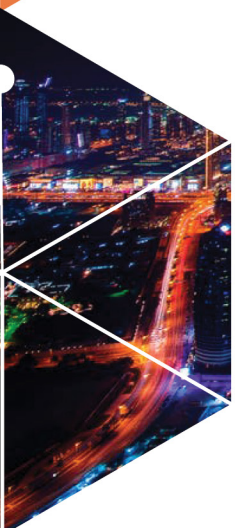
INCLUSIVE AI:

TECHNOLOGY AND POLICY FOR A DIVERSE URBAN FUTURE

JULY 2017

Brandie Nonnecke, PhD
Tarunima Prabhakar, MPP
Chloe Brown, MPA
Camille Crittenden, PhD

CITRIS & the Banatao Institute
University of California, Berkeley



CITRIS AND THE BANATAO INSTITUTE

Microsoft

This white paper and a corresponding public symposium were co-sponsored by Microsoft and CITRIS and the Banatao Institute.

The public symposium “Inclusive AI: Technology and Policy for a Diverse Urban Future” was held at CITRIS and the Banatao Institute on May 10, 2017.
Video from the symposium is available at tiny.cc/inclusiveai.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Artificial Intelligence (AI) holds great promise for governments and their citizens. For city leaders, AI-enabled technologies may provide decision support to increase efficiency and equity in the delivery of public services and resources, identify emerging opportunities and risks, and enable targeted interventions. This white paper focuses on the urban context, providing examples of AI for policing and law enforcement, labor and workforce development, and public services.

While AI holds great potential to improve civic domains, these technologies can exacerbate negative effects when they reinforce social biases and inequalities—whether by design or unintentionally. Far from remaining a "virtual threat," the consequences of ill-considered algorithms can have deleterious effects in the real world. In law enforcement, these tools can increase patrolling and biased legal decisions against protected groups. In the labor sector, they can inadvertently increase existing workforce disparities and disproportionally increase unemployment. In public services, biased decisions can lead to inequitable allocation of resources and social disenfranchisement.

Our understanding of the full benefits and risks of AI-enabled technologies on social, political, and economic inclusion in the urban context will continue to evolve with technological advancements. Yet, city leaders can better prepare themselves to take advantage of the benefits of AI while minimizing potential risks by carefully evaluating the data it collects on its citizens, improving the quality of data collection, and educating their workforce and residents about the underlying principles and implications of AI-enabled decision-making. Before incorporating AI, tradeoffs in transparency, efficiency, effectiveness, and human agency should be considered and critically evaluated. In these incipient stages, detecting and remediating false classifications and decision errors must be prioritized to better ensure that the benefits and risks of AI are more equally distributed across society.

INTRODUCTION

Half the world's population currently lives in cities. By 2050, it will be nearly 70%.[1] This influx will put substantial strain on public service agencies, necessitating adoption of innovative approaches and systems. Cities generate vast amounts of data from diverse populations and devices. These data offer opportunities to utilize Artificial Intelligence (AI)—algorithmic models that enable machines and systems to automate decisions and processes—to support the efficient and equitable distribution of public resources and services, reveal emerging opportunities and risks, and inform targeted intervention strategies. However, these innovations also risk unintended consequences due to false predictions, errors, and biased decision-making.

This paper presents examples of AI's application in three key sectors in urban governance: safety and law enforcement, labor and workforce development, and public services. These examples highlight the benefits as well as the complicated cultural, social, political, and economic effects of algorithmic decisions in the urban environment. This paper details the influence of human subjectivity in the design and impact of AI-enabled technologies, including the potential for discrimination based on education, income, gender, race, age, ethnicity, ability, creed, and sexual orientation. We conclude with recommendations for city leaders and public service agencies incorporating AI tools so they may better manage the emerging risks while benefiting from AI's potential to make their processes more transparent, efficient, and inclusive.

DECODING ARTIFICIAL INTELLIGENCE

While the concept of AI has existed since the 1950s, recent advances in computational power, data availability, and high-speed networks have increased its application for a range of tasks, including image recognition, domain-specific risk-prediction, and navigation. Early AI developments "tackled problems that were intellectually difficult for humans but relatively straightforward for computers—problems that could be described by formal, mathematical rules."[2] Current state-of-the-art techniques aim to replicate more complex and intuitive human problem-solving capabilities, presenting a unique set of ethical dilemmas. These technologies will be realized far into the future and are not the focus of this paper. Rather, we focus on the benefits and challenges from recent developments in 'narrow AI,' designed to solve specific tasks such as risk prediction, speech recognition, and facial recognition.

AI-enabled technologies can increase human ability to perform tasks more efficiently and accurately, complementing rather than replacing human decision-making. UC Berkeley's Professor Ken Goldberg has urged that collaboration between diverse groups of people and machines can lead to better problem solving, a concept he calls "Multiplicity."[3] In fact, human-machine teaming has already been shown to be more reliable than what humans or machines could achieve on their own. For example, a study that provided physicians with images of lymph node cells and asked them to determine whether the cells were malignant found that an AI-based approach had a 7.5 percent error rate whereas a human pathologist had a 3.5 percent error rate; a combined approach, using both AI and human input, lowered the error rate to 0.5 percent, representing an 85 percent reduction in error.[4]

> "Multiplicity is collaborative instead of combative. Rather than discourage the human workers of the world, this new frontier has the potential to empower them."
>
> – Ken Goldberg, Professor, UC Berkeley and Director, CITRIS People and Robots Initiative

AI systems, like all other computation systems, execute tasks from sets of instructions, or algorithms. Over the last sixty years, a variety of algorithms have been invented to replicate intelligent behavior. Some recent algorithms do not provide specific rules for decision-making

but rather give instructions for machines to develop their own rules. Such algorithms are categorized as machine learning.

Machine Learning for Policy Decisions

In addition to its clear applications in industry and healthcare, machine learning can support many complex public policy decisions. Machine learning algorithms find patterns in data to develop a set of rules to explain characteristics of interest. These algorithms attempt to locate a statistical relationship in a dataset and use this relationship to predict outcomes of future queries.[5] In some machine learning algorithms, the designers of the algorithm explicitly decide the factors within the data that could predict characteristics of interest. For example, a researcher may specify that building material, age of construction, and history of fires could be combined to predict likelihood that a building will catch fire. In other cases, the researcher might utilize machine learning to analyze a corpus of data and let the algorithm select features that are statistically most predictive. In the latter case, the algorithm identifies factors or a combination of factors that might not be obvious or even interpretable by humans to predict the likelihood of the building catching fire. In both approaches, patterns are identified from historical (training) data, which is then used to make predictions on future (test) data. The result of an algorithm depends both on the set of rules through which the algorithm learns patterns as well as the data from which it predicts these patterns.

Measuring Accuracy of Predictive Algorithms

Consider a machine learning algorithm deployed by a city fire department that helps identify and target inspections for buildings at risk of fire. The algorithm would use city records and assess different building characteristics, neighborhood, history of fire, and similar criteria to decide whether a building needs inspection. An ideal prediction algorithm would label all buildings not likely to catch fire as safe, and label all buildings likely to catch fire as unsafe, alerting the fire department to prioritize inspection of the latter. In practice, however, the algorithm would label some of the buildings incorrectly. The results of the algorithm could be classified as follows (see Figure 1):

1. A building likely to catch fire labeled as high risk (i.e., a true positive, where positive implies that the building is actually at risk of catching fire)
2. A building not likely to catch fire labeled as low risk (i.e., true negative, where negative implies that a building has low risk of catching fire)

3. A building likely to catch fire labeled as low risk (i.e., a false negative result)
4. A safe building labeled as high risk (i.e., a false positive result)

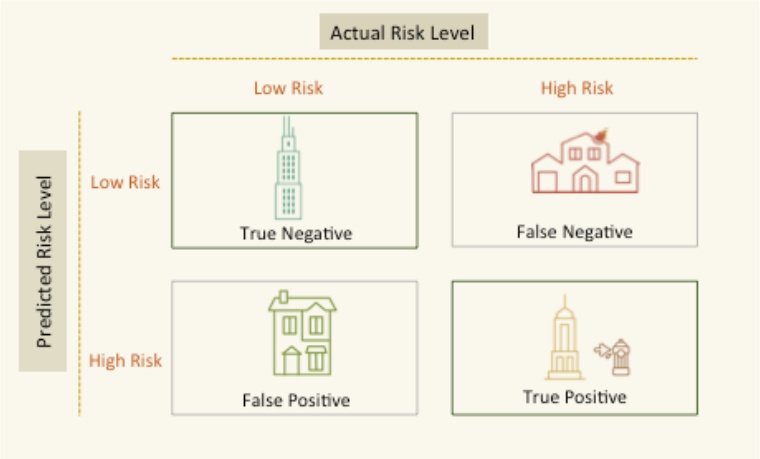


Figure 1. Labeling by Machine Learning Algorithm

In reality however, like any other predictive method, this algorithm can result in erroneous predictions. It might not identify some of the high-risk buildings or might mislabel some of the low-risk buildings (see Figure 1). Algorithms can be adjusted to accommodate different thresholds of false positive and false negative errors. These thresholds must be specified by those designing the algorithm. For example, a city fire department employing a fire risk prediction algorithm might be able to afford inspecting multiple buildings erroneously labeled as high risk for a fire. However, many cities have extremely limited resources, and these false positives could result in an undue burden of human and financial resources. If the rules are set too tightly, one might capture a lot of buildings but setting them too broadly might miss critical high-risk buildings.[6]

In another context, such as an algorithm deployed to assess the possibility of recidivism, a court might want to minimize the possibility of identifying a low-risk person as a high-risk reoffender, even if that comes at the cost of some high-risk individuals not being identified. In this case the algorithm should be designed to minimize false positives. While implementing these algorithms, cities should consider the relative risks of the different kinds of errors.

