Put AI in the Human Loop

Martha G. Russell[1], Rama Akkiraju[2]

The algorithmic age is here. Algorithms describe the mathematical relationships of chosen attributes and their values; they also describe the rules by which a sequence of specified actions will take place. Attributes, values, relationships, and thresholds are selected and brought together with rules to create algorithms that run on data. In this way, algorithmic models constitute a perspective on reality. They provide the artificial intelligence for risk analysis, outcome prediction, image analysis, natural language processing, and other routine and non-routine tasks in which statistical computations assess likelihood and predict outcomes. By their formulaic nature, algorithms represent an encapsulation of the real world, but not the whole; the encapsulation includes the biases inherent in the formula and the data, as well as in the model's objectives. Because of this, algorithmic bias in AI is ever present.

Depending on the objectives of the model, some biases may support the model's effectiveness. However, many biases, particularly the implicit ones, can be detrimental to the model as well as to the people affected by its outcomes. For example, companies are beginning to incorporate AI-based prediction models as decision support tools in many scenarios such as insurance[i] and loan approvals,[ii] and college and job application screening. The growing reliance by human intelligence systems on data-driven models using AI signals an urgent need to understand biases in AI, educate current and future AI developers and users, and fine-tune the human oversight on algorithms for AI. With this perspective, we are putting AI into the networks, loops and relationships of humans, rather than keeping humans in the AI loops.

With the increase of statistical machine learning models in real-world business applications, concerns about the predictions made by biases are on the rise. Recent academic studies have documented discriminating biases in identifying people of certain race or color from various parts

[1] Martha G. Russell is Senior Research Scholar at the Human Sciences and Technologies Advanced Research Institute and Executive Director of mediaX at Stanford University.

[2] Rama Akkiraju is Director, Distinguished Engineer, Master Inventor and a member of IBM's Academy of Technology at IBM's Watson division. She leads the Artificial Intelligence (AI) Operations mission at IBM Watson group.

of the world.[iii] Undesirable biases enter into statistical machine learning models via the business objectives that motivate computerized data analysis, the lack of clarity on those objectives, through training data used to create the models, in the features used by the models to "learn," and through the lack of diligence updating the models.

Real world scenarios are complex, generally requiring more context for decisions than is possible with the black and white – yes/no answers – that often result from computerized systems. The transfer of a model developed for a specific context for re-use in a different context carries risks that require socio-technical intelligence to evaluate. Against the pressures of time and budget, shortcuts are often celebrated. As the system evolves, known risks may be set aside; and unknown risks persist.

One approach to mitigate undesirable biases is to try to fix them after they have been discovered in working models. Another school of thought, a proactive approach, proposes techniques to build machine learning models that clearly articulate the objective of the model and are inherently representative of the subject-matter-expert opinion, thereby reducing the likelihood of undesirable biases in the first place. As AI becomes more and more pervasive, reducing and remedying unwanted bias becomes more and more important in the exercise of human intelligence.

Awareness of these issues has prompted investigations into AI bias and reviews of approaches to detect and deal with them. One of these reviews took place in the workshop, "From Humans to Algorithms to Data Biases in AI", held in January 2019 at the Hawaii International Conference on System Sciences, now in its 52nd year of multidisciplinary of convening high caliber scholars and professionals in academia, industry and government agencies. The workshop was convened by Rama Akkiraju, Distinguished Engineer, Master Inventor, and a member of IBM Watson Division's Academy of Technology. Bias-proofing tools and perspectives – from user to developer to enterprise to society – were presented by six thought leaders from academic and industry and discussed with the standing-room-only audience.

Algorithms provide the statistical rules for natural language processing (NLP) applications, increasingly embedded in automated systems to facilitate interaction, protect and assist people. To do this AI-based inferences about emotion, personality traits and physiological states, such as adverse drug reactions, are created from the analysis of words, phrases, Tweets, and sentences. Using human annotation of sentiment and emotions in such words, a lexicon of sentiment association scores has been created for such components. Over several years, the development of these scores now includes intensity and valence of emotions (positive/negative), sentiment (arousal/dominance), and stance (learnings pro- or con-) – as well as the relationships among these. The data, annotations and scores of this research are now the basis for the Equity Evaluation Corpus (EEC), one component of a Sentiment Analysis System developed at the National Research Council of Canada by Research Scientist Svetlana Kiritchenko, working in conjunction with Senior Research Scientist Saif M. Mohammad.[iv]

Importantly, the EEC includes over 8 thousand sentences containing gender- and race-associated words. It has been used to analyze the output of over 200 natural language processing (NLP) systems as part of a shared task on predicting sentiment and emotion intensity. In this assessment, Kiritchenko's team found that higher emotional intensity scores were associated with gender or race in over 75% of the NLP systems. This is considered to be a biased outcome –an inappropriately biased outcome. Using examples from search terms and results, loan eligibility, crime recidivism prediction systems, and resumé sorting systems, she provided examples in which the data and or the AI system has introduced undesirable biases that affected decision outcomes.

