“I believe this artificial intelligence is going to be our partner. If we misuse it, it will be a risk. If we use it right, it can be our partner.”

— Masayoshi Son

Artificial Intelligence and Bias
Dear colleagues,

Our April newsletter discusses a hot topic in the field of medical imaging: “Artificial Intelligence and Bias”. We will learn insights about bias in AI algorithms and the diversity of people who work on AI. As this new technology evolves, we must address related ethical and moral questions.

To generate balanced input for our newsletter from colleagues with diverse backgrounds, I reached out to 8 male non-minority colleagues, 8 female non-minority colleagues and 8 underrepresented minority (URM) colleagues. Firstly, I got an overwhelming number of positive responses. However, I also noted some differences: Of the 8 males, 7 got back to me and 6 agreed to write a story for our newsletter. Of the 8 females, 5 got back to me and 4 agreed to share a story. Of the 8 URM colleagues, 2 got back to me and 1 agreed to write a story. There was clearly a lower readiness for engagement of our female and URM colleagues. This can possibly be explained by the fact that minority groups might receive proportionally more requests than their male non-minority counterparts. It also shows the importance of increasing the representation of female and non-URM colleagues throughout our academic enterprise and computer sciences specifically. Different viewpoints can identify different strengths and weaknesses of machine learning algorithms. We need a diverse group of people to generate a multi-faceted discussion around new technologies in order to generate tools that can serve everyone in our community.

As outlined by our authors, mounting evidence suggests that AI will likely transcend almost every aspect of our work and private life. It is up to us to define, who these computers represent and which problems they will tackle. In addition, interactions with AI may raise philosophical questions and enable us to recognize qualities that are fundamentally human: What is the difference between answering a question and asking a question? What is the difference between innovation and imagination? What is the difference between intelligence and consciousness?

Thank you for joining the conversation!

Heike E. Daldrup-Link, M.D, Ph.D.
Associate Chair for Diversity
Department of Radiology

Meet the “ems” -- machines that emulate human brains and can think, feel and work just like the brains they're copied from. Futurist and social scientist Robin Hanson describes a possible future when ems take over the global economy, running on superfast computers and copying themselves to multitask, leaving humans with only one choice: to retire, forever.

https://www.ted.com/talks/robin_hanson_what_would_happen_if_we_upload_our_brains_to_computers?referrer=playlist-alternate_timelines_of_the_future
Over the last several years, we have witnessed both the incredible rise of AI algorithms as well as examples of its failures and challenges. Deep learning in particular offers the potential to solve complex problems that have so far resisted our best efforts, but at the cost of being so immensely complex that we cannot easily visualize or understand why they are so successful. It is becoming clear, though, that AI can produce biased results, and that some of these biases have real world consequences. This may be due to the demographics of the people creating these algorithms – described recently by Fei Fei Li, professor of computer science at Stanford, as predominantly “white guys in hoodies” – or it could be due to inherent biases in existing databases used to train these networks. One example of the latter is the Google search engine – perhaps the most common way many of us get our information about the world. A widely publicized example of bias was the kind of results obtained when entering the search terms “three black teenagers” versus “three white teenagers,” where the top ranked images for the black teenager search were police mug shots while the white teenager search were happy, smiling kids at play.

This pointed out something that should have been obvious to us: that if training data contains unrepresentative data, this bias will be perpetuated into the algorithm’s results. What does that mean for us in medicine? If our definition of bias is that a classification method performs poorly on some patient subsets, it is hard to avoid this outcome; patients differ based on many characteristics: gender, ethnicity, different genetic profiles and proclivities to certain diseases, etc. Clearly, the narrower the training population, the more likely that the method will perform poorly on subsets of patients who do not share the same underlying features. One example we have explored in our research is using AI to predict gold-standard PET cerebral blood flow images from MRI using simultaneous PET/MRI. We had two distinct cohorts: healthy normal subjects and patients with Moyamoya disease, a cerebrovascular disease that causes arterial stenosis and chronic ischemia. In such a simple system, we can assess the quality of the results directly: how accurate are the blood flow results predicted by the algorithm?

We found that a deep learning model trained on healthy subjects performs well when tested on other healthy subjects, but that its performance in patients with disease is significantly poorer. Interestingly, models trained only on patients
seem to perform equally well on healthy subjects. While a very basic example, it shows bias is not just an issue for classification algorithms but also for image reconstruction. This is beginning to be recognized by radiologists, who have shown that algorithms trained at one institution or with only one vendor’s imaging machines perform surprisingly poorly when transplanted to different settings, a phenomenon sometimes known as “brittle AI.” While initial studies have been content to demonstrate great performance in relative narrow domains, the current focus of much medical AI is how to make these algorithms more generalizable.

