual experiments evaluated in a common setting can leverage a relatively modest investment by the government into a significantly greater amount of research and development.

Peter Bajcsy of NIST described the work of TrojAI, an evaluation for detecting trojan behaviors in AI models. The experiments were designed for image classifications, with an image having a background with a superimposed foreground, and then some triggers (such as polygons) put in the foreground images. In addition to submission and evaluation, they prepared an interactive web-based trojan simulator, a baseline trojan detector, and a survey of relevant publications.

Jonathon Phillips of NIST discussed his thoughts on “Evaluation to Experiments” and a discussion of face recognition evaluations, including the FERET evaluation as well as the series of Face Recognition Vendor Test (FRVT) Evaluations, which NIST designed. One of the key elements of FERET was data collection. The data was partitioned into two sets: one data set was given to researchers to develop algorithms, and the other was sequestered for testing of the algorithm under proctored evaluations. Jonathon then talked about how they applied a psychological experiment to facial recognition algorithms to measure the racial bias of facial recognition algorithms. Last, Jonathon talked about how using empirical methods can help build insight into these AI methods and how they would perform.

Michael Sharp of NIST discussed an overview of Industrial Artificial Intelligence (IAI) Management and Metrology. Industrial Artificial Intelligence is a subset of AI applied to industry. The difference that distinguishes IAI from other Artificial Intelligence is that IAI is meant to solve a known problem or provide an explicit benefit to an industry or company. As a consequence, practical use is a higher priority than philosophical elegance, and justifications for modeling choices must come from the context of the application. Michael then discussed how many domain goals can be broken down into the framework of risk management. The same risk can be measured in monetary loss, it can be measured in performance output loss or gain, and it can be in terms of the impact to the public perception of an organization. Risk can be informed with access to these metrics.

Jonathan Fiscus of NIST discussed metrology lessons from past evaluations. The most striking lesson is that often we have to invent new evaluation protocols for new technologies. This requires a community effort. Going more granular, start with a small problem. The concept of beginning with a small, solvable problem and scaling it as technology improves has been used multiple times. Also, accuracy does not imply understanding. The biggest point, learned from the multimedia event detection evaluation, is that big data doesn't mean understanding; they are inversely related. We need to focus on small problems and solve those to get better technology.

Megan Zimmerman of NIST discussed test methods, metrics, and evaluation of AI in Applied Robotics. These evaluations of AI are grounded on specific applications, based on tasks where benchmarks and tests are provided. In Robotics, there are three main processes: perception, cognition and reasoning, and physical control. Most of their work has been benchmarks provided either with task boards or simulated environments. They discussed that their work and collaborations aim to help establish more benchmarks for assessment tasks.

Panel Discussion: The Panel Discussion with these panelists was combined with that of Panel 3.

Slack Discussion: During this discussion people raised issues of ethics and trust in AI and the necessity for ethics and trust in AI. The conversation mentioned that one needed to think about who the benefits are accruing to.

Panel 3 - Discussion of NIST/Community Future Work

The panel discussed different AI domains, such as ML/Neural Nets/NLP, and the requirements for different evaluation techniques in each domain. There are fundamental limitations of what each domain does, based on the algorithm but especially on the problem sets of each domain. The metrics used in evaluation of one domain may not apply to another. Some AI domains are very task oriented, and the expectations and abilities of the systems change so much that evaluating these systems will always lag the underlying changes. The panel suggested there is no single innovation for the evaluation of AI for instilling trust in the use of AI. Some suggested approaches are adding transparency to the system and demonstrating the value of an AI system.

The moderator and the panelists in Panel 3 are the same as in Panel 2. Panel 3 is the discussion component of Panel 2.

Panel Discussion: The panel started with the discussion of how the underlying approach of AI can change when being evaluated in different areas. The physical limitations (the speed of a car, etc.) can be used and considered in the evaluation design. There is also a goal to make an environment that is specific enough to be evaluable yet general enough to be compared to other environments (sometimes the environment can be too specific). Last, the panel mentioned the desire to separate out the application-agnostic parts (such as packaging and receiving submissions) and the application-specific components (such as having evaluation-specific metrics).

Another core concept discussed is what would have the largest impact of trust on the use of AI. The panelists first mentioned that having ethics guidelines are important. Another panelist mentioned that whether a system is trustworthy is different from if users trust a system: users could trust a system that is not trustworthy and vice versa. The panel mentioned that showing users the value of the tool or providing ways for users to interact with the tool can help users trust systems. Last, the panel mentioned that there are metrics within human-robotic interactions which attempt to determine the trust humans have in a particular system through the use of surveys to develop a quantifiable measure of trust.

The panel ended with a discussion of designing evaluations for application-agnostic tasks vs. evaluations of specific systems in particular domains. One panelist discussed a case where they had to simplify a task to a stark abstraction of the original task in order to make progress, and that this abstraction may be necessary at the beginning in order to make progress with some speed. Additional domain specific evaluation can and will attempt to bring the full-complexity of the problem back to focus once advances on the core abstract problem have been made. Another panelist mentioned the challenge that the pace of change in the frontier for the development of new models is faster than the development of evaluations that can adequately perform an evaluation to reassure stakeholders that the models are of high-quality.

Slack Discussion: The slack discussion of Panel 3 was in the Panel 2 channel, see Panel 2 for slack discussion.

Panel 4 - Evaluating AI During Operation

This panel focused on making a connection between evaluating AI and operation as well as the broader problem in machine learning (ML): retraining. The panelists discussed a need for ML systems that are continually learning. A key challenge is defining how, when, and which data is used to retrain the model. The panelists suggested that another challenge is model drift. Panelists discussed the trade-off of cost and other impacts to an organization using an AI system in operating environments. Panelists highlighted the use of poor models because they're doing a good enough job and the core question of what is good enough?

Antonio Moretti of Walmart moderated the panel on evaluating AI during operation and introduced Clarence Agbi of Capital One, Sergey Karayev of Turnitin, and Josh Tobin of Gantry, the panelists. He introduced the panel members and gave highlights on AI evaluation during production and operational environments, Machine Learning Operations (MLOps), operational evaluation metrics, model quality, data drift, latency, throughput, scalability issues, and issues around security and robustness.

Josh Tobin of Gantry described that static ML models are used in academic settings and continuous ML and active learning models are used in production environments, where models are regularly trained on new data, since data is never static and is constantly changing and there is data drift compared to training data set. He also discussed the use of MLOps in production and continuous learning and continuous evaluation. He co-organized the full stack deep learning class program and posted few references related to it.

Sergey Karayev of Turnitin described the ML work that he is doing for an education tech company that uses AI assisted grading of tests and exams automatically. He described the model development and how they handle privacy and security based on regulations, how they train and retrain their models based on data drift, and how they monitor the system and the correctness of their prediction. He also co-organized the full stack deep learning class program.

Clarence Agbi of Capital One described evaluation of models at a level that goes beyond the model itself. He then mentioned that all the development is to support or improve the business and the business metrics are at the heart of evaluations in production. He talked about his work in highly regulated industries, where you can’t deploy a system without thinking about its effects on users or the effect on business. He then described the development of ML systems as mainly based on business use cases. He finally highlighted the use of poor models in business because they're doing a good enough job and making profits for the business.

Panel Discussion: The first question discussed: how does the business model affect the issues around model deployment? And can you give examples of metrics that you will use? The discussion revolved around different applications and the importance of evaluation metrics and the cost of creating annotations.

The next discussion was on continuous ML and active learning models vs. static ML and the importance of it in business applications. There was general agreement that continuous ML is important for business applications, where data is constantly changing compared to static models, which are popular for kaggle challenge data sets.

Then, the next discussion topic was regarding the core question of what is a good enough AI application for a business.

Finally, there was discussion of the question, for a business, how does ML or data science relate to the core product?

Slack Discussion: There were discussions related to continuous ML, scalability issues, and issues around security and robustness. Another discussion was on business applications and metrics. Additionally, online references were posted on a number of topics discussed in the panel.

Panel 5 - Evaluation Design Process

The panel discussed the evaluation design process, robustness, difficulties in bias, fairness, and subpopulations. It's important to look beyond one accuracy number in robustness. There are many important parts to doing an evaluation which is not just running a single number on one, calling it "benchmarking" and calling it a day. It's important to report negative results when it comes to machine learning (ML) models. It is also important to continue to improve the existing complex data sets and recognize their shortcomings.

Nicholas Carlini of Google Brain moderated the panel on evaluation design process and introduced the panelists which consisted of Matthias Hein of University of Tübingen, Deborah Raji of Mozilla Foundation, Shibani Santurkar of Stanford, and Ludwig Schmidt of University of Washington. He then discussed topics related to the evaluation design process, how to set up an evaluation, benchmarking, data sets, robustness and evaluation metrics beyond accuracy.

Matthias Hein of University of Tübingen discussed his main research area, which is to make machine learning robust, safe, and explainable, and his main focus at the moment is robustness, in particular, adversarial robustness. He is working on elevating and providing tools to the users to automatically assess adversarial robustness, and his other main focus is reliability and certainty quantification, in particular, detection of out of distribution inputs.

