AI Policy Matters

Larry Medsker (The George Washington University; lrm@gwu.edu)
Farhana Faruqe (The George Washington University; faruqe@gwmail.gwu.edu)
DOI: 10.1145/3402562.3402566

Abstract

AI Policy Matters is a regular column in *AI Matters* featuring summaries and commentary based on postings that appear twice a month in the *AI Matters* blog (https://sigai.acm.org/aimatters/blog/). We welcome everyone to make blog comments so we can develop a rich knowledge base of information and ideas representing the SIGAI members.

AI and DC

News Items for February, 2020

OECD launched the OECD.AI Observatory, an online platform to shape and share AI policies across the globe.

The White House released the American Artificial Intelligence Initiative: Year One Annual Report and supported the OECD policy

Bias, Ethics, and Policy

The Policy Matters blog has started a series on AI and Bias, with posts on background and context of bias in general and then focused on specific instances of bias in current and emerging areas of AI. The information is intended to inform ideas and discussions on public policy. We look forward to your comments and suggestions. Extensive work such as "A Survey on Bias and Fairness in Machine Learning" by Ninareh Mehrabi *et al.* is one of the background resources for the conversation. Additional resources are provided by Barocas, *et al.* The guest co-author of this column is Farhana Faruqe, doctoral student in the George Washington University Human-Technology Collaboration program.

AI Bias and Discrimination

Discrimination, unfairness, and bias are terms used frequently these days in the context of AI and data science applications that make decisions in the everyday lives of individuals and groups. Machine learning applications depend on data sets that are usually a reflection of our real world in which individuals have intentional and unintentional biases that may cause unfair actions and discrimination. Broadly, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making.

Direct Discrimination. As described by Ninareh Mehrabi *et al.*, "Direct discrimination happens when protected attributes of individuals explicitly result in non-favorable outcomes toward them". Some traits like race, color, national origin, religion, sex, family status, disability, marital status, recipient of public assistance, and age are identified as sensitive attributes or protected attributes in the machine learning world. It is not legal to discriminate against these sensitive attributes, which are listed by the FHA and Equal Credit Opportunity Act (ECOA).

Indirect Discrimination. Even if sensitive or protected attributes are not used against an individual, still indirect discrimination can happen. For example, residential zip code is not categorized as a protected attribute, but from the zip code one may find out about race which is a protected attribute. So, "protected groups or individuals still can get treated unjustly as a result of implicit effects from their protected attributes".

Systemic Discrimination. In the nursing profession, the custom is to expect a nurse to be a woman. So, excluding qualified male nurses for nursing position is an example of systematic discrimination. Systematic discrimination is defined as "policies, customs, or behaviors that are a part of the culture or structure of an organization that may perpetuate discrimination against certain subgroups of the population."

Statistical Discrimination. In law enforcement, racial profiling is an example of statistical discrimination. In this case, minority

drivers are pulled over more often than white drivers. The authors define "statistical discrimination is a phenomenon where decision-makers use average group statistics to judge an individual belonging to that group."

Explainable Discrimination. In some cases, discrimination can be explained using attributes like working hours and education, which is legal and acceptable as well. In a widely used dataset in the fairness domain, males on average have a higher annual income than females because on average females work fewer hours per week than males do. Decisions made without considering working hours could lead to discrimination.

Unexplainable Discrimination. This type of discrimination is not legal as explainable discrimination because "the discrimination toward a group is unjustified". Some researchers have introduced techniques during data pre-processing and training to remove unexplainable discrimination.

AI Bias and Fairness

In terms of decision-making and policy, fairness can be defined as "the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics". Six of the most used definitions are equalized odds, equal opportunity, demographic parity, fairness through unawareness or group unaware, treatment equality.

The concept of equalized odds and equal opportunity is that individuals who qualify for a desirable outcome should have an equal chance of being correctly assigned regardless of an individual's belonging to a protected or unprotected group (e.g., female/male). Along with other concepts like "demographic parity" and "group unaware" are illustrated by the Google visualization research team with nice visualizations using a "simulating loan decisions for different groups". The focus of equal opportunity is on the outcome of the true positive rate of the group. On the other hand, the focus of the demographic parity is on the positive rate only. Consider a loan approval process for two groups: group A and group B. For demographic parity, the overall number of approved loans should be equal in both group A and group B regardless of a person belonging to a protected group. Since the focus for demographic parity is on overall loan approval rate, the rate should be equal for both groups. Some people in group A who would pay back the loan might be disadvantaged compared to the people in group B who might not pay back the loan; however, the people in group A will not be at a disadvantage in the equal opportunity concept, since this concept focuses on true positive rate. As an example of fairness through unawareness "an algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process". All of the fairness concepts or definitions either fall under individual fairness, subgroup fairness or group fairness. For example, demographic parity, equalized odds, and equal opportunity are the group fairness type; fairness through awareness falls under the individual type where the focus is not on the overall group.