Attention on how to accomplish this is a hot topic. Larger, more diverse datasets are an obvious solution, but often either do not exist or are challenging to acquire. Another idea, “transfer learning,” in which algorithms are trained on the largest possible datasets regardless of bias, followed by adjustments to the model weights using smaller datasets more representative of local populations is being actively explored. Other questions surround whether to “handicap” model performance on better represented populations such that all subgroups demonstrate similar quality, though this goes against our inherent desire to do the absolute best we can at prediction. What fairness means in these contexts where every data point represents a patient is a deep question. How do we square the imperative of personalized medicine with our desire to avoid bias? We are just beginning the exciting journey of healthcare AI; making sure that its rewards accrue equally to all patients will be an important challenge.

Greg Zaharchuk, MD, PhD
Professor
Stanford Medicine | Radiology
Let's get something out of the way quickly: while I am writing for a radiology magazine, I do not have competency in this discipline. Luckily, "Artificial Intelligence and Bias" is a broader topic, and one that interests me closely.

As the saying goes, life is a series of “decisions, decisions, decisions.” We make them with partial information, trying to satisfy multiple constraints and maximize diverse criteria of value, aware that multiple choices are acceptable, and that unforeseen circumstances might change the outcome we obtain. What got me to study statistics in graduate school was the realization that one could take advantage of the clear logic of mathematical reasoning and of the power of computation to formalize how we learn from experiences and how we make well thought-out decisions in the presence of uncertainty. Rationality and precision do not need to be thrown out of the window when we deal with a world that is not as black or white as the one described in high school math problems: we can deal with complexity, noise, partial information, and more and still obtain a meaningful, quantitative and sharable assessment of the different options in front of us.

For most of my career the principal domain of application of statistical reasoning has been scientific decisions: have we gathered enough evidence to reject the standing hypothesis? Can we conclude that a given treatment is more effective than placebo?

During the last decade we have witnessed an increased application of data analysis to support decisions outside the scientific domain: from recommendation systems for movies to evaluations of the likelihood that a potential parolee will commit another crime and automatic interpretation of medical images. These are interesting and exciting developments, but, unsurprisingly, they bring up a number of challenges, like that of racial or gender bias in AI. To discuss these challenges, it is helpful to start by clarifying the meaning of the terms at play.

The workshop at Dartmouth College that is credited with introducing the phrase “Artificial Intelligence” was based on the premise that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.” The overarching goal of the program was described as “to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” (Cite: McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence., http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf August, 1955)

The term “Artificial Intelligence” indeed evokes the idea of a “thinking machine”: a machine that understands rules, meanings, objectives, constraints just as we do and can leverage superior computing power to produce “optimal” decisions. While a lot of research has been devoted over the years to construct such machines, and some progress has been made, by and large what are currently considered as the successes in AI stem from a very different premise. Consider the quite astounding progress in natural language processing that allows for machine translations and verbal interactions with devices. Good quality translations have not been achieved by explicitly teaching the machine rules of the languages, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. The machine does sophisticated pattern matching and, seeing that the word “house” is translated as “casa” in Italian in
Artificial Intelligence and Bias

The Story of Humanity: To Create Machines that Augment our Abilities

most instances, learns that this is the outcome to produce. (Cite: Alon Halevy, Peter Norvig, and Fernando Pereira, The Unreasonable Effectiveness of Data, IEEE INTELLIGENT SYSTEMS, 2009). There is nothing wrong with this (in fact, it is arguably closer to how humans learn languages than memorizing dictionaries and grammars), but being aware of how the machine works helps us understand what its limits and possibilities are.

We still are not able to describe precisely “every aspect” of how we learn, or “any other feature of intelligence,” but we can get computers to imitate our performance in many tasks very well. Machine learning, a field at the intersection of statistics and computer science, has developed increasingly sophisticated tools to train machines on the basis of labeled data. But by and large the biggest determining factor for how computers are successful at a given task is the size of the training data available, rather than the algorithm used. Machines do better at translating a document than at listing the topics it covers even if the first is a more difficult task simply because, at the training stage, there are much more instances of translations than there are of topic description. This reliance on available training data has important implications: the performance of an algorithm will be a reflection of the data on which it is trained. Let's investigate some.