Deborah Raji of Mozilla Foundation described her interest in algorithmic auditing, and discussed how the community has been formulating or articulating that goal, which is by evaluation or assessment with the objective of holding institutions accountable. We all know during the ML development process, ML engineers are making lots of different decisions, and she wants to design evaluations so that we can assess the quality of those decisions from the outside or from the inside. She has worked on projects with teams internally within corporations trying to understand these decisions and to map them to certain consequences and articulate that through evaluations. She also mentioned her work with the Algorithmic Justice League for people that want to evaluate the system as outsiders without access to that system and how they go about designing those evaluations.

Shibani Santurkar of Stanford talked about her focus on making ML models more robust and reliable. She started with thinking about adversarial robustness, but recently she has been thinking about the broader notions of robustness in terms of how to ensure that models work, not only on the test benchmarks, but also on real world applications.

Ludwig Schmidt of University of Washington talked about his interest in AI evaluation starting with the release of ImageNet benchmark and data set. One thing that has always worried him is the way the community evaluated ImageNet, and that is the opposite of how evaluation is taught in ML class. The research community usually reuses the validation set and test sets many times, which could lead to overfitting. This process has helped him figure out how not to overfit the model, but many other interesting questions of ML data sets and evaluation have guided his research recently.

Panel Discussion: The first question discussed was, why do we need to think about evaluations? Why can't we just say, all of our vision models are going to be evaluated on the ImageNet test set, call it a day and just move on with our lives?

The panelists responded that it's possible that this would lead to overfitting on the ImageNet test set and not making progress on types of images not present in ImageNet.

Then the panelists were asked: one piece of actionable advice from each of you - If you could have evaluations done however you want it to be, what would that one thing be that people do?

The panelists offered the following advice: (1) Do not concentrate on a single metric, instead agree on a set of metrics (e.g., several robustness metrics) and then measure using all the agreed upon metrics. (2) Perform disaggregated analysis, including on sub-populations and intersectional sub-populations, instead of focusing on only a single measure. (3) Make evaluations as easy and reproducible as possible. (4) Build a set of evaluation benchmarks where it is the norm to evaluate on all the benchmarks to assist in comparing algorithms (rather than picking a choosing).

The panelists had a discussion about algorithmic auditing, robustness and fairness and evaluations.

Then, the panelists also had discussions about privacy, health data and privacy, federated learning, differential privacy, etc. and privacy and regulations.

Finally, the complex question was discussed: How do you apply auditing in complex spaces, such as medical devices or software regulation, where even the initial evaluation of these products has unresolved biases?

Deborah Raji responded that it’s a difficult and still open question, but it involves a multidisciplinary approach, asking and heavily documenting, especially asking and answering qualitative questions.

Slack Discussion: One thread of discussion focused on robustness and evaluation metrics beyond accuracy. Another thread was a discussion on overfitting the model and on adversarial robustness.

Panel 6 - Metrics and Measurement Methods

The panelists discussed a set of components which could be used to measure AI systems. These components are: a clear and defined task, a very good metric to measure how well the system is doing what it is designed to do, and data. The range of metrics goes from the Measures of Performance (represents quantitative measure of core technology — easy to repeat, defined objective, cheap, suitable to a machine consumer) to Measures of Effectiveness (more application focused, represent quantitative/qualitative measure of how technology helped final application — hard to repeat, fuzzy objective, expensive, suited to human consumer). The panelists discussed a potential conceptual loop about machine learning (ML) research: collect data, train the model, evaluate, and then deploy. The notion of what should be measured depends on the purpose of the evaluation. The rise in the benchmarks will typically be at the beginning of new technology; the panelists discussed whether or not the AI space is in the benchmarking phase.

Craig Greenberg of NIST moderated the panel. In his opening remarks he introduced the panelists and then discussed the goal of the panel, which is to discuss which aspects of the AI systems can and should be measured and evaluated and how that can and should be done.

José Hernández-Orallo of Universitat Politècnica de València discussed measurements and metrics: what and how to test. José argued that for AI, evaluation should move from being task-oriented to capability-oriented. In a capability-oriented evaluation, each system has a capability profile, and systems are matched with problems whose problem profiles overlap with the capability profiles. José discussed a project of measurement layouts, which tries to identify not only the capabilities, but capabilities in terms of probabilities, and skills of a system. More nuanced metrics (beyond additive metrics) must be developed and used.

Douglas Reynolds of the National Security Agency (NSA) / MIT Lincoln Laboratory focused on the practical side of evaluations, which drives ideas from inception to being something useful. The virtuous research and development (R&D) cycle involves a mission or need that is turned into a challenge problem, unclassified R&D, classified R&D, a prototype, to deployment, to a new challenge problem. The final number is not the purpose. It is a motivator, but more importantly, the number needs to promulgate out the best ideas. More than just accuracy needs to be measured, additional things which might need measurement include throughput, memory, and other system aspects. The components needed for an AI/ML evaluation are a clearly defined task, a very good metric that measures how well the task is accomplished, and the data for learning and for performance evaluation. The types of metrics range from measures of performance (technology focus, quantitative error rates, etc.) to measures of effectiveness (quantitative or qualitative measures of how technology helps a final application).

Sameer Singh of University of California, Irvine talked about testing and explaining NLP models as a way to evaluate them. Sameer described a mismatch between what the accuracy models say and what the systems look like in practice, and provided examples involving different queries to the system. Sameer would like to make this process of identifying these problems as early in the pipeline as possible, and would like to make it as part of the evaluation of these AI systems. The first solution Sameer has been working with is called behavioral testing. Behavioral testing involves taking an original test instance and having a trained system produce a prediction on that instance. You then change the test instance in a specific way, obtain a prediction from the system, and then compare to see if the prediction on the new instance is the expected prediction. Sameer also looks at explanations and sees (through human explanations) if those explanations of the systems are correct or not.

Panel Discussion: The first question discussed what are the properties of AI systems that can and should be measured. Jose mentioned that we want the traits that predict or explain system behavior (how to identify these metrics is a question). Doug mentioned that what should be measured depends on the purpose, audience, and needs of the evaluations. As evaluations mature, things other than accuracy matter. Sameer mentioned that some things are not addressed in evaluation. One of those is calibration: are those probabilities meaningful? Another is how a system would behave on future unseen instances (which may mean getting more data or changing data)

The panel discussed what properties metrics might possess. Desirable properties include: simplicity (the metric be as simple as possible); additive (the sum of the metric on two smaller sets is the same as the metric on the combined set); that the metric have a ratio scale; and that what the numbers mean is easily understood (easy to explain the meaning to a human).
The panelists were then asked to offer concrete suggestions for NIST with the following responses. To identify the language of specifications for ML is needed. What are the aspects for which accuracy measurements are desirable? What are our tolerances? What can be tested? Have benchmarks with multi-modality (different distributions, not necessarily different data types). NIST has a position of a reference for the community to help provide guidelines. How was that result achieved? How much data and compute was used?

Slack Discussion: One discussion was on behavioral validity and its definition. Another was a discussion on the differences between tasks and capabilities.

Panel 7 - Data and Data Sets

The panelists discussed the needs for machine learning (ML) to be robust, private, and fair. Privacy is contextual so privacy-preserving data sets are problematic. Balancing privacy with useful ML models while securing what genuinely needs to be private is a central challenge. The panelists pointed to a concern that benchmarks/data used in academics are unrealistic; resolutions, sample-set size, and perspectives may not match reality. Simulations might provide better benchmarks, allowing for testing of counterfactuals. The panelists suggested synthetic data can be useful for building robustness, but it is still difficult to train an ML with worst-case guarantees. The panelists concluded with a discussion on the value of realistic data sets.

Aleksander Madry of MIT moderated the panel on Data and Data Sets and introduced the panelists. He then discussed topics related to the data sets, how to create data sets with privacy in mind, simulated data sets, differential privacy and federated learning, and issues related to benchmarks/data sets used in academic settings.

Marzyeh Ghassemi of MIT discussed her research focus on what is called health ML, which is ML in a high-stakes context, where you need to make sure that it's robust, private, and fair. She then talked about how ML is often powered by data, and data unfortunately carries with it the personal and systemic biases that were part of the generative process. In health, we start from problem selection, what kinds of problems we think about or focus on. It goes through team formation, it goes through cohort selection, who is able to participate in studies? What kind of data we're able to collect, algorithm design, whether we do a maximum or whether we remove outliers when we are looking in the learning process. Then in post deployment considerations, thinking about what kinds of evaluations we do, are we looking at average performances? Who has that disadvantage? In many settings in ML, we focus on doing well on an average case. But that doesn't really track well in a health setting because often those who are maybe far from the average are often minority or minoritized populations who already are underserved or disadvantaged. We don't want to propagate or worsen biases that may already exist. There are some socio-technical solutions to this, and a lot of her work focuses more on the potential technical solutions. She highlighted health data sharing and access, and privacy is a concern with data, especially health data. But health data is often de-identified to a HIPAA standard, at least in the United States, and sometimes sold to private companies that are able to use that data to build really great prediction models. She thinks it is a huge disadvantage to not have large data sets that are openly accessible to accredited academics in an equal normalized way.