The ongoing objective of Kiritchenko team's is to identify the extent to which various machine learning architectures accentuate or mitigate inappropriate biases. Emotion intensity is an important characteristic in the evaluation of social media messages, such as Twitter, and is used by some services to assess public opinion, as well as to identify and take down inappropriate messages - for consumer protection and, in some cases, for censorship.

By adding sentences associated with country names, professions, and field of study, the EEC will be extended to study the source of bias in word embeddings, sentiment lexicons, lexical semantic resources, and the like. Kiritchenko emphasized that there are no simple solutions for comprehensively dealing with inappropriate biases, and she encouraged attention to machine learning architectures in addition to biases introduced by humans or by AI models.

"Our challenge is how to heighten intelligence about the uses of AI and to deepen the sensitivity about the source of data and how it will influence those decisions. Algorithmic bias is solvable," said Meeta Dash, Project Director for Figure Eight – a Human in the Loop Machine Learning Platform, whose work involves assessments to reduce unwanted bias. Developing AI models for decision making requires crisp definition of the decision specifics for which the model is solving.

Reducing bias in this type of AI depends on enlightened consideration of the potential impact of the model's results on a protected class or status of people. Because both training data and the structure of the model's task can bias algorithms created for these models, Dash encouraged development teams to minimize unwanted bias in a three-step process: balancing critical attribute-level training data, configuring specific attribute-level thresh holds, and routinely invoking a case-based diversity review of models.

She urged decision makers using the results of AI-based tools to deeply consider the attributes of the embedded AI models, taking into account whether humans are involved and, if so, assessing whether the model accounts for human well-being in a wholistic way. With examples of race and gender bias in image detection, Dash emphasized that sensitivity to diversity is essential in the many applications in which subjective opinions matter: sentiment analysis, judgment of relevance, data categorization, content moderation, image moderation, audio/text collection for speech AI. She encouraged AI developers to actively seek real-world feedback on their models, annotate with diversity, have options to gracefully fallback for revisions, and expect continual iterations.

From the perspective of services, new methods are being developed to discover unseen biases. Citing Buolamwini and Gebru's 2018 study, which documented errors in more than a third of the cases involving dark-skinned females, Anbang Xu, Researcher at IBM Watson AI Operations, described three types of biases and potential methods to address them. Model biases, in which the formula itself favors producing errors in the outcome, can be remedied by proactive testing frameworks in which analysis tools identify errors, followed by use of crowd forces[v] to validate and categorize the errors, which are then returned to developers for analysis and repair. Data bias, introduced through the selection of training data that is unbalanced with respect to the objective function of the algorithm, can benefit from feature extraction based on a general framework - to quantify both predefined and unknown biases in the data. Xu suggested that this be followed by a quantification of the distribution of the discrepancy,[vi] after which corrective actions can be taken to mitigate the bias. He cautioned that both explicit biases (observable by developers and users) as well as latent biases (anticipated but only discoverable with statistical procedures) be considered.[vii]

A third type of bias discussed by Xu concerns technical operations and the annotation tasks assigned to developers. Significant professional and managerial judgment is exercised in the choice and assignment of annotation tasks (to experts or to crowd workers[viii]) for algorithm development. The curation of those annotations requires the human task router to consider the annotators' expertise, demographic factors, and overall objectives of the algorithm.