Since more data is better, it is generally considered a good idea to capture as much as possible of the “data available in the wild” and use it for training purposes. But this data, observational in nature, can reflect errors and distortions that we might not want to perpetrate. In the US, currently, the majority of computer programmers are male. It is not surprising that a machine that learns the characteristics associated to computer programmer by summarizing the data available in online texts ends up thinking that this is a “male profession.” (Cite: Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (2016) “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” NIPS). However, if we were to use this information to select the best computer programmers out of a pool of applicants, we would do poorly: there is nothing substantial that makes males better programmers. Indeed, while we are playing with stereotypes, it is quite possible that the patience, precision, and ability to listen that are traditionally associated with females would make them better programmers. By training machines to imitate the status quo, we risk to rectify stereotypes.

Furthermore, not all data is equally representative of the populations of interest. For example, image processing software trained to distinguish males from females on photos of individuals from Northern Europe is likely to have different performance on images of subjects coming from sub-saharan Africa or east Asia. And if machines rely so crucially on training data, how robust are they going to be? Which action will they suggest when presented with a setting that is outside the space of what they have seen so far? How can they be mindful and respectful of difference?

The dictionary on my computer defines bias as “prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.” The discussion above provides some examples of how bias might be introduced in recommending systems based on machine learning: women might be at a disadvantage in being selected for some jobs, the performance of some classification algorithm might be poorer on members of different ethnic groups, etc. These issues clearly need to be addressed if we are to rely on the use of these computerized tools to inform our decisions. The dangers are particularly acute because automated rules can be quickly employed on extremely large scale, affecting unprecedented numbers of subjects. Indeed, researchers in machine learning have

“By training machines to imitate the status quo, we risk to rectify stereotypes.”
started to think about these problems. Because the term “bias” has a special technical meaning in statistics, to avoid confusion, the discussion of these issues goes under the umbrella name of algorithmic “fairness.”

“Given that human beings are different in a myriad of ways,” defining what makes a treatment “fair” is not simple. We can look at the debate in political philosophy, where, for example, Dworkin argues that “the state is obligated to treat all of its members with equal concern and respect” (cite: D. Satz (2010) Why some things should not be for sale, Oxford University press). To guarantee that this is the case, a number of United States laws specify that it is illegal to discriminate on the basis of some protected characteristics (including age, race, sex). Credit decisions, for example, have to comply with state and federal fair lending rules, which in some cases prohibit differential treatment on the basis of a customer’s race and gender: credit card companies are not allowed to ask applicants what race they are when applying for credit and they have to show that any credit score (possibly derived by machine learning) that they might use is blind to race. On the other hand, in many contexts, in the practice of medicine, characteristics as age, race and sex cannot be ignored if one is to provide equally effective care to members of these different groups: knowledge of them is necessary to calibrate the results of tests, for example.

It is not a surprise, then, that defining algorithmic fairness is challenging, and that multiple perspectives have been put forward in this nascent literature (see the review in Chouldechova and Roth (2018) “The frontiers of fairness in machine learning”). Consider an algorithm that use a subject’s characteristics to recommend them for a desirable outcome (such as being granted parole or a loan): what should we require of it to be fair? A first naive prescription to make sure that the algorithm reflects no racial bias could be that race is not one of the variables inputted to the procedure. However, race correlates with a series of socio-economic indicators, with names, with addresses, and more, and it is quite possible to reconstruct it with some accuracy from other available measurements, so that simply not providing it as input does not guarantee that the outcome might not correlate with it. What other requirements can then guarantee fairness? Some have interpreted an aspiration to “equal opportunity” as the request that the proportion of subjects recommended for the preferred treatment be the same in any race group. If, however, it so happens that the proportion of individuals that would not violate parole is truly different across race groups, this policy could lead to unfair treatment. The algorithm would not recommend for parole subjects that are truly qualified for it or would recommend for parole subjects that are likely to violate it, simply to “even things out” across race groups. Instead, we would like that subjects that are equally qualified for parole are recommended for the same outcome irrespective of their race. A couple of different formalizations of this requirement have been put forward. One states that we want subjects that are similar in the truly relevant characteristics to be treated similarly. Another states that the probability of assigning deserving individuals to the preferred outcome should be the same across races. Another, yet, suggests that to abstract from race in evaluating how much an individual is likely or not to violate parole, we should do the thought experiment of assigning her different races at birth and derive what her currently measured characteristics would be under the different scenarios.

The effort to translate a notion of “fairness” into a precise quantitative property of an algorithm is important, as it is really the only opportunity we have to check if the computerized rules we are considering satisfy what we consider
acceptable. However, in a setting where there is no universal agreement on what fairness means, this is clearly challenging. A number of the definitions that have been put forward are not yet practical, in that they require knowledge of “relevant variables” or counterfactual experiments that we do not have. Moreover, we should not limit ourselves to checking if an algorithm is fair or not, but we should make sure we can construct fair procedures. We have been taking only the very first steps in this direction.