AI IN URBAN GOVERNMENT: A SURVEY OF THREE SECTORS

The following sections present examples of the application of AI-enabled technologies in law enforcement, labor, and public services. We conclude with recommendations and key considerations for city leaders to help them better identify and manage the potential risks of AI-enabled technologies in urban settings.

Law Enforcement

AI is commonly applied in law enforcement through "risk assessments" that make predictions about a person or area based on algorithmic analysis of underlying factors and data. Law enforcement personnel—including police officers and judges—use these assessments to supplement their decision-making. Stakeholders are often interested in AI because they regard evidence-based algorithms as tools to deploy law enforcement resources more efficiently, effectively, and equitably than using personal judgments or simpler formulas alone. However, these algorithms may pose challenges related to potential discrimination and due process. For example, considering factors such as location, employment, and family characteristics may serve as a proxy for race or income, resulting in disproportionate impact on racial minorities and low-income individuals, while the proprietary nature of most algorithms may violate an individual's constitutional rights (e.g., violations of due process or protections against unreasonable search and seizure).

Predictive Policing

Across the nation, police departments are increasingly using AI in an effort to allocate resources more effectively and reduce potential bias by identifying promising targets for police intervention. Police departments have traditionally relied upon less sophisticated programs using spreadsheets, maps, and manual reviews of information—such as historical crime data, incoming gang/criminal intelligence reports, and criminal records—to identify areas and people at increased risk for being involved in crime. Police then patrol or conduct outreach based on this information, hoping to prevent crimes before they occur. In "predictive policing," police departments use analytical programs that incorporate machine learning algorithms to mathematically extend or automate existing analytical techniques, combining historical and up-to-the-minute crime information to do the work of numerous traditional crime analysts and produce real-time targeted patrol areas for police to conduct interventions (see Figure 2).[7]

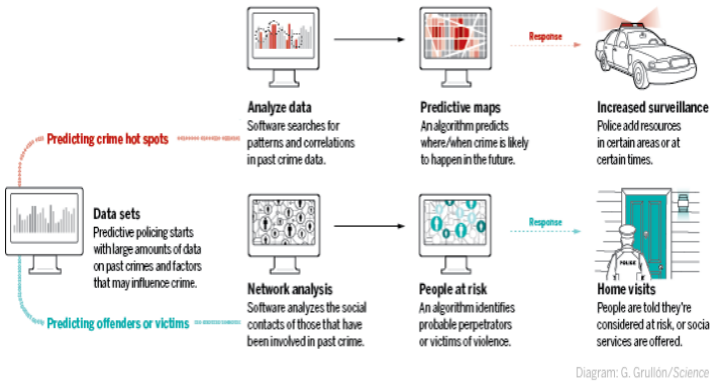


Figure 2. Predictive Policing Models. Source: Science Magazine[8]

Many U.S. police departments use AI-based predictive policing to help identify where and when crimes are likely to occur.[9] These machine-learning models use factors such as historical data on crime type, location, and date and time to generate maps of predicted crime hotspots that change temporally (see Figure 3). As a result, jurisdictions such as Los Angeles and Atlanta report that predictive policing allows them to patrol more effectively and reduce real crime numbers.[10]

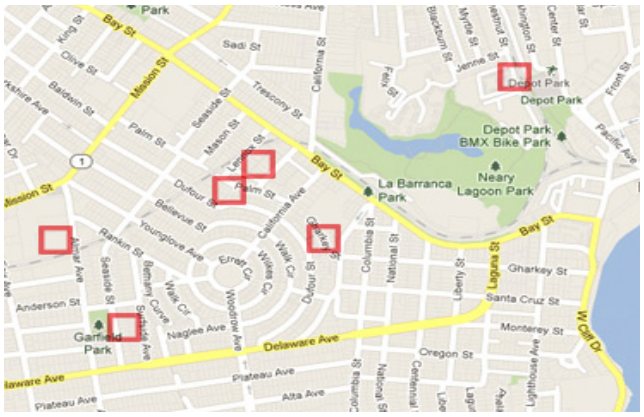


Figure 3. A PredPol map showing predicted hotspots in Santa Cruz, California. Source: PredPol.com

However, concerns remain whether location-based predictive policing programs may discriminate against racial minorities and other residents of low-income areas, and whether they are effective. The historical crime data at the core of predictive policing algorithms reflects reported crimes, an incomplete subset of actual committed crimes and more likely to represent heavily patrolled areas. As a result, using reported crimes to determine patrolling patterns may

perpetuate or amplify racial bias because areas where more people of color live have historically been more heavily patrolled. Furthermore, even if location-based predictive policing is effective at deploying police to prevent crimes, communities may begin to question the legitimacy of police presence when no crime has occurred. Although few external evaluations have been conducted, a RAND evaluation of predictive policing aimed at reducing property crimes in Shreveport, Louisiana, found the program did not significantly reduce such crimes.[11] These concerns, together with public demands to explore alternatives such as community-oriented policing, have driven some police departments, including several in the Bay Area, to decide against using location-based predictive policing techniques.

AI-based predictive policing is also used to identify individuals at increased risk of being involved in crimes, whether as victim or perpetrator. For example, the Chicago Police Department is using machine-learning algorithms for this purpose, enabling officers to target public safety interventions with the aim to turn individuals away from criminal behavior.[12] Similarly, the Fresno, California, Police Department piloted a proprietary algorithm to determine "threat scores" aimed to identify the likelihood that a person calling 911 may pose a threat to patrolling police officers and help the officers tailor their responses.[13] Although such individual-based predictive policing approaches offer potential benefits, they have also raised concerns. For example:

- Because the public—and sometimes the police, as well—knows neither exactly how the algorithms make their determinations nor the extent to which the underlying data may be accurate, civil liberties may be at risk, especially for minorities who are more likely to be targeted.[14]
- If officers use predictions to detain someone without other reasonable suspicion, the detention may violate the individual's constitutional rights.[15]
- The algorithms may be ineffective at predicting who will be involved in crime.[16]
- Even if the algorithms are effective at predicting involvement, related interventions must also be effective to actually prevent crime.[17]

Pretrial Risk Assessments

Courts increasingly use AI to supplement judges' decisions regarding pretrial release and bail bonds. Traditionally, judges have made decisions by formally or informally weighing factors like criminal record, employment status, community ties, and substance-abuse history to predict

recidivism. In contrast, AI can parse massive datasets to determine which factors are actually most relevant to recidivism and make more empirical judgments regarding risk.

AI can be used to supplement judges' decision-making regarding pretrial release in ways that improve results. For example, more than 30 U.S. cities and states now use an algorithmic assessment tool developed by the Laura and John Arnold Foundation to determine whether an individual should be detained or released on bail before trial. Researchers created the tool by analyzing hundreds of factors across a database of over 1.5 million cases drawn from more than 300 U.S. jurisdictions. They identified nine factors that best predict whether a defendant will commit a new crime of any kind, commit a new violent crime, or fail to return to court.[18] In the state of Kentucky and in Lucas County, Ohio, evaluations have shown that use of this assessment has increased the percentage of pretrial defendants released without bail while simultaneously reducing pretrial crime and increasing the percentage of defendants who appear in court as scheduled, without generating unequal results by race or gender.[19] As a result, these courts have been able to use their resources more effectively and efficiently.