Tom Goldstein of University of Maryland discussed his work on security for data sets. He talked about how bad actors are able to manipulate data set, what kind of issues in your model can elicit strange behaviors and the ways of protecting against those kinds of attacks. Recently he has been interested in issues of bias and fairness in ML and how they arise from issues related to data set creation. Finally, he is also interested in private data set beliefs and in particular differentially private or other notions of privacy and ways of creating data sets so that institutions can share data sets through different groups. He mentioned that we have this tendency to set academic standards or ethical standards higher, in such a way that we deprive academics from having access to the same data sets when perhaps companies are using them for research. He then mentioned to be careful about what kind of standards we impose on academics.

Emre Kiciman of Microsoft Research discussed how to advance the integration of causal reasoning methods with ML and applying these to problems of robustness in AI generally, but then using these technologies to expand the use of AI and causal methods for decision-making in a variety of critical domains. He then highlighted problems in the area of evaluating causal methods. He then talked about the obvious one which is the challenge of getting data about counterfactuals. Then he talked about how causal reasoning is about understanding maybe what would have happened if you did something versus if you did something else, and we only ever get to observe one outcome and we have to find ways of thinking quantitatively about the comparison between these two observations. He then discussed that for a hard data set that's grounded in the real world, where we only get to observe outcomes following an action that we actually choose. But he thinks there's another interesting challenge in evaluating causal methods and it's one that's related to problems that we're seeing recently. Causal methods are really interesting in that they start to reason formally about some of the assumptions that we make about how the underlying data generating process works. How well that level of abstraction matches the raw data that they have on hand. Finally, he discussed how this could open up very interesting opportunities to think about what algorithms do well under different situations.

Nicolas Papernot of University of Toronto discussed his research spanning a number of areas that have been already mentioned going from security with topics like adversarial examples, but also poisoning of data sets and then finally, the privacy aspects that we've discussed so far. He then gave the definition of robustness in ML. For instance, in health care it is beneficial to have metrics that go beyond the average case in evaluating the model's performance. Then he talked about robustness and in security in general, we have the same need for a worst-case analysis of the performance of the system. Then he discussed, what is missing from a lot of data sets currently is that the data sets really look at an isolated functionality of the system. He then talked about how the model itself is being evaluated directly at its inputs and outputs, but we are not capturing how the model itself is integrated into a system and then deployed in a real-world environment. He then discussed the difficulty of evaluating how much progress we are making because we could very well be getting a perception of progress at the level of the model but when it comes to deploying that model in the realistic production pipeline, then that robustness would not carry any real-world worst-case guarantees with respect to the performance of the system that is deploying itself. Finally, he mentioned benchmarks need to capture that part as well as the system if we really want to evaluate the end-to-end performance of the system.

Panel Discussion: The first question discussed what properties of AI systems that can and should be measured. The panel discussed the data ecosystem - creation, ownership, access and the notion of regulation and what role regulation has played in the data set creation, especially for health care data in term regulation.

Then the discussion was about privacy-producing data sets and regulation, about synthetic data and issues and the importance of data sets for health care. There was discussion about causal reasoning methods and ML.

Then the discussion was on the data ecosystem - creation, ownership, access and so on, the notion of regulation and what role regulation plays in why things are the way they are.

Slack Discussion: One discussion in the slack was on privacy-producing data sets and regulation. Another was a discussion on robustness and security.

Panel 8 - Limitations, Challenges, and Future Directions of Evaluation

[Note: This panel was not recorded, therefore there are gaps in the report for this panel.]

The panelists began by highlighting a central challenge: lack of transparency involved with machine learning (ML) makes it hard to understand why a system made a particular decision. The panelists highlighted the need for decision making tools that ensure trust in the operations of the AI system. Panelists highlighted that the AI systems work in very different ways than humans do and work in very different ways than traditional explicitly coded software does. AI systems are not classifying things the way humans do and that's where we have problems with robustness. The panel discussed the potential for a single general-purpose definition of fairness. Panelists suggested fairness should be defined on a case-by-case basis.

Soheil Feizi of the University of Maryland moderated the panel. After some introductory remarks the moderator asked the following questions: What are the limitations of current AI evaluations? What are the best ways to address challenges? What are the important policy factors that need to be considered?

Eric Horvitz of Microsoft Research responded first that the current AI systems are too brittle, the failure modes are not well described, and the systems are dependent on the context in which they were developed and do not transfer well outside of the training environment. Horvitz observed that a reliance on local evaluations will be necessary along with a continuous monitoring of these systems to ensure conformance with specifications and that measurement needs to extend beyond classical measures which average across test cases and expand to also focus on pockets of failures where there are potentially significant costs. In response to the policy factors question Horvitz emphasized that it is necessary to describe and solve issues around civil liberties and civil rights.

Daniela Rus of MIT replied next that the lack of transparency involved with ML makes it hard to understand why a system made a particular decision. Rus offered that the lack of transparency leads to brittleness and that new tools may be needed to enhance the system in order to develop a better understanding of models. Rus observed that ML is not like software, as a ML program changes in response to the data fed into the training model. Rus identified several challenges including: measuring ML models for safety critical systems, understanding and measuring uncertainty of a model, and how a model is correlated with the data, including any bias found in the data. Rus also suggested strategic use of tooling within the decision-making process to ensure trust in the operation of the model.

Kamalika Chaudhuri of University of California San Diego echoed previous comments by observing that current ML models work in very different ways than humans do and are also different from how software functions. ML models do not classify things the way humans do and this leads to problems with robustness. Chaudhuri then added that there is a lack of specification and there is a need for different benchmarks targeted to different use cases.

Percy Liang of Stanford commented that we should start by thinking about the goals of evaluation: is it to provide feedback and incentives for improving a system or is it to assess the absolute quality of a system? The nature of the goal determines the type of evaluation that is appropriate. Some types of benchmarks encourage direct improvement on a system, whereas some benchmarks encourage the development of general ML techniques and thus have indirect impact on downstream systems. Both are valuable, but one should be intentional about what the goal is.

Chris Meserole of Brookings began by asking the question “What are the evaluation and measurement goals that would be more useful for policy makers?” Meserole continued that there are a broad variety of evaluation metrics that should be examined, including the ability to view a black-box sample, evaluate a model for heterogeneous effects, and measure the performance in real-world settings. Along with the increasing size of models is an increase in the state space of the model and a metric that is scale invariant is needed.

Panel Discussion: The moderator then commented that there are a lot of assumptions as part of constructing a system: assume there is a cost function, a reward function, etc. These systems are then deployed into an open world and are observed as being fragile. How can this gap be closed? The discussion emphasized caution and that there was likely no perfect solution and that robustness may be difficult to achieve. Suggested solutions included randomized control trials, sensitivity studies, and careful use of benchmarks.

The next question inquired about how to perform an evaluation of the interpretability of a system. A panelist offered that a user study could be conducted to show decisions made by the model and determine whether the user understands why that decision was made, but that ultimately the method for evaluating interpretability is still an active area of research. Another panelist discussed that interpretability has to be about the principles under which a system is constructed and to understand how to build large systems of components that individually make sense. A final thought on interpretability was that interpretability only needs to match the legal regimes and that a full granular understanding of the algorithms was not completely necessary.

The next discussion area touched upon fairness and developing a definition of fairness which prompted the question of whether there is a universal definition of fairness. The response was no, but a reasonable definition is needed to make progress on this issue. Another panelist offered that fairness should be process-oriented with the definition of fairness used being relevant to the situation for where the system is applied.

Slack Discussion: The slack discussions touched on the point that robustness claims can be hard for users to interpret as well and how can users be helped in making sense of them. Another discussion mentioned that benchmarks may not be testing for the desired outcomes and may themselves be flawed.

Panel 9 - Measuring Concepts that Are Complex, Contextual, and Abstract

The panelists highlighted a challenge in the AI system measurement space: unknown-unknowns (the things we do not know, that we do not know) in ML cause cascading problems in the models. Resolving false positives and false negatives is difficult in many cases. Measuring the "understanding" of an AI/ML system presents challenges and opportunities. Understanding is the interpretation of a situation including context, insight and foresight. Humans expend enormous effort to build and manage shared understanding. That effort is a continuous, interactive process. Humans can discuss and maintain a shared context of understanding, repairing and aligning a shared understanding frequently in conversation. Existing models often simply classify, and cannot expose a model that humans can understand. The problem for measuring useful understanding is expensive, but not impossible to scale. Panelists suggested there is a lot to learn from embracing uncertainty, controlling internal validity less, and relying less on statistical conclusions. Instead, panelists suggested considering external validity and how results from experiments could generalize.

Ellen Voorhees of NIST moderated, stated the panel topic as the measurement of concepts that are complex, contextual, and abstract, and introduced each panel member, in turn, directly before each of their presentations.

