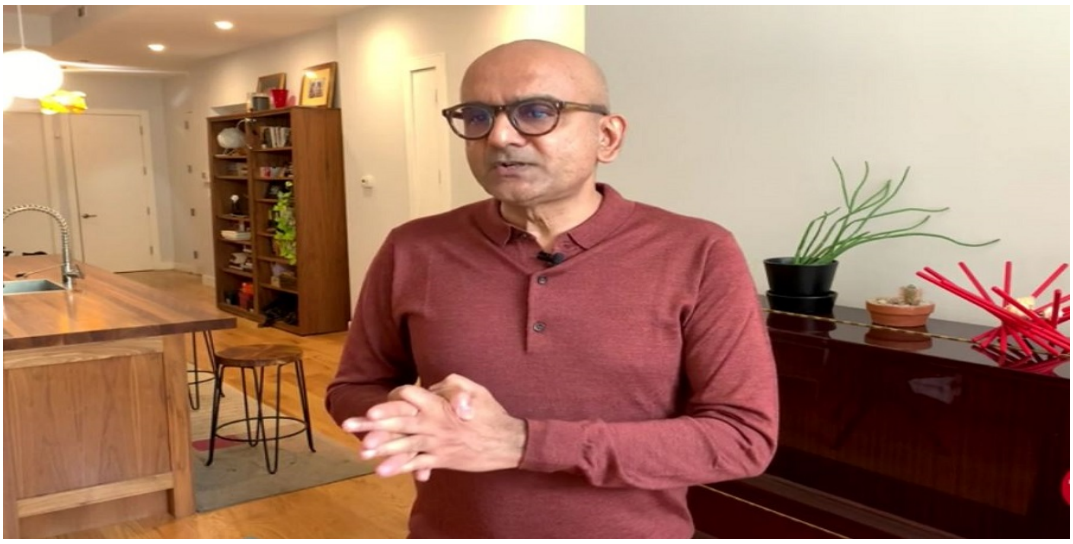
How bad data keeps us from good AI

劣质数据如何让我们远离优质AI?

2023-07-07



AI could add 16 trillion dollars to the global economy in the next 10 years. This economy is not going to be built by billions of people or millions of factories, but by computers and algorithms. We have already seen amazing benefits of AI in simplifying tasks, bringing efficiencies and improving our lives. However, when it comes to fair and equitable policy decision-making, AI has not lived up to its promise. AI is becoming a gatekeeper to the economy, deciding who gets a job and who gets an access to a loan. AI is only reinforcing and accelerating our bias at speed and scale with societal implications. So, is AI failing us? Are we designing these algorithms to deliver biased and wrong decisions?

人工智能（AI）会在接下来的十年为全球经济注入16万亿美元。这样的经济不是由几十亿人或者几百万个工厂建造，而是由电脑和算法创造。我们已经看到了AI的惊人优势：简化任务、提高效益、改善生活。然而，谈到政策制定的公平公正时，AI并没有兑现承诺。AI成为了经济的守门人，决定谁能拿到工作，谁能拿到贷款。AI只是在大规模和迅速地巩固和加速我们对社会影响的偏见。那么，AI辜负我们了吗？我们设计这些算法是为了输出带有偏见的错误决定吗?

As a data scientist, I'm here to tell you, it's not the algorithm, but the biased data that's responsible for these decisions. To make AI possible for humanity and society, we need an urgent reset. Instead of algorithms, we need to focus on the data. We're spending time and money to scale AI at the expense of designing and collecting high-quality and contextual data. We need to stop the data, or the biased data that we already have, and focus on three things: data infrastructure, data quality and data literacy.

作为一个数据科学家，我可以告诉你，不是算法的问题，而是偏置数据，它们得对这些决定负责。为了让AI可以被人类和社会使用，我们急需一次重置。我们该关注的不是算法，而是数据。我们花了大量时间和金钱，以设计和收集高质量的背景数据为代价，扩大AI的规模。我们需要停下这些数据，或我们已有的偏置数据，去关注三件事：数据基础设施、数据质量和数据素养。

In June of this year, we saw embarrassing bias in the Duke University AI model called PULSE, which enhanced a blurry image into a recognizable photograph of a person. This algorithm incorrectly enhanced a nonwhite image into a Caucasian image. African-American images were underrepresented in the training set, leading to wrong decisions and predictions. Probably this is not the first time you have seen an AI misidentify a Black person's image. Despite an improved AI methodology, the underrepresentation of racial and ethnic populations still left us with biased results.