However, policymakers should exercise caution as some AI-based risk assessments used in pretrial release decisions may not perform as well when measured by efficiency and equity criteria. For example, Northpointe's COMPAS risk assessment asks defendants to respond to more than 130 survey questions—a process that critics say is both time-consuming and potentially gameable—and then uses a formula to create risk scores.[20] A 2016 study of this assessment showed that although the tool correctly predicted recidivism for about 60 percent of offenders, it mislabeled the remaining 40 percent of offenders in a way that demonstrates racial bias; black defendants were significantly more likely to be labeled as high risk and not re-offend, while white defendants were significantly more likely to be labeled as lower risk and actually re-offend.[21]

Sentencing Risk Assessments

Courts increasingly use AI to supplement judges' decisions and sentencing guidelines regarding front-end sentencing (i.e., determining sentence when convicted of a crime), parole eligibility, and back-end sentencing (i.e., determining sentence for parole violations). Since at least the 1980s, judges have been instructed to make equivalent sentencing decisions for offenders who commit similar offenses and have comparable criminal histories; such decisions involve a judge's professional judgment and often rely upon sentencing guidelines. However, more

recently, some criminal justice systems have begun to incorporate AI-based risk assessments into sentencing in an attempt to use resources more efficiently and relieve prison overcrowding without jeopardizing public safety.[22]

Pennsylvania and Virginia are both using AI in sentencing risk assessments. In 2010, the Pennsylvania state legislature directed the state's sentencing commission to develop a set of algorithms that could supplement front-end sentencing guidelines and judges' decision-making in an effort to balance concerns regarding public safety, fairness, and resource allocation.[23] Specifically, the algorithms identify cases in which a judge may wish to seek more information before determining a sentence because a defendant might have a particularly low or high risk of recidivism, but the algorithms are not used to recommend the sentence to be imposed. Following extensive testing, researchers found that the statistical significance of factors that could be used in a risk assessment varied by category of offense, and therefore created a series of algorithms that feature different factor weights in relation to the type of crime.[24] Four Pennsylvania counties piloted the algorithms in 2016, and early results show that the algorithms may help judges reduce unnecessarily heavy sentences. The state plans to submit the tool for external review and vote on phased adoption in summer 2017.[25] In contrast, Virginia uses risk assessments to adjust sentencing guidelines for non-violent offenders. Rather than using the risk assessment to determine whether judges should request more information before sentencing, Virginia's risk assessments aim to divert low-risk offenders into alternative punishments such as community service or probation, while high-risk offenders proceed with their sentence recommendations unchanged.[26]

However, AI-based risk assessments used in sentencing may also raise concerns about due process and discrimination. For example, in 2014, the U.S. Department of Justice requested that the U.S. Sentencing Commission study the use of data-driven analysis in front-end sentencing and issue policy recommendations due to concerns regarding ways in which such assessments may move sentencing policy away from being based on the crime committed and toward decisions based on group characteristics and likelihood of recidivism.[27] Although experts have identified ways that developers can address potential discrimination due to historical data that differs by race or other protected class, these concerns are now beginning to play out in the courts.[28] For example, a Wisconsin defendant whose judge said he had determined the sentence in part based on the risk assessment score is appealing the ruling. He claims the process violated his due process rights: 1) because the algorithm is proprietary, neither the

defendant nor the court can tell if it uses inaccurate information, and 2) the algorithm makes its decision based, in part, on gender, which is discriminatory.[29] Although the Wisconsin Supreme Court has ruled that judges could use AI-based risk assessments to inform sentencing decisions, in March 2017, the US Supreme Court asked the U.S. Attorney General for an opinion on this case.[30]

> ***"By basing sentencing decisions on static factors and immutable characteristics – like the defendant's education level, socioeconomic background, or neighborhood – [risk assessments in sentencing] may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society…They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place."***
>
> **– U.S. Attorney General Eric Holder[31]**

Community-Police Interactions

Police departments are increasingly applying AI to address challenges like negative community-police interactions. Many police departments use a traditional three-part early warning system.[32] First, the system tracks the number of incidents—e.g., citizen complaints, on-the-job vehicular chases and accidents, uses of excessive force, and civil suits—involving an officer over a given time period, and notifies a supervisor if the number exceeds a set threshold.[33] The supervisor then decides what kind of corrective action to take. Finally, the department monitors the officer's subsequent behavior. In general, departments do not test or monitor the effectiveness of such early warning systems, due to allegations that they may be ineffective or easily manipulated.

AI could be used to create early warning systems for adverse interactions and prevent police misconduct before it occurs. For example, starting in 2015, the Charlotte, North Carolina Police Department began collaborating with researchers from the University of Chicago to process at least 10 years of data on interactions between police officers and the public to predict potential misconduct, allowing supervisors to intervene in risky circumstances.[34] In addition to confirming that past complaints are good predictors of future misconduct, the pilot has identified other factors, including involvement in stressful incidents such as responses to suicide or domestic violence reports. According to researchers, the pilot has been more efficient and accurate than Charlotte's existing system.[35] Specifically, the pilot reduced both false positives and false

negatives for officers at risk of being involved in adverse interactions, and helped the department prioritize interventions. As a result, the Charlotte Police Department has decided to deploy the system force-wide, and similar programs will be piloted in Nashville, Pittsburgh, Los Angeles, and San Francisco. Future pilots may be able to refine best practices for de-escalation by identifying different risk profiles and successful response strategies.

AI could also help incorporate massive amounts of audio and video data increasingly generated by body cameras, which otherwise remain largely inaccessible, into early warning systems by automating analysis. For example, the Oakland Police Department conducted a pilot study with researchers from Stanford University to utilize AI to analyze footage from officer body cameras for broad patterns in community-police interactions, including whether and how these interactions differ by race. According to officials involved in the pilot, the audio footage was first transcribed to text, and then officer language, tone of voice, and other indicators of the content and quality of their interactions were analyzed; researchers carried out the rest of the analysis manually. The study generated recommendations that the department use AI both to identify footage of exemplary interactions involving de-escalation for use in developing materials for officer training, as well as identify footage of negative interactions for integration into early warning systems.[36]

Even in these applications, local policymakers and law enforcement officials will face potential challenges to implementing these tools. Police departments must still obtain and use high quality data on officers and incidents to better ensure non-discriminatory outcomes. Also, the interventions based on the system's risk assessment must still be effective. Such interventions must be implemented early enough that they have the potential to act as true warnings rather than punitive measures, which are more likely to raise opposition from officers and less likely to protect public safety.

Labor

Recent advances in AI, like previous waves of technological changes, will affect the shape and functioning of the modern labor market—changing the nature of work and how people are hired, evaluated, and compensated.[37] AI can enable greater inclusion and introduce consistency into hiring decisions, helping to highlight and correct for employer and workplace biases. Governments as employers can leverage these systems to streamline their own hiring and work processes to support a more equitable work environment and improve the efficiency and effectiveness of their internal processes.

Changes in Employment

Recent advances in AI have renewed concerns regarding effects of automation on employment. While it is difficult to identify the precise effects of AI on employment, the most significant effects will likely be seen in low- and mid-skill level jobs, such as those in transportation, office administration, and agriculture.[38] In 2016, an autonomous truck made its first delivery in Colorado; and autonomous vehicles will likely be used throughout the trucking industry within the next decade, affecting employment rates in an industry that currently employs 3.5 million people.[39,40] The effects of automation will be disproportionate on demographic groups overrepresented in these occupations. For example, the trucking industry employs 4.2% of the black workforce in the U.S., and provides a higher median annual wage than non-driving jobs.[41] In preparing for the effects of automation, governments should consider the disproportionate effects on certain demographic segments and design policies and strategies accordingly.

While AI may automate or replace certain tasks such as customer support and office administration, it will also create new employment opportunities. Advancements in AI have increased demand for skilled workers in fields such as data science, software development, and machine learning. Software development, for example, continues to see labor shortage.[42] Service sector industries are also likely to get a boost. In order to remain competitive, workers will need skills that focus on creativity or personal interactions such as those required in nursing and education.[43] Governments should consider alternative methods for equipping their communities with the skills and resources necessary to deal with the anticipated transformation of the labor market, including development of programs that support continued education and digital skills training.

"Every time we invent something, we make it easier to invent other things using the previous technology."