Lora Aroyo of Google Research (NYC) gave a short position statement on uncovering unknown unknowns in machine learning (ML), where unknown unknowns are defined as high-confidence mistakes. The speaker explained that unknown unknowns are important in practice, in part since these errors can cascade downstream in a system, and that there are approaches to addressing them, e.g., with automated approaches and human reporting, as well as with crowd-sourcing. She described a project titled "Crowdsourcing Adverse Test Sets (CATS) 4 ML", which set out to scale human efforts for detecting unknown unknowns, and showed some examples of unknown unknowns that were detected and described some of the challenges discovered.

David Ferrucci of Elemental Cognition described his work in machine language understanding, defining understanding as the interpretation of a particular situation in order to provide the context, insight, and foresight required for effective human decision-making. The speaker noted that understanding is hard for both humans and machines because it is complex, contextual, and abstract, that it sometimes fails, and that it may be an interactive process that needs to take humans into account. The speaker then gave examples of errors machines made when performing various tasks that demonstrate the challenges of machine understanding: even machines that perform well on current benchmarks. The speaker argued that current benchmarks inadequately assess understanding, and presented an approach to defining and evaluating understanding grounded in the content needed to perform tasks. The speaker described an experiment that looked at machine performance on tasks that required understanding spatial, temporal, causal, and motivational content. The speaker ended by noting the important but challenging nature of measuring understanding, both in the context of humans and machines, and that large investments are potentially necessary to address the inherent challenges.

Ben Carterette of Spotify described his efforts to measure delight, which he emphasized as challenging, even to put into words, and contrasted measuring delight with simpler measures of positive experiences, for example clicking “thumbs up”, which users often do not interact with, and more indirect measures, such as (not) skipping past the media they are engaging in. The speaker noted that delight can be contextual, varying for people, e.g., across cultures, and individuals, e.g., by mood, and then posed the question of measuring delight or engagement within the framework of experimental validity (broken out into statistical validity, internal validity, external validity, and construct validity). The speaker ended with the position statement that a lot could be gained from embracing uncertainty, exerting less control of internal validity, relying less on statistical conclusions, and thinking more about external validity and how results from experiments can generalize; measures used are often distant proxies for what is actually of interest, and rather than tighten up those measures, it is worth instead trying to model what is unknown about how to approximate what is of interest (despite there being more uncertainty and it perhaps no longer being possible to determine winners of leaderboards).

Panel Discussion: The moderator then asked, “What is your approach to abstracting away from a task in order to make an evaluable task?”. The discussion began with acknowledging that it is a difficult challenge and the suggestion of starting with concrete examples and seeing how things can generalize. It was then suggested that considering how humans naturally perform tasks and including in evaluation data sets data points where humans disagree could be fruitful, drawing in to question the reasonableness of assuming there is a single gold standard/ground truth. An additional suggestion is to leverage previous efforts at task abstraction to understand what are the levers that can be pulled.

The moderator agreed that, by definition of a complex task, it is necessary to consider data where humans disagree, noted that this can happen in surprisingly simple instances, and shared that this is something acknowledged by the information retrieval community and dealt with by choosing a single person’s response, and that this approach has not been universally accepted (including, for example, by the natural language processing community). She then asked, how do you deal with the challenge of not having a single right answer when running an evaluation. The panel responded by noting a tension between doing science and publishing results vs effectively addressing practical needs, particularly when someone is required to take responsibility for the decisions made by an AI, the value and inherent challenge of considering distributional (rather than binary) truth (both in system development and evaluation), and that uncertainty modeling as well as suites of metrics might be a good approach to addressing these challenges.

The moderator continued by asking two questions from the audience: how often do offline evaluations generalize to online evaluations at Spotify? (Ben Carterette responded that it's variable), and in the CATS4ML challenge do you ask participants for a reason why they think a datapoint will be difficult? (Lora Aroyo responded yes, and that this information was helpful.)

The panel concluded with the panelists each emphasizing the importance of the challenges being addressed and their appreciation for the efforts being made to address them.

Slack Discussion: There was some discussion regarding the challenges of crowdsourcing and the potential introduction of additional bias, as well as the challenges of dealing with inter-annotator disagreement and multiple correct answers when labeling ground truth.

Panel 10 - Measuring with Humans in the Mix

The panelists suggested a trustworthy AI system needs to be fair, easy to understand, that it's not being compromised, and it needs to be able to communicate how certain it is. Panelists discussed the understanding and measurement of the societal effect of AI systems. Current issues include current jargon and metaphors. The presentation of trust to different people might be very different and how the system explains trust will be different depending on the audience. The panelists suggested that explanations could be viewed as an education effort -- trying to educate on a particular thing at a particular time.

Margaret Burnett of Oregon State University, panel moderator, introduced the theme of the panel: how AI meets actual humans, including who do you measure, what do you measure, and where do you measure. She introduced each panelist in turn directly before their introductory remarks. After the opening remarks by the panelists, the moderator introduced herself, and noted the discrepancies between the demographics of AI developers (predominantly white or asian men, all college educated and well above the socio-economic median) vs US AI consumers (who are gender balanced, roughly 50% college educated, and are much more racially and economically diverse), and that the needs of many US AI consumers are not being met. She concluded by noting that different people have different learning and information processing styles and attitudes towards risks, and it is important for AI to be able to support the range of approaches and needs, and not just those that are indicative of the AI developers.

Madeleine Clare Elish of Google began by sharing a story of a specialized nurse at the Duke hospital ICU, who the panelist shadowed in order to study how an AI driven system, Sepsis Watch, actually worked in practice. The panelist noted that sepsis is the leading cause of death in hospitals and that it is notoriously difficult to diagnose and treat quickly, and then described how the Sepsis Watch system worked: that the deep learning model reported the patient’s risk of sepsis to the nurse, who relayed that information to a doctor, who then instructed the nurse on how to care for the patient. The panelist emphasized that Sepsis Watch is an example of a socio-technical system, which means that the technical component of the system cannot effectively be evaluated outside its social and organizational context, which is needed to measure the actual impact the system has on patient care outcomes.

Rachel Bellamy of IBM started by asking the question, when measuring AI with humans in the mix, who are the humans in the mix? The panelist answered that clearly the humans are end users interacting with the AI systems, e.g., the people at a bank when someone applies for a loan where the decision is made by an AI system, the hiring or admissions committee when people apply for a job or college admission where the applications are filtered by an AI system.

The panelist continued that of course the people affected, i.e., those applying for the loan, job, or college admission, are also in the mix, as is society at large. She described some of the failures of AI systems with respect to humans, e.g., gender bias in recruitment software, along with others, and highlighted the need to measure and evaluate the societal effects of AI systems. The panelist continued by describing work at IBM on trustworthy AI, which consists of asking whether the AI is fair, whether it is easy to understand, whether it has not been tampered with, and whether it is certain, noting the challenge of measuring each of these. She concluded by emphasizing the importance of involving people who understand the important contextual and societal information relevant to the AI system at the beginning and throughout the development of the system.

Robert Hoffman of IHMC started by asserting that several of the issues that came up in day 3 of this workshop form clusters of psychometric issues, including: (i) questions about parametric significance testing (noting that the notion of *practical* significance is a little over 100 years old, though work describing how to measure practical significance is only being done now), (ii) the distinction between a measure and a metric, (iii) the need for operational definitions, (iv) conceptual issues and metaphors (e.g., a machine cannot be a “team member”, but is instead a tool), and (v) cognitive issues involving explanation and sense making. He emphasized that you can measure human performance using known measures and machine performance using known measures, but this alone is insufficient–instead the measurement needs to be done at the work system level, and it must include the human-machine interrelationship.

Panel Discussion: The discussion began with an audience member stating that viewing machines as team members is an example of anthropomorphism, which is common in AI and can lead to over trust. Robert Hoffman responded in agreement, sharing examples of machine “attention” and “perception”, noting that machines cannot “pay attention” nor “see” things. and that they lack concepts (e.g., confusing chopsticks with a hat rack because they have no concept of what chopsticks are). The AI makes mistakes no human would make.

The next question from the audience was whether there are challenges to integrating AI systems into human processes, e.g., as described in the Sepsis Watch example. Madeleine Clare Elish responded that it is mistaken to believe that any AI system is not integrated into a human process, and emphasized the importance of language, noting that describing the deployment of an AI system gives the wrong impression (that they can be dropped in without consideration to context), and instead we ought to talk about the integration of an AI into a system (which implies the need to consider context). Rachel Bellamy added that there are huge benefits to integrating AI, since humans are limited in ways that machines are not. Robert Hoffman added that in the introduction of new technologies, it is always the case that new forms of error, roles, structures, and challenges are also introduced, and that this is inevitable.

The next question asked whether the level of trust in an AI system ought to be affected by who does the trusting. Rachel Bellamy responded that how the factors affecting trust are presented to the person might need to be very different for different people. The moderator noted that it is also important to consider what is being trusted, e.g., if it’s a bank manager trying to decide on a loan application vs the person applying for the loan, the trust questions will be different. Madeleine Clare Elish added that trust is not quantitative, it is a relationship. The panelists and moderator noted that trust is a complex and situational concept and emphasized that the goal is not to increase trust, but instead to accurately calibrate trust in AI.