At the enterprise level, businesses are transitioning from organizing as "intelligences apart" to "intelligences augmented," advised Ajay Chander, Vice President of Research at Fujitsu Laboratories of America, Inc. This reflects a major shift in the human-machine collaboration continuum as we journey from being apart to being augmented to being alongside autonomous machine intelligences. For successful use of AI in human sensing, decision making, and action, enterprises need their employees to co-create with AI, to manage AI, to select and use AI appropriately, and to edit AI as needed for objectives set by humans. The drive to "just make it work" can produce a problematic opaqueness that occludes the transparency necessary to identify and correct perspective bias, data bias, and/or engineer bias. In contrast, Transparent AI enables human-AI co-creation by encouraging the user to ask the trained AI model a question, receive an answer with an explanation, and iterate this process (permitting the AI model to be "tuned") until the question and answer (with explanation) align appropriately. Chander anticipates that bias testing will become an essential part of AI development for relevant, robust and resilient AI systems. This will include testing for access to opportunity-blocking biases (impact on the receiver) as well as risk-reduction/return optimization biases (critical for decision makers).

Further, Chander elaborated that transparency into algorithms and AI systems is necessary in order to have interactions points that reveal and enable reaction to both types of bias. Accessible AI allows a diverse and inclusive set of questions to be asked of AI; explainable AI reveals bias; and tunable AI allows response to biases that warrant remedy. It is easier to detect bias, considered a structural property by Chander, in deterministic systems –those in which the antecedents control the outcome. It is harder to detect bias in non-deterministic systems – those in which non-rational factors are combined with AI predictions to influence outcomes. The crux of the AI challenge, he asserted, is aligning the rationality of artificial intelligence with human intelligence, which is both rational and irrational. With a case study of optimizing sales for a large global services unit, Chander illustrated how tunable AI-driven exploration of strategic options can bridge the awareness of options to yield confidence in value-based decisions.

Taking the definition of bias an inclination or a prejudice for or against a person or group so as to be considered unfair, the issue of fairness becomes a concern for services whose systems use AI algorithms, as well as for the people who use those systems. A fair system treats people equally, or in a way that considers everything that reasonably has an effect on a situation. This presents a challenge, according to Martha Russell, Executive Director of mediaX at Stanford University and Senior Research Scholar at Stanford's Human Sciences and Technologies Advanced Research Institute. "People understand fairness differently. There are great differences in personal versus community understandings of what might be considered 'fair'." According to Russell, the debates on fairness currently focus on the instruments. There is a need to look beyond whether an AI algorithm is considered fair, she urges, to also examine fairness in the context of the decisions on which the AI is being applied, such as medical contexts (in the quest for personalized medicine[ix]) or political contexts (to analyze demographic makeup of neighborhoods[x].) In legal contexts (harm caused by automated devices), for example, fairness determination has been based upon intent, which is nearly impossible to determine with an algorithm[xi]. Definitions of fairness are deeply related to the relative values of a specific community[xii]. Moreover, Russell emphasized, one can often identify a bad outcome, but there is no objective measure of good. Equal opportunity is not the same as equal outcome.

Because of this, she continued, algorithmic fairness is an issue for policy and ethics, as well as engineering[xiii]. Ethical developers need to think in a utilitarian mode, beyond abstract concepts of fairness, focusing on the well-defined outcome objectives of algorithms that have been developed for specific contexts. This requires a case-based approach - identifying a set of test problems against which algorithmic outcomes can be evaluated – in context.

A case-based approach requires from both developers and users the kind of critical thinking that comes from interdisciplinary education, which has high priority at Stanford University. Russell emphasized that algorithms and AI systems are authored texts[xiv], written by individuals and carrying with them the implicit values, biases, and ideologies of their authors. She shared a co-evolution perspective - that humans and technologies are co-evolving and that adopting new technologies involves an iterative process of social learning. Devices and their AI's teach and influence the cultural values and societal world-views of people, as we use them. The preferred role of AI is to assist with insights and synthesis, not to be adversarial, to replace human wisdom and judgment, or to create dependency on the artificial intelligence. AI needs to be carefully positioned in the loops and networks of people working together, who – Russell suggests – have unlimited capacity for contextual integration, wisdom and judgment.

"Everyone is impacted by bias in AI," cautioned Jim Spohrer, Director of DBG/DEG and Measuring AI Progress Cognitive Opentech Group (MAP COG) at IBM "and it requires lots of different people to combat it." It is encouraging, he said, to witness the growing interaction between the Service Science community, in which trust is built through value co-creation and transdisciplinary endeavors) and the OpenTech AI community – in which trust is developed through ethical, safe, explainable, and open communities. Spohrer highlighted a megatrend of open development evidenced by Tensorflow, PiTorch, Onyx, Redhat, the github contributions of Microsoft and the Linux Foundation's Deep Learning, as well as the many adjacent open source communities providing explanations and responding to adversarial attacks. These are crucial steps forward, Spohrer argued, because "one of the biggest benefits of AI may be persuading us all to be more ethical as we understand the historic origins and sources of our biases in complex decision-making; many of these work against us and against humankind in the world today."