– Eric Brynjolfsson, Director, MIT Initiative on the Digital Economy[44]

Hiring and Work Management

Online Labor Marketplaces

Online professional talent search platforms such as LinkedIn are now standard tools to connect prospective employers and employees. Companies like Gild use data from LinkedIn and other sites, along with employer data, to find suitable candidates for open positions.[45] As online platforms become the dominant method for looking for jobs, it becomes essential to ensure impartiality in these results. The search results in online platforms are driven by machine learning algorithms that often tailor information to a person's browsing history and known characteristics. A 2015 study from Carnegie Mellon University showed that personalized search recommendations could be discriminatory, finding that Google displayed fewer ads for high paying jobs to female users than their male counterparts.[46]

Online platforms for on-demand labor recruitment such as TaskRabbit and Fiverr have been shown to exhibit bias inadvertently. Since the sites' recommender systems take into account users' feedback when suggesting an individual for a task, negative reviews or lack of reviews can cause individuals to fall in the rankings, reducing their likelihood to be selected. For example, on both TaskRabbit and Fiverr, users were found to leave more negative reviews for black workers. Similarly, people who hired women were less likely to leave feedback on their performance on TaskRabbit, a criterion considered in the platform's job recommender system, decreasing the number of job opportunities presented to female workers.[47]

While the systems are not designed to be discriminatory, recommender systems can reflect the biases of those using these platforms. Algorithms designed to incorporate recommendations and feedback from users into the functionality of the recommender systems could provide biased results. Companies like Airbnb have made efforts to reduce conscious and unconscious bias through their platforms by deemphasizing functionalities such as photographs that explicitly identify platform users.[48] While demographic characteristics may make the online platforms safer and trustworthy, they also increase the opportunity to perpetuate biases observed offline.

Identifying and Mitigating Discrimination in Recruitment and the Workplace

A 2003 study showed that resumes of individuals with Caucasian-sounding names received 50% more callbacks for interviews than resumes with African-American-sounding names even though the resumes were nearly identical.[49] One rationale for introducing automated decision-making into the recruiting and hiring process is to replace subjective human decisions.[49] Startups like Gild, Ideal, and HireVue use machine learning to automate parts of the hiring process with claims of making the process more efficient and equitable. Natural language processing can be applied to screen resumes for relevant skills, decreasing effects of reviewer bias and yielding a more diverse applicant pool. GapJumpers and HireVue use digital assessment tools that are potentially less biased than human judgment to conduct performance evaluations and digital interviews for applicant screening. However, a 2016 White House report claims that "if a machine learning model is used to screen job applicants, and if the data used to train the model reflects past decisions that are biased, the result could be to perpetuate past bias. For example, looking for candidates who resemble past hires may bias a system toward hiring more people like those already on a team, rather than considering the best candidates across the full diversity of potential applicants."[50]

AI can also be applied in novel ways to highlight and address employers' unconscious bias in the workplace. Platforms like Joonko offer digital 'diversity coaches' that allow "organizations to address workplace bias as it occurs."[51] The platform uses data from a company's internal data management system to detect unconscious biases in how tasks are allocated and work is rewarded. The platform also provides suggestions to employers on how to address possible discriminatory behavior. Platforms such as these use AI to close gaps in opportunity for tasks leading to merit and promotion between various demographic groups.

Task Streamlining

With increasing urban population density and decreasing budgets, city governments must formulate new strategies to streamline their operations. AI offers cities an opportunity to make their work more efficient and equitable by optimizing task allocations for their staff and agencies.

Making Public Services Accessible

One of the most evident gains from AI has been in language translation tools. Just as Google automatically translates webpages, AI platforms such as Unbabel can translate business operations into over 14 languages. City governments can use such systems to make their services accessible to the linguistically diverse communities they serve. These tools can help fulfill federally mandated responsibility to ensure non-English speakers can access programs and services that receive federal funding.

AI-enabled “chatbots” can provide the public with quick answers to important service questions, reducing backlogs and costs while enabling government employees to focus on more complex tasks. Chatbots have been piloted to help residents apply for government housing, answer non-emergency public health questions, and apply for a business license.[52,53] In addition to streamlining tasks, Chatbots also feature the ability to communicate in foreign languages, enabling more personalized and effective citizen-government interaction.

AI can also be used to predict work demand and create efficient task schedules. Engineering tasks on Hong Kong’s subway system are scheduled and managed by AI. An analysis by the Brookings Institution showed that “during a typical week, about 10,000 workers carry out around 2,600 engineering tasks ranging from smoothing rails to replacing tracks. The cognitive system saves about two days per week by optimizing scheduling of tasks and allocation of resources.”[54] Microsoft’s ‘Connected Field Service’ uses IoT to monitor field sites and consequently improve scheduling of employees in the field.[55]

Public Services

AI can help governments better manage and allocate resources and services effectively by reducing backlogs, overcoming resource constraints, and optimizing response efforts. Predictive risk assessment and response optimization can help cities better prepare for future contingencies and allocate resources accordingly.

Risk Assessment and Support in Social Service Programs

Social service programs are targeted to assist the most disadvantaged populations. Such programs typically require continuous follow-up from caseworkers, yet resource constraints often limit the individual attention that program beneficiaries need. Incorrect determinations by caseworkers and poor adherence to recommendations by participants not only hinders progress, but can further compound the constraints of the administering organizations. IBM’s Human Outcome Analytics attempts to improve the performance of social service programs by combining individual assessment data with historical records to formulate strategic plans for participants. Applying AI tools to the social services sector should be approached with caution, however, since decisions are highly context-specific. Risk assessment tools that use individual-

specific data are likely to capture features of race or gender and incorporate them into the decisions, which could lead to discriminatory results that are hard to detect.[56]

Prioritizing Inspections

Health and safety inspections are essential urban governance tasks, and enforcing compliance with local regulations is critical to ensuring the safety of city residents. Yet for most units, the workload is substantially greater than the available inspectors can meet, necessitating use of random selection of inspection locations. In this case, predictive analytics can help prioritize inspections by identifying high-risk locations.[57]

The New York City Fire Department deployed the Risk-Based Inspection System (RBIS), an AI model that predicts fire likelihood based on a variety of variables such as building structure and past violations.[58] Data from multiple cities add to the richness of the model. Application of an RBIS can help any city address the shortage of inspectors vis-à-vis the large number of buildings by better targeting inspections at high-risk locations.

Social media and crowdsourced data can also play a role. To improve food safety and public health, Las Vegas experimented with using Twitter data. The city analyzed approximately 16,000 tweets daily to identify keywords indicative of food-borne illness in relation to specific restaurants. These insights were used to create a high-priority list for inspection. The targeted inspections outperformed random inspections, resulting in citations in 15 percent of targeted inspections compared to 9 percent in randomly selected inspections.[59]

Natural Disaster Prediction, Management, and Response

Effective emergency response can save lives, minimize civilian injuries, and reduce property and infrastructure damage. The ubiquity of sensors and increased computing power allow cities to monitor several ecological factors that can help predict natural disasters or provide early warnings. Current systems such as ShakeAlert, built by the Seismology Lab at UC Berkeley, have developed predictive algorithms that can provide an alert for an impending earthquake a minute before it strikes. The Bay Area Rapid Transit (BART) system in the San Francisco Bay Area uses this alarm to slow down its trains to prevent derailment.[60] Scientists at Los Alamos National Laboratory are also using machine learning techniques to predict earthquakes months in advance.[61]

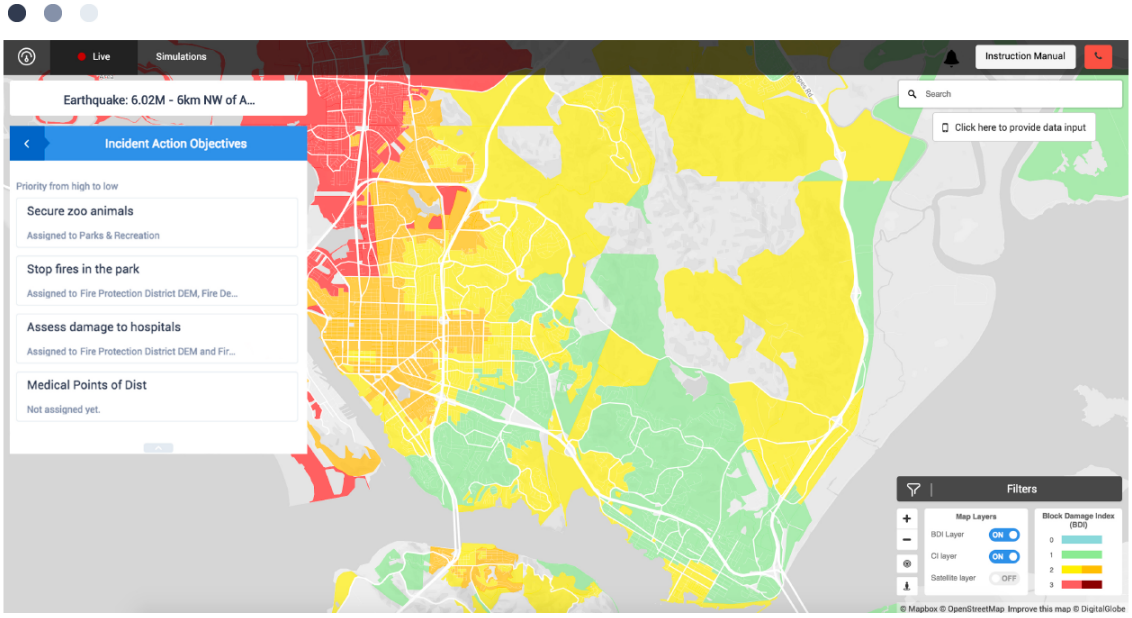


Figure 4. The One Concern platform for disaster prediction and response optimization. Source: OneConcern.com