As we envision to automate more and more decisions, we need to realize that this is a far more complex task than tasks like identifying if an image includes a cat or not. On the one hand, in most cases, the machine simply does not have training data with "ground" truth of comparable quality to learn a good classification rule. On the other hand, decisions have consequences that we want to evaluate in the context of a social good that goes beyond fairness. For example, in qualifying individuals for loans, we might want to influence the home-ownership rate of a community, encourage speculation or not, or exert some control on home prices. As we adopt an automated decision rule we want to think about the impact it will have in the long term and what possible feedback loop mechanisms it will generate. These include the effect on the human capital: what are the consequences on pilots' abilities to respond to situations of emergency when we substantially increase the proportion of time in which aircrafts are flying automatically?

Clearly, statisticians and machine learning researchers cannot address these questions on their own. Their responsibility is to contribute some of the building blocks, to make sure the tools they provide have the advertised properties, and to educate everyone about their error rates, striving to avoid unintended consequences. This includes investigating the qualities of the “training data,” highlighting when it is not representative, and making sure that the guarantees associated to a classification rule hold for each “protected” group, highlighting the possible discrepancies in precision.

It is part of the responsibility of researchers in Artificial Intelligence to clarify to the public what exactly machines can or cannot do at a given point in time, and what is the effectiveness with which they can execute certain tasks. And we all have to be conscious that to satisfactorily automate decisions we need to be able to specify clear and complete objectives; simply adopting off-the-shelf algorithms is not enough.

The story of humanity is the story of our ability to create machines that augment our abilities and reach. We can fly, travel under water, construct cities in inhospitable environments; we have even been to the moon. It is not hard to imagine that computers will outperform humans in an increasing number of tasks. We simply need to be able to understand when this is truly the case. A lot of the enthusiasm around the performance of algorithms like neural networks is really about the machine's ability to imitate humans. The consensus of a group of radiologists defines the ground truth we use to train a machine to recognize an anomaly. While we might observe that the machine, having seen a large enough number of examples, might learn to perform better than “the average radiologist” on a comparable test set, it is hard to consider this truly “surpassing human ability.” Especially because humans are extremely more robust: they will have no hesitation in realizing that the picture of an apple does not belong in the medical record of a patient, while the machine might be quite confused. As our algorithms and our training sets get better—allowing the machine to truly match or surpass humans in classification—before we rely on automatic decision rules, we will need to make a serious collective effort to explicitly formulate the goals of common good we intend to pursue.

Chiara Sabatti, PhD
Professor
Stanford Medicine | Biomedical Data Sciences and Statistics
As designers, we are often tasked with creating the future. We are trained to consider all the possible ways the choices we make might impact the user. Products evolve, user's needs change, and we try to understand them. We apply principles of psychology and behavioral science towards due diligence. We try to remain objective as we learn from our tests and we use those insights to make good choices on behalf of the users. Empathy is engrained in the design process but everything that is made is not always intentionally designed. AI and machine learning is becoming increasingly accessible to the masses. AI has already made its way into many complex and nuanced decision-making processes like deciding the fate of potential parolees, informing college admissions, and diagnosing diseases.

With the rise of Big Data and increasingly complex neural networks we have more ways to automate our lives than ever before. Such far-reaching applications mean we have even more responsibility to ask important questions about the potential impacts of AI making some of our most important decisions. AI learns from what we teach it. Who gets to be the teacher? We tell AI how to interpret the data it encounters. Whose perspective guides that interpretation? We allow AI to curate the information that informs our viewpoints. Who decides what information receives priority? Accountability starts by asking difficult questions.

I work in technology, an industry where many glorify the “move fast and break things” ideology. It's also an industry that is still overwhelmingly White-- and overwhelmingly male. There are some things that are too important to be broken, like deciding a patient's treatment options or an inmate's future. There are some choices that are nuanced by more than statistical factors like cultural conditioning and systems of oppression. Making AI decisions for the masses should involve diverse teams applying empathetic and future-oriented design principles. Even with all of its positive potential and far reaching capability, AI is still a tool. That tool should work well for every person that it impacts. We must constantly correct our assumptions and check the biases that we build into these products if we truly want AI to transcend our collective knowledge.