今年六月，我们看到了一个杜克大学的AI模型，PULSE，产生了令人尴尬的偏差，它增强了一张模糊的图像，使它呈现了一张可辨人像。这个算法错误地将一张非白人图像增强至了一张白人像。在训练集中，没有足够的非裔美国人的图像，导致了错误的结果和预测。也许这不是你第一次看见AI误判黑人的图像。即使AI方法已经改进过了，对有些种族的忽视依旧会产生有偏差的结果。

This research is academic, however, not all data biases are academic. Biases have real consequences.

研究是学术性的，但是，不是所有数据偏差都是学术性的。偏差有真实的后果。

Take the 2020 US Census. The census is the foundation for many social and economic policy decisions, therefore the census is required to count 100 percent of the population in the United States. However, with the pandemic and the politics of the citizenship question, undercounting of minorities is a real possibility. I expect significant undercounting of minority groups who are hard to locate, contact, persuade and interview for the census. Undercounting will introduce bias and erode the quality of our data infrastructure.

以2020年美国人口普查为例。人口普查是很多社会和经济政策制定的基础，所以人口普查应当记录美国人口的100%。然而，由于新冠疫情和国籍背后的政治问题，实际上非常有可能少算了少数群体。我预计会有大量少数群体的少计，很难定位、联系、说服、采访他们加入人口普查。少计会带来偏差，威胁我们数据基础设施的质量。

Let's look at undercounts in the 2010 census. 16 million people were omitted in the final counts. This is as large as the total population of Arizona, Arkansas, Oklahoma and Iowa put together for that year. We have also seen about a million kids under the age of five undercounted in the 2010 Census.

我们来看2010年人口普查的少计。在最终统计中，有1600万人被忽略了。这相当于亚利桑那州、阿肯色州、俄克拉何马州、爱荷华州当年的人口总和。我们还可以看到在2010年人口普查中，少计了大约100万低于五岁的儿童。

Now, undercounting of minorities is common in other national censuses, as minorities can be harder to reach, they're mistrustful towards the government or they live in an area under political unrest.

如今，少数群体的少计在其他国家的人口普查中也很常见，因为很难联系到这些少数群体，他们不信任政府，或者生活在政治动荡区域。

For example, the Australian Census in 2016 undercounted Aboriginals and Torres Strait populations by about 17.5 percent. We estimate undercounting in 2020 to be much higher than 2010, and the implications of this bias can be massive.

比如，2016年的澳大利亚人口普查少算了大约17.5%的澳洲土著和托雷斯海峡人口。我们估计2020年的少计会远远超过2010年，这种偏差可能造成的影响是深远的。

Let's look at the implications of the census data. Census is the most trusted, open and publicly available rich data on population composition and characteristics. While businesses have proprietary information on consumers, the Census Bureau reports definitive, public counts on age, gender, ethnicity, race, employment, family status, as well as geographic distribution, which are the foundation of the population data infrastructure. When minorities are undercounted, AI models supporting public transportation, housing, health care, insurance are likely to overlook the communities that require these services the most.

我们来看人口普查数据的影响。人口普查是最权威的、可公开获取的丰富数据，它记录了人口组成和特征。虽然商家会对顾客收集私密信息，但是美国人口调查局会报告完整的、公开的年龄、性别、民族、种族、就业情况、家庭情况信息，还有地理分布情况，这是人口数据基础设施的基础。当少数群体被少计时，支撑公共交通、住房、医疗、保险的AI模型更可能忽视最需要这些服务的群体。

First step to improving results is to make that database representative of age, gender, ethnicity and race per census data. Since census is so important, we have to make every effort to count 100 percent. Investing in this data quality and accuracy is essential to making AI possible, not for only few and privileged, but for everyone in the society.

改进结果的第一步就是让每一次人口普查数据的年龄、性别、民族、种族数据库具有代表性。既然人口普查如此的重要，我们就要不遗余力地让它记录100%。在数据质量和准确率上投入精力对AI的可行性至关重要，不但是为了少数有特权的人，还是为了社会中的每一个人。

Most AI systems use the data that's already available or collected for some other purposes because it's convenient and cheap. Yet data quality is a discipline that requires commitment -- real commitment. This attention to the definition, data collection and measurement of the bias, is not only underappreciated -- in the world of speed, scale and convenience, it's often ignored.