Emergency response requires coordinated action by several agencies. One Concern, a U.S.-based startup, enables formation of targeted response strategies by offering simulations of an earthquake's effects based on building characteristics and adjacent social and ecological features such as soil data and location of water bodies, population density, and geographic changes over time (see Figure 4). After an earthquake strikes, the platform can be used to allocate tasks to relevant agencies and track their actions in real time. One Concern is currently being used in San Mateo County, California, and its predictive capabilities have proven effective at identifying priority disaster risk areas at the city-block level.[62]

Decision Support Systems for First Responders

AI systems can assist first responders in the field as decision support systems. The Jet Propulsion Laboratory at NASA has developed AUDREY, an AI-enabled system that tracks individual firefighter's movements and makes recommendations on how the crew can better coordinate its response. AUDREY not only increases the effectiveness of firefighters during crises by helping identify objects among debris, it also monitors firefighters' health. This can potentially address issues such as the high incidence of heart attacks among firefighters.

AI tools can also be used to accelerate response among different agencies and increase lead time. IBM's Intelligent Operations Center for Emergency Management draws data from disparate sources such as street cameras and police reports, to identify a range of possible threats and alert multiple departments simultaneously, enabling a more concerted response among relevant emergency services. [63]

Tracking Epidemics and Catastrophic Events

User-generated content on social media platforms allows real-time event tracking. CrowdBreaks is a disease surveillance system that uses Twitter feeds and hashtags relevant to a given disease to identify at-risk locations and populations.[64] Such systems can be used to provide early warnings, which can save critical time in managing epidemics. During the Ebola outbreak in 2014, AI helped track population movement patterns and outbreak locations to inform public health tactics. These AI systems used cell phone data to estimate the spread and extent of infection rates in the region and to help health agencies target health screenings and security checkpoints at airports.[65]

Banjo, a startup based in California, analyzes mass amounts of social media data worldwide to identify and explain events as they occur in real time. The software combines the analysis of text and metadata (e.g., location and time) with computer vision algorithms to enable real-time analysis and insights of images posted through social media. The company has mapped the world into a grid of 35 billion squares. If abnormal posts or images start emerging from within one of the squares, the data are flagged and analyzed. For example, Banjo was used to assemble photos in near real-time from the scene of the Boston Marathon bombing, revealing insights into the location and impact of the detonated bombs. Recently, it identified an Amtrak train collision in Philadelphia within five minutes of its derailment and was able to identify a fire at an Amazon data center in Virginia before the fire department arrived.[66]

PREPARING FOR AI IN GOVERNMENT

> ***"Artificial intelligence presents a cultural shift as much as a technical one. This is similar to technological inflection points of the past, such as the introduction of the printing press or the railways."***
>
> **- Kate Crawford and Meredith Whittaker, AI Now [67]**

AI offers great potential to revolutionize how cities function—making them more efficient, transparent, and inclusive. However, these gains will not be achieved without a corresponding change in governance culture.[68] This cultural shift will not only involve city leaders but also citizens who access public services through these new systems. It will be necessary to train city leaders on capacities and limitations of these new tools. Programs like the Bloomberg Harvard City Leadership Initiative, which offers executive education and coaching for city mayors and their senior staff, can enable focused training on application and management of AI-enabled technologies. It will also be important for citizens to understand the benefits and of risks of AI and become accustomed to interacting with AI-enabled tools as a point of contact for public services.

Coding Inclusivity into AI

AI can be applied to some of the toughest challenges in civic governance. This paper has highlighted applications in the domains of public safety and law enforcement, labor, and public services. In addition to enabling efficiency and financial gains, AI tools can help detect and mitigate human bias in decision-making.

Recent applications of AI have also brought to light multiple cases of discriminatory outcomes for protected groups. These anecdotes serve as early warning signs against unbridled optimism about the use of AI for urban governance as AI can also serve to perpetuate existing social biases and exacerbate inequalities. Public understanding of AI technologies and their implications for society is still evolving. Placing inclusivity as a core goal in the design and application of these tools can better ensure that city leaders are aware of and responsive to emerging benefits and risks. At the very least, careful consideration of potential discriminatory

effects of AI should be recognized and corrected for, to the extent possible, before integration into government processes.

While AI systems can help reduce bias from human subjectivity in day-to-day decisions, subjective judgment can still be engrained into these systems. Figure 5 details different points in the design process that involve human subjectivity and are thus vulnerable to bias. Even in the absence of malicious intent, discrimination could permeate AI systems through incorporation of value-laden data, prioritization of subjective factors, and faulty decision-making processes. Detecting and mitigating unintentional discrimination through AI systems is difficult. The following section details some of the overarching steps that city governments can take to better prepare themselves to benefit from AI systems in their daily processes while mitigating risks of discrimination.

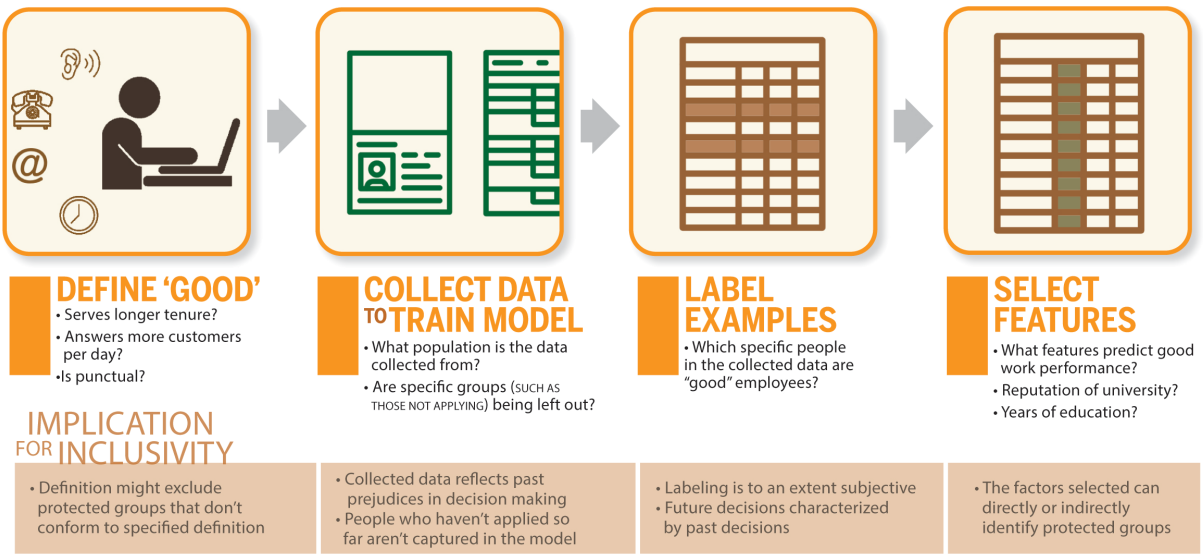


Figure 5. Considering Design Implications for Inclusivity in Predictive Models

Invest in Data

Machine learning algorithms classify, identify, and predict based on relationships identified in past observations. Since data are at the core of machine learning algorithms, its quality determines the performance of the model. It is important to consider how data can perpetuate discrimination.

Unequal Representation of Different Classes

Ideally, the data on which the AI model is built should represent the population targeted. In reality however, this might not be the case. For example, populations with limited access to smartphones or the internet might be poorly represented in data collected from online sources. Decisions based on predictions from such data would continue to ignore the needs of these communities.[69] For example, an AI model that uses only information from previously submitted resumes to train which candidates to prioritize might erroneously neglect individuals with unconventional backgrounds.

Street Bump

"Boston's Street Bump app, collects smartphone data from drivers going over potholes. However, if cities begin to rely on data that only come from citizens with smartphones, it will necessarily have less data from those neighborhoods with fewer smartphone owners, which typically include older and less affluent populations.