Danielle Beecham

Stanford Radiology Alumni
UX Design Consultant and MFA Design and Technology Candidate
Parsons School of Design

Danielle recently completed her training at graduate school at Parsons School of Design. Danielle's thesis is exploring ways that tech can support low-income patients in their search for affordable resources.
Understanding, assessing, and mitigating bias is one of the most important issues in machine learning today. Ensuring that models perform well on under-represented populations and do not fail catastrophically on rare cases is critically important to effective and equitable deployment of these techniques in practice. On the other hand, machine learning models are often reliant upon large amounts of labeled data, which carry an inherent bias from the human labeler. Given recent movement towards weaker, noisier forms of supervision, where humans provide higher-level inputs to supervise machine learning models, we have an opportunity to design systems that are not reliant upon static datasets, but rather can be adjusted and tweaked in a tight iteration loop that will help address issues of systematic bias in performance on various data subsets. While there remains much work to be done, there has been substantial progress towards integrating notions of fairness into various AI algorithms, and future of this area of work looks both promising and intensely interesting.

Jared Alexander Dunnmon, PhD
Postdoctoral Research Fellow, Rubin Lab
Stanford | Computer Science
With the hype of deep learning, Artificial Intelligence (AI) research is becoming increasingly focused on applying machine learning techniques to solving complex healthcare problems. Commercial giants are rushing to push AI tools into medical decision making. In my opinion, the idea that physicians and algorithms are designed to function independently is a main repulsive factor why AI is not widely adopted in daily clinical practice.

Being a computer science researcher, my job is to teach AI models to make predictions and diagnoses from large amounts of patient data, by learning their own associations. I developed models that can predict short-term survival of metastatic cancer patients, progression of disease, imaging outcomes, and treatment effects. To be reliable, these data-hungry algorithms need humongous amounts of data for training and, given the fact that protected health information (PHI) is not easy to access and share, most of the time the models learn from data acquired from a single or no more than three different healthcare systems.

Being trained in such a restricted knowledge space, AI algorithms applied to clinical decision support may be subject to important biases based on racial, socio-economical and disease disparities within the available database, among others. Bias in the AI algorithm is not limited to healthcare. For example, AI language models which are used to facilitate website searches and machine translation, reflect societal biases as well. For example, associating searches for jobs with the search terms “female” and “woman” leads to suggestions for openings in the arts and humanities professions, whereas searches that include the terms “male” and “man” leads to suggestions of math and engineering occupations. Another example is an AI model that was designed to aid judges in sentencing prisoners by predicting an offender’s risk of recidivism, which has shown an unnerving propensity for racial discrimination.

From my healthcare AI model building experience, whenever we train a model only on data from the Stanford population, it is always hard to generalize on the entire population in North America and beyond. But my question is: do we really need the AI model to generalize on each-and-every human being on the earth? Given that the geographical distribution of diseases and best treatment practice is also a variable factor, it is close to impossible to make AI work on every single patient in every environment.

Humans adjust to their local environment as well. In contrast with highly-trained physicians with multiple years of learning experience, training an AI model is somehow analogous to teaching a child a new language. If she knows only English, she will not be able to speak with her new French friend immediately. But with time and proper training, she may be able to communicate with her new friend. Similarly, trained only on the Stanford population, my AI models are currently generating good performance on Stanford as well as other healthcare systems. But they may not be immediately suitable for a healthcare system in the Midwest, because they have not been exposed to data from that area. However, with proper re-training and additional information, they can work on a more diverse environment. AI models are only as good as the data that we used to train them. I believe that bias can be tamed and that the AI user can tackle bias properly by joining data from different populations and environments. Sharing data will be highly beneficial for AI advancement.

Thus, one of the most important considerations for introducing AI into a healthcare system should be “knowing your AI”: Understand who trained it and which data were used to train your AI system, including geographical and population factors, such as age, race and clinical conditions. Apply the model accordingly.

Imon Banerjee, PhD
Instructor
Stanford Medicine | Radiology

Artificial Intelligence and Bias
Rapid advances in machine learning (ML) over the past decade have led to burgeoning interest in its applications to medical imaging. While the impact of ML on radiology is just beginning to be felt, in many cases these technologies have already been incorporated in other facets of our society – arguably to greater effect. A recent Wall Street Journal editorial by Christopher Mims (March 23rd, 2019 edition) explores the implications of ML technologies on our everyday lives. Algorithms are presently being used to verify our identities through facial recognition, to judge our creditworthiness using financial records, and to calculate the cost of health insurance based on our medical histories. Algorithms such as COMPAS (Equivant) have even been used to determine the severity of sentencing after someone is convicted of a crime. As applications of ML technologies continue to expand, many have begun to scrutinize not just the utility of these algorithms, but also their “fairness” – and whether they are susceptible to the same underlying biases as their human creators.