Summing the advice of panelists – and the discussion with the audience - on preventing, diagnosing and fixing bias in algorithms used for artificial intelligence:

1. Carefully define the decision that you're asking your model to solve, the context of the decision and the potential impact on human beings.
2. Think broadly about end users, test widely, and acknowledge the difference between equal opportunity and equal outcome.
3. Be transparent on the nature of the data the system is trained on, as well as the data used by the AI model.
4. Ensure that the model and its output are appropriately positioned in the human decision making system, actively learning, with new real-world data as soon as it becomes available.
5. Be empathetic and understand that your end-users will use the system differently. Build user experiences for failure/edge cases.
6. Test for errors at every step of the development and application processes.
7. Take feedback and have options to gracefully fall back.
8. Allow for ambiguity and difference of opinion.
9. Educate current and future algorithm developers for critical thinking, across multiple social and technical issues.
10. Accelerate progress through the development of trust by contributing to the service sciences and open tech communities.

Acknowledging the continuing co-evolution of new issues and solutions in the fast-paced quest for intelligent science and technology systems, scientists and educators in both business and academia have a shared responsibility for thoughtful consideration of the human, social, technological, and organizational factors for preventing and dealing with undesirable biases in algorithms for AI. The need to do so underscores the need for transdisciplinary education and for lifelong learning at the speed of change. And it highlights the responsibilities of humans to include AIs appropriately in their networks, relationships and business practices.

Many thanks to the presenters and attendees of the HICSS Workshop, January 8, 2019
"From Humans to Algorithms to Data Biases in AI"
Hawaii International Conference on System Sciences (HICSS) 52, Maui, Hawaii

[i] Diana Hope (2018) "How AI is transforming Lending and Loan Management, Smart Data Collective, December 14, 2018, https://www.smartdatacollective.com/how-ai-is-transforming-lending-and-loan-management/

ii Boan Yang, Ling Xi, Qinghua Xie, Jing Xu, (2001) Development of a KBS for managing bank loan risk, *Knowledge-Based Systems*, 14:6, 2990302.

[iii] Joy Buolamwini and Tmmit Gebru (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification, *Proceedings of Machine Learning Research*, 81:1-5.

[iv] Svetlana Kiritchenko and Saif M. Mohammad (2018) Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of International Workshop on Semantic Evaluation* (Sem-Eval-2018), New Orleans, LA, USA, June 2018.

[v] Siwei Fu, Anbang Xu, Xiaotong Liu, Huimin Zhou, Rama Akkiraju (2018) Challenge AI mind: A crowd system for proactive AI testing, arXiv.org>cs>arXiv:1810.09030.

[vi] Anbang Xu et al. A General Methodology to Quantify Biases in Natural Language Data, in submission.

[vii] Anbang Xu et al. Evaluating Group Biases in Text, in submission.

[viii] Chenguang Wang, Alan Akbik, Laura Chiticariu, Yunyao Li, Fei Xia, Angang Xu (2017) CROWD-IN-THE-LOOP: A hybrid approach for annotating semantic roles, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1913–1922, Copenhagen, Denmark, September 7–11, 2017.

[ix] W. W Yim, A. J. Wheeler, C. Curtin, T. H. Wagner, T. Hernandez-Boussard (2018) Secondary use of electronic medical records for clinical research: Challenges and opportunities. *Convergent Science Physical Oncology,* 4 (1).

[x] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, Fei-fei Li (2017) Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, *PNAS*, December 12, 2017, 114:50; 13108-13113.

[xi] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, Aziz Hug (2017) Algorithmic decision making and the cost of fairness, arXiv:1701.08230v4 {CS.CY] 10 June 2017.

[xii] Daniel E. Ho and Kristen M. Altenburger, (2018) When algorithms import private bias into public enforcement: The promise and limitations of statistical debiasing solutions, *Journal of Institutional Theoretical Economics*, November 14, 2018.

[xiii] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel (2017) The problem of infra-marginality in outcome tests for discrimination, *The Annals of Applied Statistics*, 11:3, 1193-1216.

[xiv] Aylin Caliskan-Islam, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, and Rachel Greenstadt (2015) De-anonymizing programmers via code stylometry, 24th Usenix Security Symposium, USENIX.