Boston's Office of New Urban Mechanics has made concerted efforts to address these potential data gaps, less conscientious public officials may miss them and end up misallocating resources in ways that further entrench existing social inequalities."

Kate Crawford, Think Again: Big Data, *Foreign Policy*

Human Biases Captured in Data

Many AI models aim to reduce inconsistencies in human decisions resulting from implicit or explicit biases. Yet the data on which these models are trained often capture these biases. For example, neighborhoods more likely to be patrolled in the past may have seen more arrests. Predictions from such data could continue to highlight the neighborhood as a crime hotspot without legitimate reasons for excessive patrolling. While models using historic data may reflect prejudices of past human decisions, models using real-time inputs such as recommendation systems can reflect the prejudices of currently active users.

Cities can take steps to address concerns of discrimination from data through the following practices:

- **Critically Evaluate and Correct Data-Collection Processes.** If cities have reason to believe their data are selectively under- or over-representing certain groups, they can target their collection to compensate for skewed representation. Agencies should also carefully consider the specific information they collect. In absence of good predictors, algorithm designers might use features of the limited dataset that might not only be less accurate but also revealing of group membership and could result in discriminatory

results. Thus, agencies should critically evaluate the predictors in their models, checking and correcting for discriminatory results.

- **Retrain Models on Newer Data.** Public service agencies should also consider periodically re-training their models on fresh data to ensure that any patterns based on historic data are not repeated in future decisions.

Clarify Subjective Design Criteria

While algorithms make consistent decisions, these decisions are based on criteria specified by the designers of these models. It is important that the model's design reflects the values and intentions for its application. Thus, we recommend that public service agencies carefully decide error thresholds, define target variables, and define fairness, to the extent possible:

- **Decide Error Thresholds**

  Machine learning algorithms can be designed to penalize misdetections (false negatives) and false alarms (false positives) differently. While it might be more important for a city to detect all the buildings at risk of fire, another city might consider it more important to minimize false identification of safe buildings to reduce staff time and expense. This minimization of false identification might come at the cost of a few more misdetections. Thus, cities must not only decide on the maximum error rate for the predictive algorithm, they must also weigh the relative importance of the two kinds of errors. These thresholds are likely to depend on the specific application and resource constraints in a city.

- **Define Target Variables**

  The first step of machine learning requires translation of variables to be considered into more formal terms in the model. The behavior to be predicted must be captured in a finite set of values of a variable. While this translation can be better performed through contextual understanding, it is necessarily a subjective process. For example, in designing an AI tool that selects prospective employees from a pool of applicants or decides whether an employee should be promoted, the designers must first define what makes a "good" employee in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example.[5] It is critical that the effects of how target variables are defined be considered and evaluated for discriminatory implications.

- **Define Fairness**

  Whether with the intent of mitigating bias in human decision-making or monitoring the discriminatory aspects of newer algorithms, cities must clarify the factors they consider discriminatory. This is especially important for applications in law enforcement, as there are multiple definitions of fairness and not all may be satisfied simultaneously.[70]

  Within law enforcement, fairness as defined by demographic parity insists that an equal proportion of defendants are detained in each protected group.[71] Such a definition can be problematic since one might incarcerate women who pose no public safety risk in order to release the same proportions of men and women on probation. Another definition of fairness could be that conditioning on legitimate factors, the algorithm is equally accurate for all classes. For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates. In summary, cities must first define their criteria for fairness in order to inform the model's design and explicitly share these definitions to enable assessment of whether the models are discriminatory.

Support Equity in Innovation

A commitment to inclusivity not only mandates parity in decisions made by AI models but also in access to resources to participate in this new wave of innovation. Most research in the field of AI is concentrated in academia and a few large technology companies. Currently, writing and reading code is a specialized skill.[72] As AI systems become more complex, they threaten to widen the existing digital divide; those affected by decisions of AI tools might not have adequate technical literacy to understand or contribute to their development, application, evaluation, and outcomes.

Developing AI tools also requires access to vast amounts of data and computational resources, which tend to be expensive and proprietary. Economic constraints limit access to these necessary resources, potentially excluding some demographic groups from innovating in the field. Greater diversity among developers is not only necessary to share future economic opportunities equitably, but will also increase the likelihood that these systems reflect the needs of diverse communities. Governments should invest in technical literacy and foundational infrastructure to enable more equitable access to rapid developments in the discipline. Some of

the ongoing efforts such as the U.S. Department of Education's program to provide financial support to students in coding programs and online courses can help expand coding literacy. [73] A number of non-profits such as Code.org and Girls Who Code are working to make coding skills more accessible.[74] Technical literacy will enable people to understand the assumptions made by AI models, question their application and impact, and suggest and develop alternate use cases.

Managing Uncertainty

Criteria to Consider when Choosing AI

As with other new technologies, using AI tools in civic decision-making is accompanied by a degree of uncertainty. While technologists in the field are building and analyzing underlying algorithms, other researchers are exploring the impact AI will have within society at large. Given this uncertainty, public service agencies considering using AI should clarify their reasons for doing so. These reasons could be efficiency gains, effectiveness, or greater inclusivity in service delivery through consistent and transparent decisions. Yet AI-enabled technologies do not automatically guarantee these outcomes. AI works well for problems that lend themselves to formalization in a way that computers can understand. Thus, not all policy challenges lend themselves to the utilization of AI. Furthermore, while gains in efficiency and effectiveness from these technologies might be quantifiable, the inadvertent discriminatory outcomes may be harder to observe and have wide-reaching negative effects.

The following section lists four qualitative features of AI-enabled technologies that cities should consider. These criteria can help public service agencies navigate the uncertainty regarding risks of AI while phasing in these tools to support their ongoing work.

Transparency

Technology companies building AI tools often guard the underlying algorithm as proprietary, claiming that secrecy is necessary to maintain their competitive advantage and to incentivize innovation in the field. Secrecy might also be necessary from a security perspective—if the parameters of decision-making through an algorithm are known, the system may be gamed to yield specific results.

On the other hand, intentional secrecy makes it harder to inspect or regulate the systems, making erroneous behavior harder to detect—the social, political, and economic ramifications of discriminatory prediction or classification could have deleterious effects that are not easily identified or corrected. Secrecy reinforces power asymmetries between the developers of the algorithms and those affected by their decisions.

Opening access to the algorithm, or to the extent possible the factors considered within the model, could increase its transparency and trustworthiness by allowing greater scrutiny among a diverse group. Some of the concerns regarding gaming of an open-source algorithm can be addressed by using factors for prediction that are less susceptible to manipulation. For example, in medical applications "it is preferable to base risk-adjustment systems on diagnosis-related health data, rather than on treatment data," because the former is more resistant to manipulation.[75] However, restricting the number of parameters on which an algorithm predicts future risk can come at the cost of predictive performance.

Some machine learning algorithms—specifically those in which the application developer does not specify the predictive features (such as neural networks)—can select features that might make limited or no sense to humans designing these algorithms. While the predictive accuracy of such algorithms might be high, it would be difficult to explain the classification of the target factors by the model. In several decisions affecting daily civic life, the results must be explainable. Consequently, results from AI models must be interpretable. In some domains such as the credit market where explaining the logic of decisions is legally mandated. The European Union's new General Data Protection Regulation, has tried to address this issue by creating provisions for a 'right to explanation' meant to allow individuals to demand explanation about an algorithmic decision made about them. Yet, this right has been argued as legally only providing a 'right to be informed' that a decision has been made through the use of automated decision-making in order to enable the individual to contest the decision, not for the individual to receive a full explanation of *how* the algorithm made its decision.[76]

Autonomy

AI technologies can automate in several ways—they can relieve workers by taking over mundane tasks, they can complete some of the sub-tasks, they can augment skills, or they can replace a job previously done by humans (such as language translation).[77] As cities incorporate

AI systems into their daily processes, they must consider whether the systems supplement or reduce human agency in decision-making processes and the implications of the latter.

Clearly, human agency is critical in some domains. For example, delegating legal decisions to autonomous systems could challenge fundamental assumptions about legal due process and could produce spurious outcomes that current ethical and legal frameworks are ill equipped to address. Greater mechanical autonomy may make it harder to allocate responsibility in instances of failure. Thus, AI systems should serve as advisory tools rather than final arbiters.

Indeed, several researchers have proposed that human accountability be designed into AI tools. 'Human-in-the-loop' techniques in machine learning leave room for human input at critical decision-making junctures. Besides adding to the accuracy of the algorithms, such techniques also ensure some human agency in the final decision. In cases where AI tools are replacing human decision-making, cities might consider algorithmic designs that ensure accountability in human decisions.