The next question was about progress in explanations provided by AI systems, and it was noted that explanations are not things in themselves, but exist with respect to the person receiving it, and that there may be some value in viewing explanations through the lens of education.

The next question was about measuring human acceptance of AI. Madeleine Clare Elish responded that ongoing and iterative engagement with stakeholders throughout the AI development process is key for acceptance. Rachel Bellamy noted the existence of work on measuring socio-technical systems, and how this work applies to AI, but that there are a few additional challenges for AI.

The discussion continued with a question about qualitative vs quantitative methods. Robert Hoffman responded that the distinction between these is historical and misguided and what are considered qualitative methods are going to be essential. Rachel Bellamy shared the importance of interacting with real users, and Madeleine Clare Elish shared that measurement is both necessary as well as problematic, and that it’s important to hold space for uncertainty and not quantify too soon. The moderator added that qualitative methods are useful for finding unknown unknowns (you don’t have to know the question, which is needed for qualitative methods) and that qualitative methods provide more information at the expense of not being as good for comparison.

The panel concluded with the question: are there challenges in helping developers understand that there are societal implications for the work that they are doing? Rachel Bellamy responded that there are many metrics, but knowing which metrics to use requires knowledge of the domain and context, and that developers need to be open to working with people with that knowledge. Robert Hoffman added that conferences aren’t enough, there needs to be a (funded) task force.

Slack Discussion: There were several comments in the slack about the role humans should play in AI systems and the role trust and context play.

Panel 11 - Software Infrastructure Overview, Existing Tools and Future Desires

The panelists discussed the need for new techniques to be developed to improve adversarial robustness of AI models. There has been a deep learning revolution, that is ushering in a new AI revolution, but one which we are not prepared for in terms of security and adversarial attacks in particular. The panelists discussed the tradeoff between accuracy and robustness of models in light of robustness. Models are more sensitive, and thus less robust. If you measure and evaluate based on accuracy alone, the models will not be robust enough. The panelists highlighted a potential approach to explainable AI benefits through concepts such as “Model Cards” that represent the abilities of a model in an easily digestible format.
The panelists discussed the desire to go beyond simple benchmarks in AI evaluation, going beyond a simple quantitative measure. There is an appetite for having open APIs and tools that can help with different kinds of evaluation. Lots of small, diverse measurements are the way to go forward.

Harold Booth of NIST moderated the panel and began by describing the panel topics covering software infrastructure, existing tools as well as future desires, and the landscape, challenges, and needs of tools and infrastructure for the purpose of measuring, testing, and evaluating AI systems. He introduced each panelist in turn directly before their introductory remarks.

Pin-Yu Chen of IBM presented on the topic of holistic adversarial robustness of AI models. The speaker noted the excitement around using deep learning in various settings, as well as the new sources of vulnerability it introduces to the systems using it, for which a majority of organizations are unprepared. The speaker showed several adversarial examples of images and noted that this type of “Achilles heel” for AI is present in a variety of modalities, and explained that improvements in accuracy do not correspond to increased adversarial robustness. The panelist continued by emphasizing the importance of adversarial robustness and its real-world impact. He then gave a view of adversarial robustness from the perspective of the model life cycle and offered an appeal toward a holistic view of adversarial robustness, likening the attack and defense of the model to the bug finding and fixing process that many software engineers are more familiar with, and suggesting that lessons learned from autonomous vehicles could be useful. The panelist closed his opening remarks with his view of how the challenges can be approached at multiple levels (distributional shifts, single threat models, multi threat models, and global robustness).

Harsha Nori of Microsoft began by describing InterpretML, a tool for training intrinsically interpretable models as well as for explaining models that are not intrinsically interpretable. He then began to describe the challenge of choosing among the metrics for accuracy, and noting that there are orthogonal concepts that people also wish to measure (e.g., fairness) that each have their own challenges and complex interrelationships, some of which are incompatible with one another, and that a dashboard with metrics describing each is insufficient for a holistic view of a model. The speaker then expressed his view that interpretability can enhance model evaluation in a way that is metric agnostic, and then walked through a case study identifying which edits to wikipedia were malicious, showing that an intelligible model was useful for this circumstance.

David Pitman of Google described what Google is doing for explainable AI and how it relates to evaluation, particularly how to develop tools that allow non-experts in explainability to utilize model explanations to evaluate their models. These range from tools for model developers to debug their models to tools that are meant to be easily digestible in order for less technical people to gain some information about the model, including where it performs well and where it performs less well. The speaker emphasized the need to view these things through the lens of responsible AI and that having a range of tools for this is useful, since a programmatic way of looking at things might, for example, be able show how inflection points in the data have impact, but this is a local view, versus a more holistic view of how the model is performing overall.

Panel Discussion: The panel discussion began with the moderator asking: what methodologies are used to build tools that enable developers to improve their models? David Pitman began by describing how he decides which tools to build, how to assess the tools, and then described how to make tools as useful as possible. Pin-Yu Chen agreed and added that end users have overly high expectations of AI models, in part due to misconceptions on the meaning of the system measurements, and how users can better calibrate their trust in AI systems. Harsha Nori also agreed and added that tool developers have a responsibility to encode best practices into their tools in order for users to get the correct interpretation of the tools output.

The moderator then asked: what capabilities need to be added to evaluation tools in order to measure beyond just benchmarks? Harsha Nori responded that benchmarks are restricted by the ability to define metrics. It will be necessary to keep advancing metrics, and something that is helpful about benchmarks (but is not necessarily limited to benchmarks) is that they make tools easy to compare with one another, and interoperability and standards (and openness generally) can help facilitate that for non-benchmark style tools. Pin-Yu Chen added that a lot of progress has been driven by not only reporting results on benchmarks but also reporting measures of robustness, e.g., using perturbation tests. David Pitman responded that we’re at a watershed moment in our ability to evaluate models, and one possibility is to continue to seek to represent performance using a single value (e.g., p-value or area under receiver operator characteristic (ROC) curve), however these systems have become so complex that this is no longer a sufficient way to evaluate them and it would be nice to see tools develop that go beyond a single quantitative measure.

The moderator then asked what sorts of standards are needed for these tools, what is ready to be standardized today, and what needs to wait for the future? David Pitman expressed his positive view of model cards then added the importance of the standards being able to describe and compare a wide variety of different types of models and of being able to communicate the evaluation results to a wide variety of non-technical stakeholders. He added that there does not yet seem to be much consensus and that NIST working with industry to build such a consensus would be valuable to all. Harsha Nori agreed and added that it’s hard to standardize this early in the game, so a pursuit of standards would need to be flexible enough to accommodate the changes that are sure to come, and it’s worth thinking about what we want to know about a model that is not contained in the model itself so that information can be captured. Pin-Yu Chen agreed and added that if standardization is too difficult at the moment there may still be an opportunity to benefit from some formalization of guidelines.

The discussion turned to tools for probably approximately correct (PAC) bounds, how to find out more about model cards, the relationship between explainability, privacy, and model complexity, and how to reconcile ethical principles, single number quantification types of evaluation, and more qualitative evaluations.

The panel concluded with a discussion of the challenges of having evaluations be inductive (e.g., that results in one area apply to others), and how it’s important not to lose sight that the objective of evaluation is to understand a model, and the importance of considering multi-objective evaluations.

Slack Discussion: There were several threads on the challenges and importance of robustness and interpretability, as well as the connections and analogs between model testing and software testing.

Panel 12 - Practical Considerations and Best Practices for Measurement and Evaluation

Panelists discussed the notion of having a system which is tested initially for performance, but then can move through the steps of evaluation. AI is not just an algorithm, but is part of an embedded system. Panelists highlighted a central challenge that data can be hostile. There can be threats within the data. Another key challenge is having access to a sufficient amount of data in the test environment to give an accurate picture of live data. There is a need for a process to keep the model updated. There is a need for a human baseline. Panelists discussed the need for a baseline, and suggested with one, it's difficult to make an argument on how much time and money are saved through AI systems.

William "Bill" Streilein of MIT Lincoln Laboratory moderated and introduced the panel. The panel was described as focusing on the practical considerations needed to manage an AI system and mechanisms and constraints when performing test and evaluation of such systems.

Matt Gaston of SEI Emerging Technology Center, Carnegie Mellon University (CMU) began by introducing the idea of the creation and definition of an AI Engineering discipline that would be engaged with the deployment of AI systems as reliably and responsibly as possible. SEI is focusing on three areas of AI Engineering: scalable AI, robust and secure AI, and human centered AI.

Sven Krasser of CrowdStrike began by describing the goal of CrowdStrike is to stop breaches and they use AI models in the service of that goal. Some measurements of CrowdStrike's performance are the result of third-party measurement regimes. Some of the challenges with this particular task include the adversarial nature of the task, and the idea of concept drift as new attacks, threats and techniques are developed. Identifying what makes sense to measure as part of the development of a good model is challenging, and some measurements may also extend to the modeling process itself to ensure that the process is designed to produce a good outcome. Some questions that need to be answered and that measurement can help with include: what to do with unknown data, consistency of the model, and comparability of solutions.