The potential for bias to impact machine learning has not been thoroughly explored in the past, possibly because ML technologies were developed to answer academic questions, or because their generalizability outside specific tasks (such as recognizing written characters) was limited by available computing power. Recent computational advances have enabled development of robust algorithms that are extraordinarily complex – for example, Google’s Inception-v3 network has more than 20 million free parameters. The complexity of these modern networks has led to the requirement that they be “trained” with very large amounts of data to achieve reasonable performance. It is in part through selection and annotation of training data that bias can become incorporated into these algorithms. For example, algorithms trained using image data exclusively from US hospitals may completely neglect factors contributing to pathology that are common in other parts of the world. Binary classification algorithms, which may seem less susceptible to bias when they rely on identification of unequivocal image features, can potentially become biased if their sources of training data are not sufficiently diverse. Even algorithms themselves can become susceptible to bias if they fail to take into account factors that influence susceptibility to disease, such as genetic background, socioeconomic status, or local geography.

Outside of medical imaging, the applications of modern ML algorithms have been magnified by a deluge of data available as a result of lax privacy regulations in many countries – including web browsing histories, financial transactions and in some cases health histories. The availability of these data presents many opportunities for exploitation by entities whose intentions may not be benevolent. Consider the Chinese Social Credit System, which has been enabled by ML technologies such as facial recognition and is already being used by the Chinese government to regulate the behavior of its citizens. Even in the US, there have been instances where a biased algorithm has led to the rights of certain groups of people being curtailed. For example, Facebook was recently sued by the Department of Housing and Urban Development regarding practices which enabled its clients to engage in housing discrimination.

Given these advances, the moment to consider the impact of bias on machine learning has very much arrived. ML is clearly susceptible to bias, not only in the selection of data used to train ML algorithms, but in the design of algorithms themselves. Over the next decade, as applications of machine learning continue to expand, it will be incumbent on practitioners of ML to understand the potential for bias to influence their work. However, taking action to limit bias in ML is not an easy task, as no straightforward metric to measure bias exists. A first step towards mitigating bias might be to increase transparency by making algorithms and their training data publicly available. However, this is unlikely to occur outside of academia, as corporations utilizing ML will aim to protect their proprietary technologies. Furthermore, governmental / legal regulation of bias in ML is likely to be challenging, as complexity of modern ML-based algorithms presents a significant barrier to thorough assessment by regulators. Awareness of the potential impact of bias on ML will thus be of paramount importance going forward. If the stewards of ML lose sight of this task, we run the risk that biased algorithms will begin to impact the many facets of our society where machine learning plays a role – potentially resulting in marginalization of its most vulnerable members.

Eric Bultman, MD, PhD
Resident
Stanford Medicine | Radiology
“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, . . .”
– Charles Dickens, A Tale of Two Cities

As AI systems are being integrated into everyday life, intrinsic, often unintentional, biases are being discovered. Prominent companies have stumbled in their implementation of AI platforms across multiple sectors. Not long ago, Google's image recognition algorithm was unable to recognize African American males as humans. Voice recognition software still demonstrates poor performance in understanding women and people with foreign accents. Criminal recidivism risk assessments such as COMPAS, a tool commonly used by law enforcement throughout the United States, have been shown to discriminate against African Americans, who were twice as likely to be labeled as high-risk re-offenders compared to their white counterparts. These cautionary tales put a damper on the enthusiasm that is often encountered when considering the tremendous potential AI has in transforming the human experience.

Such public blunders have ignited efforts to combating bias in AI. The fundamental tenants to overcoming bias lie in embracing diversity and inclusivity. Ubiquitous integration of these principles will be key in the evolution of AI. One commonly referenced area of improvement is in the diversity of training sets. The adage “garbage in, garbage out” reflects AI's vulnerability to not only propagating existing biases, but at times, even amplifying them. For example, Amazon's well-intentioned efforts to develop an automated hiring and recruitment system fell flat when it was discovered to be biased against women. The program that was trained using a decade of prior applications consistently rated male candidates higher than females, unwittingly perpetuating the bias that it was designed to combat. Similarly, facial recognition software has been criticized for its poor performance in identifying females, particularly those with dark skin. This should come as little surprise when training sets are primarily composed of light-skinned males. In light of such pitfalls, companies like Google and IBM have been amassing large image datasets with increased representation of minority groups. Construction and utilization of datasets must continue to be mindful of inclusivity. Moreover, encouraging transparency by researchers and companies on the training data used in creation of AI platforms is important when considering the potential biases in the platform's eventual deployment.