Efficiency

As some of the cases mentioned in this paper demonstrate, AI can achieve dramatic improvements in efficiency by adequately allocating human and financial resources in places where these are scarce. In an age of data abundance, AI can facilitate decision-making by finding and presenting the necessary information in a timely manner.

AI can help governments become more efficient by enabling prompt emergency response, managing workforce, and supporting the work of judicial systems. Risk-assessment tools aimed at increasing accuracy of decisions—whether in law enforcement or disaster preparedness—enable large financial savings by streamlining current processes and preventing future events that could be a considerable cost to society. Where possible, agencies should consider a cost-benefit analysis of using AI-enabled tools prior to deployment. This can help cities better evaluate the utility of these tools against some of the risks they present.

Effectiveness

AI tools can help cities better fulfill their defined mandate. The Risk-Based Inspection System used by the New York City Fire Department, for example, helped the fire department prioritize at-risk buildings and inspect those that might otherwise have gone unnoticed and uninspected.

To assess the effectiveness of its work and the tools deployed, cities must precisely define what they aim to identify and predict. For example, cities must clearly operationalize what is considered 'criminal activity' in risk-prediction tools used in law enforcement. Would the system be used to predict the likelihood of committing a felony such as physical harm, or a misdemeanor such as vandalism, or both? The process of translating a policy objective into a machine learning problem is a subjective process. Cities should take care in this translation process and have a good understanding of what these AI tools are measuring.

To monitor the effectiveness of these tools, cities should plan for field trial evaluations. Cities could team up with academic institutions or independent agencies to generate and analyze performance metrics for AI tools and assess whether and how they improve on the methods they intend to replace. At the same time, cities should guard against the unintended consequences of effective measurement. For example, being able to precisely identify people with higher risk in a population could undermine traditional beneficiaries of social institutions such as insurance and other pooled social safety nets or may inadvertently target certain demographic groups. Cities should carefully consider the desirability and merits of accurate measurement and prediction.

Confronting Tradeoffs

While evaluating options for AI systems or choosing between different AI tools, cities should consider how important each of the aforementioned criteria are in the context of application. These criteria might be in tension with one another. For example, a more efficient algorithm could be less interpretable and transparent. Similarly, it might be costlier (in time and money) to build a system with continuous human inputs; however, ensuring human agency in decision-making might be critical in some domains (e.g., sentencing guidelines). Figure 6 shows possible scenarios of AI applications and considerations for each of these four criteria. Having decided the relative importance of these criteria, city leaders can select the technologies that best meet their requirements.

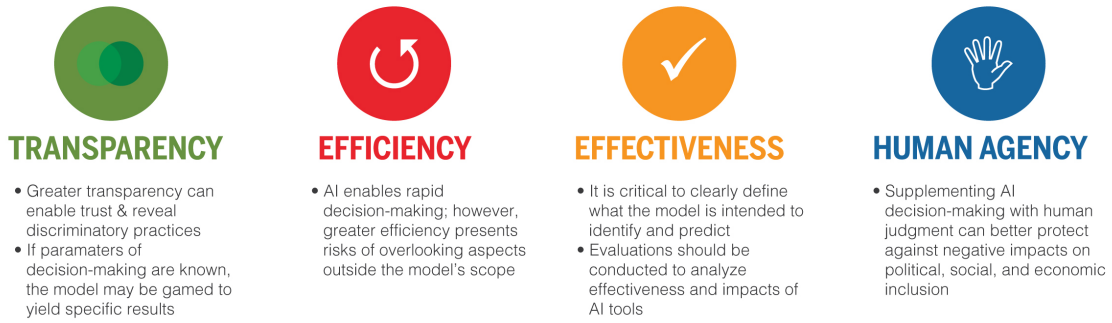


Figure 6. Criteria to Consider When Implementing AI

Managing Errors

While cities should attempt to minimize errors at the design phase, they should also craft contingency plans for false predictions or failure. In circumstances where results have unexpected or irregular predictions—such as labeling pre-trial defendants of a particular race or socioeconomic class as high-risk—agencies should implement another round of human review. Agencies using these systems should outline points of human agency and accountability to support responsibility and remediation in instances of failure. Evaluation mechanisms to assess the performance and potential discriminatory outcomes of AI-based tools can help in early detection of system failure and help minimize deleterious effects.[78,79]

MiDAS

The Michigan Unemployment Insurance Agency falsely accused tens of thousands of Michigan residents of committing unemployment insurance fraud based on the results of the Michigan Integrated Data Automated System (MiDAS). "The system had a 93% error rate and made false fraud findings affecting more than 20,000 unemployment insurance claims. Those falsely accused of fraud were subjected to quadruple penalties and aggressive collection techniques, including wage garnishment and seizure of income tax refunds. Some were forced into bankruptcy."[78] A lawsuit is now pending against the state over false fraud findings. While it is unclear if the problem is the software or the way it was used, the software development company Fast Enterprise is also being tried for the case. The company however claims that it would have limited liability.

Diagnosability

Agencies should consider the diagnosability of the AI-enabled technologies they are deploying (i.e., whether it is possible to know if a given model has failed in whole or in part).[80] While technical complexity could lead to low diagnosability of technologies, bureaucratic structures

can also add to the challenge. When designing and deploying these systems, city leaders must consider methods to make error reporting and detection more convenient so system failures may be caught with minimal negative impact.

AI technologies and their applications in public life are still in their incipient stages and should be carefully adopted. While it would be ideal to test the AI tools thoroughly before deployments, not all errors may be caught or even anticipated outside of the real-world setting. Public service agencies should partner with academic institutions and independent agencies to enable continuous evaluations and audits, especially when the underlying algorithms are not public.

CONCLUSION

Artificial Intelligence is one of the largest technology shifts happening globally. Cities are well positioned to leverage these technologies to improve their efficiency and equity but should be mindful of and responsive to potential risks. Explicit attention to fairness and inclusivity in design, application, and evaluation of these new technologies will not only minimize inadvertent discriminatory effects of these tools, but can also make these revolutionary technologies a force for greater social, economic, and political inclusion.

ACKNOWLEDGMENTS

We would like to thank Microsoft for its support of this white paper and public symposium "Inclusive AI: Technology and Policy for a Diverse Urban Future," held at CITRIS and the Banatao Institute on May 10, 2017. Video from the symposium is available at tiny.cc/inclusiveai. We also acknowledge with gratitude the distinguished experts who took the time to talk with us and share their knowledge and advice.

- Ahsan Baig, Project Manager III; City of Oakland, Systems & Database Administration, Information Technology Department
- Dan Baker, PhD Student, Goldman School of Public Policy, UC Berkeley
- Mark Bergstrom, Executive Director, Pennsylvania Sentencing Commission
- Richard Berk, Professor of Criminology & Statistics, University of Pennsylvania
- Joshua Blumenstock, Asst. Professor, School of Information, UC Berkeley
- Ryan Calo, Asst. Professor, School of Law, University of Washington
- Edward Chow, Manager, AUDREY, NASA Jet Propulsion Laboratory
- Brian Christian, Author, *The Most Human Human;* Co-Author, *Algorithms to Live By*
- Mariko Davidson, Civic Partnerships Manager, Technology & Civic Engagement, Microsoft
- Avi Feller, Asst. Professor of Public Policy, Goldman School of Public Policy, UC Berkeley
- Megan Garcia, Senior Fellow & Director, New America CA
- Rayid Ghani, Director of the Center for Data Science & Public Policy, University of Chicago
- Sharad Goel, Asst. Professor, Dept. of Management Science & Engineering, Stanford
- Ken Goldberg, Professor, Industrial Engineering and Operations Research, UC Berkeley
- Melissa Hamilton, Visiting Criminal Law Scholar, School of Law, University of Houston
- Mohit Kothari, *Independent*
- Zvika Krieger, Co-Lead, Center for the Fourth Industrial Revolution, World Economic Forum
- Fei-Fei Li, Assoc. Professor & Director, Stanford AI Lab; Chief Scientist of AI, Google Cloud
- Roberto Manduchi, Professor, Computer Engineering, UC Santa Cruz
- Nikhil Marathe, *Independent*
- Scott Mauvais, Director, Technology & Civic Innovation, Microsoft
- Naman Muley, *Independent*
- Deirdre Mulligan, Assoc. Professor, School of Information, UC Berkeley
- Robert Pless, Professor of Computer Science, George Washington University
- David Robinson, Principal, Upturn
- Jessica Saunders, Senior Criminologist, RAND
- Jennifer Skeem, Professor & Associate Dean of Research, U.C. Berkeley Social Welfare
- Costas Spanos, Director, CITRIS and the Banatao Institute
- Sonja Starr, Professor of Law, University of Michigan Law School
- Laura Tyson, Faculty Director, Institute for Business & Social Impact, Haas School of Business, UC Berkeley
- Chris White, Principal Researcher, Microsoft
- David Zaharchuk, Global Industry Research Lead, Institute for Business Value, IBM
- John Zysman, Professor Emeritus, Political Science, UC Berkeley