Sanjeev Mohindra of MIT Lincoln Laboratory spoke to AI assessment as an evaluation of AI systems which are composed of data, algorithms, computing, and the humans who will use the AI. Framing as a system forces thinking about the lifecycle of a system and the need to perform assessment before and after deployment, and, looking forward, the continuous monitoring and assessment of AI systems will become more important over time. Thinking about performance with the inclusion of the human aspect of the system is important since although an AI model may be fooled by an adversarial input it is not necessary that the whole system, which includes the human users, be fooled. Items highlighted as data concerns included sufficiency, fairness and bias, and areas of assessment for models included not only accuracy, but also robustness and resilience. Another question is how does the AI communicate to a human team member to ensure the system operates appropriately.

Jane Pinelis of Test and Evaluation of AI/ML at DoD Joint Artificial Intelligence Center (JAIC) highlighted that testing of AI brings new challenges such as dependence on data quality, representativeness of the data, and new failure modes. When systems are tested by the JAIC they evaluate properties such as operational effectiveness, integrity, robustness, resiliency, adherence to DoD AI ethical principles, and human system integration. Historically, testing of non-AI enabled systems benefited from leveraging existing best practices from academia and industry, but with the rapid development of AI technologies, the need to push forward the science, data knowledge, workforce, skills and infrastructure has surfaced. In response the JAIC has developed framework for test and evaluation of AI enabled systems with four main components: test of the AI model itself (e.g., accuracy, precision, recall), system integration (e.g., reliability, functionality, interoperability, compatibility, security) , human-system integration (e.g., trustworthiness, explainability, mental models), and operational tests (e.g. mission accomplishment).

Richard Tatum a Civilian at the US Navy, Naval Surface Warfare Center, Panama City, Florida is interested primarily in trust and developing an understanding where systems will work properly and where they will break down. When performing verification of systems as a monolithic system, testing of systems was more difficult and would have trouble passing tests when deployed into a stochastic test environment. By breaking down these systems into more atomic capabilities they wanted to simplify testing, but then needed to establish total system performance based on testing of these atomic capabilities and determine if these atomic tests are a good measure of system performance. Some of the approaches to metrics included performing a gap analysis (identifying what is missing), or comparing performance to other systems.

Panel Discussion: The moderator began the discussion with the question of what are some of the practical aspects relevant to amassing data needed to determine operational performance, as well as other data such as adversarial examples, testing for robustness, and other properties. The ensuing discussion mentioned testing only on operationally relevant data, reserve operationally relevant data that is not representative to what the system was trained on to determine out-of-domain performance, identify natural perturbations in the data such as cloud cover and dirt on the camera, measure and evaluate the data not just the system itself, track and evaluate data sets as they evolve over time as the result of augmentations, provenance of the data as it is used as input into model creation, understand the distribution of the data and evolve models to understand the tails if they are events important to the performance of the system, understanding what is in the corpus vs what is in the field, be wary of adversarial data, be aware of cases where not enough meaningful data is available to perform proper testing, synthetic data can be a challenge and is not necessarily an answer, and sensitivity to initial conditions.

The next question asked what role synthetic data has in the test and evaluation of AI systems. The discussion acknowledged the tension with having enough data for training and reserving data for test, and that synthetic data can have a role in providing an additional source of data that is operationally relevant and that one possible use would be to blend synthetic transforms with existing data sets instead of generating synthetic data from a model. An important aspect of the use of synthetic data is understanding where in the process the synthetic data is used and being aware of any legal or other constraints.

The next question asked how do you know a model is worth pursuing or evaluating to deployment, what are the metrics that are most relevant in making that decision? The responses emphasized to first understand the task to be solved and then the first question to be answered is whether the solution needs to be machine learning, and once that is answered in the affirmative other metrics of interest would be cost of inference, does the model cover new areas or fortify weak areas, what is the increment of true positives by deploying the model, what is the cost to remediate (addressing false positive/negatives), and what is the retraining complexity (difficulty of ensuring the next model is an improvement). Performing a risk assessment of the model, understanding the assumptions of the model, risk of deploying vs not deploying, cost to maintain, and value to the organization of deploying.

The next question asked what is the role of the human in the human-machine teaming aspect of the system, and at what point should the human element be brought into the test and evaluation process? Establish a baseline without the solution and then use that baseline as a measure against which the system is measured. Additionally, an understanding of the human workflows and how they will change once a new system is deployed is an important part of the process of determining whether the deployment of a particular solution is “worth it,” and “worth it” may be range from the system can perform slightly worse to the system must perform significantly better than the existing process. An important part of analysis is to ensure that the touch points into the system are meaningful and provide sufficient insight into the operation of the model and that the human is able to observe, orient, decide, and act appropriately using the system to accomplish a given task. Another concern regarded how to evolve a system over its lifecycle as the objectives for a system change and that a possible solution is to monitor the system performance over time and evaluate whether the performance degrades as the situation for which the system was initially applied changes. A final point raised was having a clear path for incorporating human feedback into the behavior of the model.

A final question asked what are the measurements that can be used to recognize whether a system performs in an ethical manner and determine if a system possesses attributes such as being reliable, governable, traceable, equitable, and responsible. The response suggested viewing responsible AI as a source of requirements and many of these requirements may already be operational requirements, and develop metrics which address these requirements. A key aspect is to help users of the system make good decisions on whether and how to use a system, through training and documentation, and providing transparency into what is going on within a model.

Slack Discussion: A major note of agreement in the thread discussion was noting that not everything should be implemented using an AI and that simpler solutions should be preferred when those solutions are able to accomplish the same tasks.

Keynote Summaries

In this section we summarize the three keynote presentations of the workshop.

Keynote 1 - Jason Matheny (Deputy Assistant to the President for Technology and National Security; Deputy Director for National Security in the White House Office of Science and Technology Policy; and Coordinator for Technology and National Security at the National Security Council)

The speaker highlighted the National Security Council's report on AI and its recommendations for NIST as they relate to AI standards and best practices. The speaker noted that as a society we spend so much time and money on back-office applications, and as such it is a prime opportunity for AI. Also, in cybersecurity where defense can be expedited by AI. The speaker highlighted the need for AI systems to be trustworthy, explainable, robust, as well as maintain privacy. Lastly, the speaker noted the importance of participating in international standards as a way to get ahead of technology governance.

Jason Matheny: The keynote speaker mentioned that our progress in AI has been a function of how we measure performance and how we set standards for performance. The speaker referenced the website ai.gov that has resources, as well as the National Security Council's 2021 Final Report. (See Appendix for the full URLs of these references).

The keynote speaker sampled from some of the key recommendations of the NSC AI Report. The first recommendation was that NIST should provide and regularly refresh a set of standards, performance metrics and tools for qualified confidence in AI models, data, and training environments and predicted outcomes. The second was that NIST's testbed program for AI should be expanded to accelerate the development and adoption of interoperable, secure and reliable AI technologies as well as to generate data for AI enabled cyber defenses in differing IT infrastructure environments. The third recommendation was that NIST should provide and regularly refresh a set of standards and tools over time as the science of how to test systems across responsible AI evolves. The fourth recommendation was to establish third-party testing centers to allow independent third-party testing of national security related AI systems that could impact US persons.

Next, the keynote speaker talked about the challenges to measurement that the speaker thinks we are facing or will face in the future, some of which intersect with security issues. The speaker also talked about which areas measurement will be most important. One application is the back-office process areas such as finance, contracts, and logistics. In those applications, the security is of lower risk. Cybersecurity is another important application of AI. Another is the application of AI to accelerate Science and Engineering. Some require new metrics.

The keynote discussed challenges in AI. The first of these challenges is causality, understanding how events are caused. The second of these challenges is explainability, i.e., explaining to a human why it generated the result that it did. The third challenge is handling rare and extreme events, as many ML systems are optimized for the average case. The fourth challenge is robustness to spoofing and adversarial attacks. The fifth challenge is privacy protection.

The speaker ended with some observations on challenges in building international standards.

Keynote Q&A: The first question was how to see that the human and AI teaming can be optimized for extreme cases. Humans sometimes have scope neglect (compressing together small probabilities or neglecting high-consequence low-probability events). It would be nice to train models better in those cases to assist us.

There was a question about the speaker's meaning for the term "testbed". It is all of these: the data, the software, the platform, and the experts that can interpret the system. The Government operates on test ranges for other technologies.

The next question was about the current work being done to address the challenges the speaker mentioned as well as recommendations for current and future directions for AI testbed and evaluation. The speaker mentioned that we should be building the infrastructure and centers of excellence with the skills, data sets, and measurement tools. The other place we need work is benchmarking. He likes the AI index, and wants to track the performance of different AI on different applications in different places in the world

Slack Discussion: There was some discussion during the keynote with some references for causality. (All slack references are included in the reference list compiled in the Appendix, which include these references.)