A definition of bias can be in the three categories data, algorithm and a user interaction feedback loop: **Data** – behavioral bias, presentation bias, linking bias, and content production bias; **Algorithmic** – historical bias, aggregation bias, temporal bias, and social bias falls; **User Interaction** – popularity bias, ranking bias, evaluation bias, and emergent bias. Bias is a large domain with much to explore and take into consideration. Bias and public policy will be discussed in future blog posts.

AI and Work

The AI and Work session in the recent AAAI FSS-19 Symposia was a good example of exploration and research that should inform public policy making. All topics are related to, or could be expedited by, good public policy and aware policy makers. The deployment of AI technologies in the future will likely require humans to collaborate with AI systems, and this realization highlights the need for more sustained research on how to design such systems. High levels of autonomy and the ability to learn and interact with other systems, including humans redesigning work and rethinking incomes with bold ideas to improve the lives of workers and provide more interesting jobs with more meaning, purpose and dignity.

1. How do we design effective human-AI teaming?

2. What does participatory design look like for AI in the context of work?
3. What training do people need to be able to work successfully with smarter systems?

Time Frame for AI Impact

An interesting IEEE Spectrum article "AI and Economic Productivity: Expect Evolution, Not Revolution" by Jeffrey Funk questions popular claims about the pace of AI's impact on productivity and the economy. He asserts that "Despite the hype, artificial intelligence will take years to significantly boost economic productivity". If correct, this will have serious implications for public policy making. The article raises good points, but many of the examples do not look like real AI, at least as a dominant component. Putting "smart" in the name of a product does not make it AI, and automation does not necessarily use AI.

On a broader note, we should care about the technology language we use and beware of the usual practices in commercialization. As discussed previously, expanding the meanings of terms like AI, machine learning, and algorithms makes rational discourse more difficult. Some of us remember marketing of expert systems and relational databases: companies do a disservice to society by claiming each breakthrough technology actually is in their products. Here we go again today, with anything counting as AI depending on the point you want to make and the products you want to sell.

Another issue raised by the article relates to startups as the leaders of economic impact, as opposed to innovations from established industry and government labs. Any technology has an adoption curve, going from early adopters through the laggards, of about seven years. If you add to that the difficulties of making a startup succeed, a decade or so is probably the minimum timescale. A better perspective on revolution versus evolution could come from longitudinal evaluations looking at trends. In that case, a good endpoint for a hypothesis about dramatic impact on productivity might be the 2030-2035 time frame. Another difficulty of using a vague and broad notion of AI is that policymakers could miss the revolutionary impact of data science, which can, but may not, involve real AI. Data science probably has the best chance of dramatically impacting society and the economy soon and has the advantage of not having to involve designing and manufacturing physical objects, and thus not waiting for consumers to adopt new products. Data Science is already affecting society and employment through obvious, and not so obvious, revolutionary impacts on our work and lives.

Please join our discussions at the SIGAI Policy Blog.

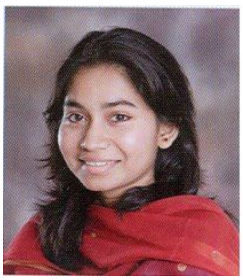
References

1. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning". CoRR, abs/1908.09635, 2019.

2. Moritz Hardt, Eric Price, and Nati Srebro. 2016. "Equality of Opportunity in Supervised Learning". In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3315–3323. http://papers.nips.cc/paper/ 6374-equality-of-opportunity-in-supervised-learning.pdf.

3. Martin Wattenberg, Fernanda Viegas, and Moritz Hardt. "Attacking discrimination with smarter machine learning." Accessed at https://research.google.com/bigpicture/attacking-discrimination-in-ml/, 2016.

Larry Medsker is Research Professor and founding director of the Data Science graduate program at The George Washington University. He is a faculty member in the GW Human-Technology Collaboration Lab and Ph.D. His research in AI includes work on artificial neural networks, hybrid intelligent systems, and the impacts of AI on society and policy. He is the Public Policy Officer for the ACM SIGAI.



Farhana Faruqe is a doctoral student in the GW Human-Technology Collaboration Lab and Ph.D. program. Her research includes work on impacts of cognitive assistants on human-technology collaboration. She investigates issues in AI and Ethics, human-machine trust, and impacts of AI system bias on individuals and society.