An AI workforce reflective of the diversity of thought, backgrounds, and perspectives present globally is desperately needed. A mere 22% of AI professionals are women, according to the World Economic Forum. Such underrepresentation has pervasive consequences, particularly when considering such a powerful and ubiquitous technology. The lack of diverse stakeholders within AI teams leave behind entire communities as others are flooded by advancements. Diverse teams have been repeatedly demonstrated to improve group performance on problem solving tasks, enhance creativity and innovation, and raising collective intelligence. The disparity present in the AI workforce has sparked the creation of multiple organizations dedicated towards recruiting and training diverse individuals, including AI4ALL, a nonprofit geared towards young high schoolers, and Women in Big Data, a forum to support and attract female talent to big data and analytics. Other companies have chosen to battle the implicit biases present in the recruitment process by developing new AI platforms, like Textio, a startup that analyzes job descriptions to identify language that would alienate potential candidates and offers alternative engaging language. The importance of continued intentional efforts to include diverse populations within the AI workforce cannot be understated.

As the bold optimistic pursuit of advancing humanity with AI continues, mindful early integration of safety nets that improve inclusivity and diversity will become integral in ceasing the perpetuation of bias.

Katie Shpanskaya, MD candidate
Medical Student | Stanford Medicine
AI Researcher | Radiology
From Elon Musk’s musings of the AI singularity to presidential candidate Andrew Yang’s propositions to allay the effects of unemployment secondary to automated manufacturing and driving, artificial intelligence is a persistent and increasing topic of conversation. Deep learning through artificial intelligence has streamlined marketing efforts, promised to bring safer vehicles to the roads and provide personalized Netflix recommendations. Thus, it is unsurprising that machine learning has found its way into medical research labs, particularly those in the Silicon Valley.

I am fortunate to be a part of one such lab, Dr. Heike Daldrup-Link’s, where we have members from varying backgrounds whose origins span from North America to Europe to Asia. Despite the diversity in our lab, though, the data we provide our algorithm in order for it to “learn” is dependent upon the demographics of the patients we have selected, not our own backgrounds. Throughout medical training, I have been taught to “treat the patient, not the disease.” However, it is difficult to account for this variable when it comes to AI training due to specific demographic variables in a given patient population. This potentially causes problems for the detection of disease processes, which are unique to or more prevalent in specific populations. For example, if the patients we would select for the training phase of our tumor detection algorithm were predominately of Scandinavian descent, what would the algorithm think of bone infarcts or a posterior mediastinal mass related to extramedullary hematopoiesis in an African American sickle cell patient?

Solutions to this problem do exist. Feedback systems can be incorporated into the final products so that practicing radiologists can mark for review studies, which are incorrectly diagnosed by the software so that the algorithm can be retrained. Harking back to Elon Musk and Andrew Yang’s apprehension with the takeover of artificial intelligence, though, I suppose the question I have is: Should our algorithms get to a point where they can correlate imaging findings with patient demographics and presenting symptoms such that we can come to a more refined diagnosis? Or will we maintain our software’s handicap of “color blindness”?

Alimohammad (Ali) Moalem, MD
Resident
Stanford Medicine | Radiology
In a recent fascinating work originating from the University of Wellington, researchers were attempting to train a type of AI, termed Generative Adversarial Networks (GANs), to learn increasing levels of smiles, using pictures of celebrities. A GAN learns the data in such a way, that it becomes able to mimic the pictures with realistic accuracy, and create pictures of new never-before-seen “celebrities”, with varying levels of smile. To the researchers surprise, the new smiling celebrities always looked female, and the less happy faces were always male. What was going on? Was the AI saying that women were happier? Well, almost. In the original dataset, the pictures were imbalanced such that female celebrities were more likely to smile, and therefore the AI learned a gender-bias, attributing female features to the human smile. In other words; the AI just didn't know any better, and only had the data at hand to learn from.

In a modern healthy society, we are often taught about fairness, equality, and the potential harms of favoritism and prejudice throughout our upbringing and schooling. Yet, despite the best of intentions, we humans tend to very easily deviate from rationality in judgement, and unwillingly introduce cognitive biases in our everyday decisions and interactions. In fact, to such an extent that human- and cognitive biases is a large field of study in both Behavioral Psychology where researchers will try and understand the unintentional behavior, and in Consumer Marketing where glitches in rational decision-making are actively exploited.

Coming back to the realm of AI, there is a lesson to be learned. We know, that the learning algorithms of AI are purely statistical written in a strict mathematical formalism with no intentions of creating unfairness. Yet bias is so easily created, by simply having the AI observe the current human world (i.e. dataset), where imbalance and inequality already exists. Humans tend to do the very same, and what’s dangerous is that we often don’t realize it. We seldom actively question whether our intuitions, and beliefs of the world are biased, or if we subconsciously observed and manifested a skewed world-view. We rest comfortably in the assumption, that as long as we don’t want to be biased, we won’t become it.