Keynote 2 - Fei-Fei Li (Sequoia Professor, Stanford University; Co-Director of Stanford's Human-Centered AI Institute)

The speaker highlighted ImageNet as a benchmark data set in the field of computer vision. It helps to see how well algorithms do image categorization. As time goes on, the speaker noted, programs need to evolve to keep pace with a changing landscape; there are new needs, new techniques. The speaker cautioned for vigilance against historical bias in society and be aware that bias is present in the entire pipeline of the AI system.

The following is a summary of the Q&A session between Elham Tabassi (Chief of Staff, Information Technology Laboratory, NIST) and Fei-Fei Li.

Fei-Fei Li provided some historical background information about the genesis of ImageNet. Li began with a quote from Einstein: "The formulation of a problem is far more often essential than its solution…" and the idea of what were the important questions in the field of computer vision and AI. The driving force of ImageNet was then to answer the question of object recognition. The goal was to create a benchmarking data set that could enable advances toward the development of object recognition capabilities.

Reflecting on what could be done differently based on the experience of assembling ImageNet, examples of things which could be improved were to push on the dimension of contextualizing objects in a more cluttered setting, improved handling of privacy concerns (e.g. blurring of faces), and to better manage bias in the data set. There is now more awareness of how to curate data and assign labels to mitigate bias. The discussion then covered steps for mitigating bias which include recognizing historical bias, understanding the source of the data, and paying attention to different attributes of the data where bias could surface. Even with these efforts bias may still be present, therefore transparency in the training pipeline and ensuring humans are still responsible for decision making.

Slack Discussion: The primary thread picked up in the slack discussion focused on bias and how to manage bias in order to provide for fairness, based upon some predetermined criteria of what fairness meant for the particular application context.

Keynote 3 - Lynne Parker (Director, National AI Initiative Office, White House Office of Science and Technology Policy)

The speaker highlighted that government agencies are encouraged to continue to use AI when appropriate to benefit the American people. Principles have been published in the form of an executive order to establish common requirements for the entirety of the civilian government. These principles emphasize that AI use must be reliable, safe and secure, regularly monitored, transparent, and accountable. The speaker discussed the plethora of tools and techniques in AI but a lack of guidance on how to use them. One challenge on human AI teaming is making sure humans have the appropriate level of trust of the AI system.

Lynne Parker: The speaker began by stating three key goals the National AI Initiative Office seeks to advance: 1) for the US to continue to lead in AI research & development, 2) for the US to lead the development and deployment of trustworthy AI, and 3) to train the current & future US workforce for integration AI technology. The talk focused on a sub-part of the second goal, namely for the US to lead the development and deployment of trustworthy AI in the public sector, and posed the key question: what needs to happen for AI to be confidently deployed in the public sector?

In response to this question, the speaker described three things that need to happen. First, there’s a need for agreed upon fundamental principles that guide the use of AI in the public sector, which she shared does exist for the Federal Government, in the form of an Executive Order (EO-13960), “Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government”, which establishes common requirements for the entirety of the civilian federal Government. These principles emphasize that the use of AI in the public sector must be: lawful, purposeful and performance-driven, accurate, reliable and effective, safe and secure, understandable, responsible and traceable, regularly monitored, transparent, and accountable.

The speaker stated that while this is an important first step, principles are not enough, and the second thing that needs to happen is to determine how to turn these principles into practice and assess whether the principles are in fact being upheld. She explained that because the Federal Government is large and diverse, there is a need for best practices that can be used by each agency to ensure that AI is adopted in a manner that works effectively and consistently delivers services to the American people and to foster public trust in AI. The speaker then highlighted and described a proposed framework from the Organization for Economic Co-operation and Development (OECD) network of experts on AI, which categorizes tools as technical, procedural, or educational, and structures information about each tool (e.g., description, origin, category, scope, target users, policy area, alignment with principles, adoption potential, and expected benefits). This breakdown allows comparisons of tools within and according to their category, and qualitative and quantitative assessments of tools are important future work.

The third thing that needs to happen for AI to be confidently deployed in the public sector is to turn best practices into formal policy guidance for federal agencies. The speaker emphasized that everyone working toward moving from best practices to policy guidance must ground their approach in a broader awareness of what AI can and cannot do. For example, a policy guidance statement that “tools must be explainable” is almost vacuous. The federal Government has prioritized AI R&D to support progress toward policy guidance. The technical community can play a key role in helping to inform non-technical policymakers about the evolving state of the field and by contributing some time and expertise to the federal government to help build up federal expertise in this area, maybe even by survey for a while in an official capacity in Government.

Keynote Q&A: The moderator asked a question from the audience, noting recent EU regulations on AI and inquiring about the plan for the US to develop similar policy and guidelines beyond Government applications. The speaker responded that last year a memo was issued regarding the guidance for regulation of AI in private sector, which laid out requirements for agencies to produce their own plans for looking at, among other things, the use cases of AI that their regulatory authorities need to pay regular attention to and then also previewing what some of those plans might be; expect more information on this from various agencies to become public, perhaps as soon as this summer.

The moderator then asked, what are the requirements and challenges for human-AI collaborations, and what sorts of measurement and evaluation do you see is needed for those requirements? The speaker responded that one of the key challenges to human-AI teaming is for humans to give the appropriate level of trust to the AI systems (and not over- or under-trust the system), so measurements of trustworthiness or proxies, such as human mental models or the reliability of the human-AI teams performance, as well as how these things evolve over time, are key challenges.

Slack Discussion: A few questions were posed in the general slack channel, including about US plans to develop policy and guidelines similar to those that were recently introduced by the EU, research in over and under trusting AI, and how to tell whether an autonomous system’s goals are shared by society.

Appendix A: References from Panelists and Participants

Throughout the workshop, references were mentioned. The majority of these are from the slack, and there are some references that were mentioned in the panels and keynotes. This list of references is a compilation provided in alphabetical order.

"Adversarial Robustness 360." n.d. Accessed June 27, 2021a. https://art360.mybluemix.net/.

"AI.Mil - Accelerating DoD's Adoption & Integration of AI." n.d. Accessed June 24, 2021. https://www.ai.mil/.

Arnold, M., R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, et al. 2019. "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity." IBM Journal of Research and Development 63 (4/5): 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288.

"Attack Discrimination with Smarter Machine Learning." n.d. Accessed June 25, 2021. https://research.google.com/bigpicture/attacking-discrimination-in-ml/.

Berner, Julius, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. 2021. "The Modern Mathematics of Deep Learning." ArXiv:2105.04026 [Cs, Stat], May. http://arxiv.org/abs/2105.04026.

Blodgett, Su Lin, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." In , 5454–76. https://doi.org/10.18653/v1/2020.acl-main.485.

Bouchard, Andrew, and Richard Tatum. 2015. "Verification of Autonomous Systems: Challenges of the Present and Areas for Exploration." In OCEANS 2015 - MTS/IEEE Washington, 1–10. https://doi.org/10.23919/OCEANS.2015.7404475.

Bowman, Samuel R., and George E. Dahl. 2021. "What Will It Take to Fix Benchmarking in Natural Language Understanding?" ArXiv:2104.02145 [Cs], April. http://arxiv.org/abs/2104.02145.

Brewer, Marilynn B., and William D. Crano. 2014. “Research Design and Issues of Validity.” In Handbook of Research Methods in Social and Personality Psychology, edited by Charles M. Judd and Harry T. Reis, 2nd ed., 11–26. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.005.

Church, Kenneth Ward. 2020. “Benchmarks and Goals.” Natural Language Engineering 26 (5): 579–92. https://doi.org/10.1017/S1351324920000418.

Church, Kenneth Ward. (2020) 2021. Kwchurch/Benchmarking_past_present_future. https://github.com/kwchurch/Benchmarking_past_present_future/blob/5524d7687416c76a37789e21dd8e94ee903129e1/README.md.

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. “Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies.” SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3551505.

“Federal Register :: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.” n.d. Executive Order 13960. Accessed June 27, 2021. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. “Datasheets for Datasets.” ArXiv:1803.09010 [Cs], March. http://arxiv.org/abs/1803.09010.

“GitHub - Interpretml/Interpret: Fit Interpretable Models. Explain Blackbox Machine Learning.” n.d. Accessed June 27, 2021. https://github.com/interpretml/interpret.

Gleave, Adam, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. 2021. “Adversarial Policies: Attacking Deep Reinforcement Learning.” ArXiv:1905.10615 [Cs, Stat], January. http://arxiv.org/abs/1905.10615.

“Google Cloud Model Cards.” n.d. Accessed June 27, 2021. https://modelcards.withgoogle.com/about.

“Google Cloud Model Cards Face Detection.” n.d. Accessed June 27, 2021. https://modelcards.withgoogle.com/face-detection.

Guedj, Benjamin. 2019. “A Primer on PAC-Bayesian Learning.” ArXiv:1901.05353 [Cs, Stat], May. http://arxiv.org/abs/1901.05353.