大多数的AI系统利用已有的或者为其他目的收集的数据，因为很方便又便宜。但是数据质量是一个需要大量投入的领域，真正的投入。对偏差的定义、数据采集和偏差测量的关注在快速、大规模、便利的世界不仅不受重视，还会被无视。

As part of Nielsen data science team, I went to field visits to collect data, visiting retail stores outside Shanghai and Bangalore. The goal of that visit was to measure retail sales from those stores. We drove miles outside the city, found these small stores -- informal, hard to reach. And you may be wondering -- why are we interested in these specific stores? We could have selected a store in the city where the electronic data could be easily integrated into a data pipeline -- cheap, convenient and easy. Why are we so obsessed with the quality and accuracy of the data from these stores? The answer is simple: because the data from these rural stores matter. According to the International Labour Organization, 40 percent Chinese and 65 percent of Indians live in rural areas. Imagine the bias in decision when 65 percent of consumption in India is excluded in models, meaning the decision will favor the urban over the rural.

作为尼尔森公司（Nielsen）数据科学组的一员，我实地走访收集了数据，拜访了上海和班加罗尔之外的零售店。拜访的目的是测量这些店铺的零售业绩。我们开出了市区，发现了这些小店——不正规、交通不便利。你可能在想为什么我们会对这些店感兴趣？我们可以在市区找一家店，电子数据可以轻松地接入数据流程，便宜、方便又简单。为什么我们对这些店的数据质量和准确率如此感兴趣？答案很简单：因为这些乡镇小店的数据很重要。国际劳工组织表示，40%的中国人和65%的印度人生活在乡村。想像印度消费的65%被排除在模型之外，政策制定的偏差会怎样，意味着这些决策会偏向对城市比对乡村更为有利。

Without this rural-urban context and signals on livelihood, lifestyle, economy and values, retail brands will make wrong investments on pricing, advertising and marketing. Or the urban bias will lead to wrong rural policy decisions with regards to health and other investments. Wrong decisions are not the problem with the AI algorithm. It's a problem of the data that excludes areas intended to be measured in the first place. The data in the context is a priority, not the algorithms.

没有这些城乡背景信息和民生、生活方式、经济和商品价值的信号，零售品牌会在定价、广告和营销上做出错误的决定。倾向城市的偏差会导致错误的乡村政策制定，包括医疗和其他方面。错误的决定不是AI算法的问题。这是数据的问题，排除了那些早就该被统计的地区的数据。这些背景数据才该被首先关注，而不是算法。

Let's look at another example. I visited these remote, trailer park homes in Oregon state and New York City apartments to invite these homes to participate in Nielsen panels. Panels are statistically representative samples of homes that we invite to participate in the measurement over a period of time. Our mission to include everybody in the measurement led us to collect data from these Hispanic and African homes who use over-the-air TV reception to an antenna. Per Nielsen data, these homes constitute 15 percent of US households, which is about 45 million people. Commitment and focus on quality means we made every effort to collect information from these 15 percent, hard-to-reach groups.

我们来看另一个例子。我拜访了一些俄勒冈州偏远的房车营地，和纽约的公寓，邀请这些家庭加入一些尼尔森论坛。我们在这段时间邀请了一些家庭参与论坛的统计，数据样本具有代表性。我们的目标是邀请所有人加入统计，让我们可以收集这些西班牙裔和非洲裔家庭的数据，他们都会收看有线电视。根据尼尔森的数据，这些家庭占美国家庭的15%，大约4500万人。对质量的投入和关注意味着我们要尽全力从这15%难以触及的人群收集信息。

Why does it matter? This is a sizeable group that's very, very important to the marketers, brands, as well as the media companies. Without the data, the marketers and brands and their models would not be able to reach these folks, as well as show ads to these very, very important minority populations. And without the ad revenue, the broadcasters such as Telemundo or Univision, would not be able to deliver free content, including news media, which is so foundational to our democracy.

为什么这很重要？这是个人数可观的群体，对市场人员、品牌、传媒公司都非常非常重要。没有这些数据，市场人员、品牌和他们的模型无法接触到这些人，也无法对这些非常非常重要的少数群体做广告。没有广告收入，像Telemundo和Univision这样的电视台就无法播送免费的内容，包括新闻媒体，我们民主制度的基础。

This data is essential for businesses and society. Our once-in-a-lifetime opportunity to reduce human bias in AI starts with the data. Instead of racing to build new algorithms, my mission is to build a better data infrastructure that makes ethical AI possible. I hope you will join me in my mission as well.

这些数据对商家和社会都十分重要。我们减少AI人为偏差的千载难逢的机会始于数据。我的目标不是争先恐后设计新算法，而是建设更好的数据基础设施，让AI更合乎伦理。我希望你能加入我的使命。

Thank you.

谢谢。