The bias in AI can be corrected by balancing the dataset; i.e. assembling an equal number of pictures with smiling men and women. Unfortunately, the distribution of real-world data cannot simply be changed from one day to another. Instead, we as humans must make active decisions of questioning our beliefs and intuitions, recognize biases in ourselves, and change our skewed patterns. Ultimately, that may build for a more balanced society.

Anuj Pareek, PhD
Postdoctoral Research Fellow, Daldrup-Link Lab & Rubin Lab
Stanford Medicine | Radiology
A popular question nowadays is if artificial intelligence (AI) can become the equivalent of a human mind. While AI holds tremendous promise for technological and medical breakthroughs, the current fascination about whether AI can have human biases is intertwined with our definition of what it means to be human. To be human isn't just about being intelligent, but also about being humane.

When was the last time you saw an animal, or an algorithm, behave in a way that suggested dissatisfaction over the quality of its own existence? I’m not talking about the ritualistic mourning of kin who have died; animals also do that. I'm referring to the need to know that our lives have meaning, whether in the form of pleasure, comfort, power, delicious food, or prestige. All of these point to the human need for meaning.

On LinkedIn, an AI scientist posted a question about how he could prove to online learners that he was human and not an AI that was trained to teach AI. While this is a great technical question for scientists to wrestle with, it challenges us to re-examine our understanding of what it means to be human.

Below was my response:

“Prove through personal example that answers to existential questions (What is the meaning of life? Where do I come from? Why does it even matter?) can cause you to go into depression if not satisfactorily answered for you. And, show that if/when you become dissatisfied by the answers that you find — or don't find — your software and hardware start to malfunction.”

If our definition of human intelligence involves only the ability to do complicated computations, then the technology we have is already human. I’ll believe that a machine is human when its software system — its “mind” — can reflect on itself, producing emotions such as joy or sadness, satisfaction or dissatisfaction. Why do we have hospitals and psychiatric clinics? Why do institutions need diversity newsletters? Why are you even reading this? Because consciousness is more than intelligence.

David H. Nguyen, PhD
Visiting Scholar
Stanford Medicine | Radiology
Computational Scientist
Tissue Spatial Geometrics Lab (www.TSG-Lab.org)
How often are you the only female in the room? As a radiologist, imaging informaticist, and former mathematician, it has been a frequent experience for me. So much so that it has become my norm. But it doesn't need to be...

Considering the number of women choosing radiology as their medical specialty, my experience is not a surprise. Despite nearly equal percentages of men and women in medical schools, men outnumber women in radiology residency by almost 4 to 11. Similarly, men outnumber women in radiology practice by almost 5 to 11. In radiology leadership, men outnumber women by almost 8 to 11. And this trend is not new. The chronic underrepresentation of women in radiology and radiology leadership has led to a dearth of female role models for young women – a fact that may contribute to the continued gender disparity in our profession.

Dr. Geraldine McGinty and Mini Peiris shared a similar experience. Too often in their careers they were the only women in the room. With a shared commitment to change the status-quo, they formed RADxx. Founded in November 2016, RADxx is an organization focused on fostering networking and mentorship opportunities for women in radiology, informatics and IT management of radiology systems. To that end, RADxx is creating a list of quality women speakers. This “Speakers Bureau” will be unveiled on June 18, 2019. If you know someone who should be included, please email their name and email address to Catherine Slotnick at cslotnick@ambrahealth.com.

Multiple female speakers were highlighted at the recent 2019 Stanford Radiology Improvement Summit. Jennifer Broder presented an insightful session about “Creating Just Culture,” Bettina Stewart presented “Communication across Authority Gradients,” Joyce Liang, a colleague of mine at Radiology Partners did a fantastic job with first official presentation titled “recoMD: Decision Support for Radiologist Recommendations,” to name a few. All of these presentations along with those from the other great speakers at the conference are available on the Stanford Medicine website at http://med.stanford.edu/radisummit/2019ProgramAgenda.html.

I am proud to be a member of RADxx. Our tweet chats have seen over 4.5M impressions and 250 individuals attended the last RADxx networking event. Together with other female and male RADxx members, we are working to increase gender diversity in radiology through mentorship, coaching and advocacy.

Join the discussion at #RADxx and the community at https://radxx.ambrahealth.com/. Together we can make a difference.

Nina Kottler, MD, MS
VP Clinical Operations
Radiology Partners

“AI is likely to be either the best or worst thing to happen to humanity.”

– Stephen Hawking