Gunning, David, and David Aha. 2019. “DARPA’s Explainable Artificial Intelligence (XAI) Program.” AI Magazine 40 (2): 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Hanzely, Filip, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. 2020. “Lower Bounds and Optimal Algorithms for Personalized Federated Learning.” In Advances in Neural Information Processing Systems, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:2304–15. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/187acf7982f3c169b3075132380986e4-Paper.pdf.

Heck, Patrick R., Christopher F. Chabris, Duncan J. Watts, and Michelle N. Meyer. 2020. “Objecting to Experiments Even While Approving of the Policies or Treatments They Compare.” Proceedings of the National Academy of Sciences 117 (32): 18948–50.

Hicks, Mar. 2017. Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing. https://mitpress.mit.edu/books/programmed-inequality.

Hind, Michael, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. 2020. “Experiences with Improving the Transparency of AI Models and Services.” In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. CHI EA ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3334480.3383051.

“Home - Data Cards Playbook.” n.d. Accessed June 27, 2021. https://pair-code.github.io/datacardsplaybook/.

“IBM FactSheets Further Advances Trust in AI.” 2020. IBM Research Blog. July 9, 2020. https://www.ibm.com/blogs/research/2020/07/aifactsheets/.

“InterpretML.” n.d. Accessed June 27, 2021. http://interpretml.github.io/.

“Introducing the Model Card Toolkit for Easier Model Transparency Reporting.” n.d. Google AI Blog (blog). Accessed June 27, 2021. http://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html.

Jones, Karen Sparck. 1994. “Towards Better NLP System Evaluation.” In Proceedings of the Workshop on Human Language Technology, 102–7. HLT ’94. USA: Association for Computational Linguistics. https://doi.org/10.3115/1075812.1075833.

Ke, Alexander, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. 2021. “CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation.” Proceedings of the Conference on Health, Inference, and Learning, April, 116–24. https://doi.org/10.1145/3450439.3451867.

Koh, Pang Wei, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, et al. 2021. “WILDS: A Benchmark of in-the-Wild Distribution Shifts.” ArXiv:2012.07421 [Cs], March. http://arxiv.org/abs/2012.07421.

Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. 2019. “Do Better ImageNet Models Transfer Better?” ArXiv:1805.08974 [Cs, Stat], June. http://arxiv.org/abs/1805.08974.

Larsen, Kai R., Roman Lukyanenko, Roland M. Mueller, Veda C. Storey, Debra VanderMeer, Jeffrey Parsons, and Dirk S. Hovorka. 2020. “Validity in Design Science Research.” In Designing for Digital Transformation. Co-Creating Services with Citizens and Industry, edited by Sara Hofmann, Oliver Müller, and Matti Rossi, 272–82. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64823-7_25.

Learning, Full Stack Deep. n.d. “Full Stack Deep Learning.” Accessed June 25, 2021. https://fullstackdeeplearning.com.

Liao, Renjie, Raquel Urtasun, and Richard Zemel. 2020. “A PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks.” In . https://openreview.net/forum?id=TR-Nj6nFx42.

Lin, Jieyu, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. 2020. “On the Robustness of Cooperative Multi-Agent Reinforcement Learning.” ArXiv:2003.03722 [Cs, Stat], March. http://arxiv.org/abs/2003.03722.

Liu, Boxiang, Jiaji Huang, Xingyu Cai, and Kenneth Church. 2021. “Better than BERT but Worse than Baseline.” ArXiv:2105.05915 [Cs], May. http://arxiv.org/abs/2105.05915.

Lütjens, Björn, Michael Everett, and Jonathan P. How. 2020. “Certified Adversarial Robustness for Deep Reinforcement Learning.” ArXiv:1910.12908 [Cs], March. http://arxiv.org/abs/1910.12908.

Meystel, A., and Elena R. Messina. 2001. “Measuring the Performance and Intelligence of Systems: Proceedings of the 2000 PerMIS Workshop, August 14-16, 2000,” September. https://www.nist.gov/publications/measuring-performance-and-intelligence-systems-proceedings-2000-permis-workshop-august.

Miller, John, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. “The Effect of Natural Distribution Shift on Question Answering Models.” ArXiv:2004.14444 [Cs, Stat], April. http://arxiv.org/abs/2004.14444.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. “Model Cards for Model Reporting.” In Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–29. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596.

“MLOps.” n.d. DataRobot. Accessed June 27, 2021. https://www.datarobot.com/platform/mlops/.

Molina-Markham, Andres, Cory Miniter, Becky Powell, and Ahmad Ridley. 2021. “Network Environment Design for Autonomous Cyberdefense.” ArXiv:2103.07583 [Cs], March. http://arxiv.org/abs/2103.07583.

Molnar, Christoph. 2019. Interpretable Machine Learning. Online Edition. @ChristophMolnar. https://christophm.github.io/interpretable-ml-book/.

Mullen, Nadia, Judith Charlton, Anna Devlin, and Michel Bedard. 2011. Simulator Validity: Behaviors Observed on the Simulator and on the Road. https://trid.trb.org/view/1114738.

National Security Commission on Artificial Intelligence. n.d. “2021 Final Report.” Final Report. Accessed June 24, 2021. https://www.nscai.gov/2021-final-report/.

Neyshabur, Behnam, Srinadh Bhojanapalli, and Nathan Srebro. 2018. “A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks.” ArXiv:1707.09564 [Cs], February. http://arxiv.org/abs/1707.09564.

Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. “InterpretML: A Unified Framework for Machine Learning Interpretability.” ArXiv:1909.09223 [Cs, Stat], September. http://arxiv.org/abs/1909.09223.

OCED, Legal. 2019. “Recommendation of the Council on Artificial Intelligence.” OECD Legal. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

“OECD Principles on Artificial Intelligence - Organisation for Economic Co-Operation and Development.” n.d. Accessed November 30, 2020. http://www.oecd.org/going-digital/ai/principles/.

Paltoo, Dina N. n.d. “Digital Health Equity , Training and Research Consortium,” 12.

“Perl Problems.” n.d. Xkcd. Accessed June 27, 2021. https://xkcd.com/1171/.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.” ArXiv:2001.00973 [Cs], January. http://arxiv.org/abs/2001.00973.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. “Do ImageNet Classifiers Generalize to ImageNet?” ArXiv:1902.10811 [Cs, Stat], June. http://arxiv.org/abs/1902.10811.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier.” In KDD 2016: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM. https://www.kdd.org/kdd2016/subtopic/view/why-should-i-trust-you-explaining-the-predictions-of-any-classifier.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2018. “Anchors: High-Precision Model-Agnostic Explanations.” Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. “Beyond Accuracy: Behavioral Testing of NLP Models with CheckList.” In , 4902–12. https://doi.org/10.18653/v1/2020.acl-main.442.

Richards, John, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. “A Methodology for Creating AI FactSheets.” ArXiv:2006.13796 [Cs], June. http://arxiv.org/abs/2006.13796.

Roelofs, Rebecca, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. 2019. “A Meta-Analysis of Overfitting in Machine Learning.” Advances in Neural Information Processing Systems 32. https://papers.nips.cc/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html.

Sagar, Ram. 2021. “Big Data To Good Data: Andrew Ng Urges ML Community To Be More Data-Centric And Less Model-Centric.” Analytics India Magazine (blog). April 6, 2021. https://analyticsindiamag.com/big-data-to-good-data-andrew-ng-urges-ml-community-to-be-more-data-centric-and-less-model-centric/.

Santurkar, Shibani, Dimitris Tsipras, and Aleksander Madry. 2020. “BREEDS: Benchmarks for Subpopulation Shift.” ArXiv:2008.04859 [Cs, Stat], August. http://arxiv.org/abs/2008.04859.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. “Towards Causal Representation Learning.” ArXiv:2102.11107 [Cs], February. http://arxiv.org/abs/2102.11107.

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. “Machine Learning: The High Interest Credit Card of Technical Debt.” In SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop).

Sendak, Mark, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "'The Human Body Is a Black Box': Supporting Clinical Decision-Making with Deep Learning." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 99–109. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372827.

Su, Dong, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2018. "Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models." In Computer Vision – ECCV 2018, edited by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, 644–61. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01258-8_39.

"SuperGLUE Benchmark." n.d. SuperGLUE Benchmark. Accessed June 24, 2021. https://super.gluebenchmark.com/.

"The National Artificial Intelligence Initiative (NAII)." n.d. National Artificial Intelligence Initiative. Accessed June 25, 2021. https://www.ai.gov/.

"The OECD Artificial Intelligence Policy Observatory - OECD.AI." n.d. Accessed June 24, 2021. https://oecd.ai/.

"Uncertainty Quantification 360." n.d. Accessed June 24, 2021. http://uq360-dev.mybluemix.net/uq360-dev.mybluemix.net.

Wu, Eric, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. 2021. "How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals." Nature Medicine 27 (4): 582–84. https://doi.org/10.1038/s41591-021-01312-x.

Yadav, Chhavi, and Léon Bottou. 2019. "Cold Case: The Lost MNIST Digits." ArXiv:1905.10498 [Cs, Stat], November. http://arxiv.org/abs/1905.10498.