REFERENCES

[1] The World Health Organization [WHO]. (2015, Sept.). Ageing and Health. http://www.who.int/mediacentre/factsheets/fs404/en/

[2]Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep Learning*. Cambridge, MA: MIT Press. http://www.deeplearningbook.org

[3] Goldberg, K. (2017, June 11). The Robot-Human Alliance. *The Wall Street Journal. https://www.wsj.com/articles/the-robot-human-alliance-1497213576*

[4] Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A., (2016, June 18). Deep Learning for Identifying Metastatic Breast Cancer. https://arxiv.org/pdf/1606.05718v1.pdf

[5] Barocas, S. & Selbst, A. (2016) Big Data's Disparate Impact. *California Law Review. 104,* 671-730. DOI: http://dx.doi.org/10.15779/Z38BG31. Accessed from: http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

[6] Hickman, L. (2013, July 01). How algorithms rule the world. *The Guardian*. https://www.theguardian.com/science/2013/jul/01/how-algorithms-rule-world-nsa

[7] Perry, W., McInnis,B., Price, C., Smith, S., and Hollywood, J. (2013). Predictive Policing: Forecasting Crime for Law Enforcement. Santa Monica, CA: RAND Corporation. http://www.rand.org/pubs/research_briefs/RB9735.html.

[8] Hvistendahl M. (2016, September 28). Can 'predictive policing' prevent crime before it happens? *Science*. http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

[9] Robinson, D. & Koepke, L. (Aug. 2016). Stuck in a pattern: Early evidence on "predictive policing" and civil rights. *Upturn.* https://www.teamupturn.com/reports/2016/stuck-in-a-pattern

[10] Mohler, G. O., Short, M.B., Malinowski, S., Johnson, M., Tita, G.E., Bertozzi, A., & Brantingham, P.J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association, 110*(512), 1399-1411.

[11] Hunt, P., Saunders, J. & Hollywood, J. (2014). *Evaluation of the Shreveport Predictive Policing Experiment*. Santa Monica, CA: RAND Corporation.

[12] Gorner, J. (2016, July 22). With violence up, Chicago police focus on a list of likeliest to kill, be killed. *Chicago Tribune. http://www.chicagotribune.com/news/ct-chicago-police-violence-strategy-met-20160722-story.html*

[13] Sheehan, T. (2016, March 31). Fresno council halts purchase of data software wanted by police. *Fresno Bee. http://www.fresnobee.com/news/local/article69337677.html*

[14] Jouvenal, J. (2015, Jan. 10). The New Way Police Are Surveilling You: Calculating Your Threat 'Score', *Washington Post.* https://www.washingtonpost.com/local/public-safety/the-new-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e5-baf4-bdf37355da0c_story.html?utm_term=.708bf594e0a6

[15] American Civil Liberties Union [ACLU]. (2016, Aug. 31). Statement of concern about predictive policing by ACLU and 16 civil rights, racial justice, and technology organizations. https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

[16] Saunders, J., Hunt, P. & Hollywood, J. *Predictions Put into Practice: A Quasi-experimental Evaluation of Chicago's Predictive Policing Pilot, Journal of Experimental Criminology, 12(*3), 347-371.

[17] Saunders, J. (2016, October 11). Pitfalls of Predictive Policing. 'The Rand Blog'. Rand Corporation. http://www.rand.org/blog/2016/10/pitfalls-of-predictive-policing.html

[18] Laura and John Arnold Foundation. (Nov. 2013) *Developing A National Risk Model for Pretrial Risk Assessment.* http://www.arnoldfoundation.org/wp-content/uploads/2014/02/LJAF-research-summary_PSA-Court_4_1.pdf

[19] Laura and John Arnold Foundation. (Aug. 2016). *Results from the First Six Months of the Public Safety Assessment – Court in Kentucky*. http://www.arnoldfoundation.org/wp-content/uploads/2014/02/PSA-Court-Kentucky-6-Month-Report.pdf

[20] Casey, P., Elek, J., Warren, R., Cheesman, F., Kleiman, M., Ostrom, B. (2014). *Offender Risk & Needs Assessment Instruments: A Primer for Courts.* National Center for State Courts. http://www.ncsc.org/~/media/Microsites/Files/CSI/BJA%20RNA%20Final%20Report_Combined%20Files%208-22-14.ashx

[21] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23) How we analyzed the COMPAS recidivism algorithm. *ProPublica.* https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[22] U.S. Department of Justice, Criminal Division. (2014, July 29). The Promise and Danger of Data Analytics in Sentencing and Corrections Policy. https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf

[23] Brambila, N. C. (2016, Sept. 28). Recidivism risk assessment as sentencing tool is controversial, *Reading Eagle*. Retrieved May 02, 2017, from http://www.readingeagle.com/news/article/recidivism-risk-assessment-as-sentencing-tool-is-controversial#.WQlKRVPyuYU

[24] Pennsylvania Commission on Sentencing. (Feb. 2016). *Risk Assessment Project II: Interim Report 2. Validation of a Risk Assessment Instrument by Offense Gravity Score for all Offenders.* http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment/phase-ii-reports/interim-report-2-validation-of-risk-assessment-instrument-by-ogs-for-all-offenses-february-2016/view

[25] Bergstrom, M. (2016). Message from the Executive Director. *Pennsylvania Commission on Sentencing*. http://pcs.la.psu.edu/publications-and-research/the-monitor/fall-2016/view

[26] Ostrom, B., Kleinman, M., Cheesman II, F., Hansen, R., Kauder, N. (2002). Offender Risk Assessment in Virginia. *National Center for State Courts.* http://www.vcsc.virginia.gov/risk_off_rpt.pdf

[27] U.S. Department of Justice, Criminal Division. (2014, July 29). The Promise and Danger of Data Analytics in Sentencing and Corrections Policy. https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf

[28] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017, March 27). Fairness in Criminal Justice Risk Assessments: The State of the Art. Retrieved April 25, 2017, frhttps://arxiv.org/abs/1703.09207v1

[29] Wisconsin Department of Justice. (2016, April 05). DOJ argues State v. Eric L. Loomis in Wisconsin Supreme Court. https://www.doj.state.wi.us/news-releases/doj-argues-state-v-eric-l-loomis-wisconsin-supreme-court

[30] Eric L. Loomis, Petitioner v. Wisconsin. (2016, October 12). https://www.supremecourt.gov/search.aspx?filename=%2Fdocketfiles%2F16-6387.htm

[31] U.S. Department of Justice, Criminal Division. (2014, July 29). The Promise and Danger of Data Analytics in Sentencing and Corrections Policy. https://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf

[32] Walker, S., Alpert, G., & Kenney, D. *Early Warning Systems: Responding to the Problem Police Officer* (U.S. Department of Justice, Office of Justice Programs, National Institute of Justice Research in Brief 188565)

[33] U.S. Department of Justice, Office of Community Oriented Policing Services. *Early Intervention Systems for Law Enforcement Agencies: A Planning and Management Guide* (Washington, D.C.: 2003).

[34] Arthur R. (2016, March 9). We Now Have Algorithms to Predict Police Misconduct: Will Police Departments Use Them? *FiveThirtyEight*. https://fivethirtyeight.com/features/we-now-have-algorithms-to-predict-police-misconduct/

[35] Gregory, T. (2016, Aug. 18). U. of C. researchers use data to predict police misconduct. *Chicago Tribune.* http://www.chicagotribune.com/news/ct-big-data-police-misconduct-met-20160816-story.html

[36] Coldewey, D. (2017, June 5). Deep analysis of policy body cam footage shows patterns of less respectful speech towards black people. *TechCrunch.* https://techcrunch.com/2017/06/05/deep-analysis-of-police-body-cam-footage-shows-pattern-of-microaggressions-towards-black-people/

[37] Purohit, S. R. (2015, July 01). How LinkedIn Knows What Jobs You Are Interested In. *Udacity.* http://blog.udacity.com/2014/05/how-linkedin-knows-what-jobs-you-are.html

[38] Osborne, M. & Frey, C. (2013). The future of employment. *The Future of Life*. Retrieved April 12, 2017 from https://futureoflife.org/data/PDF/michael_osborne.pdf

[39] Davies, A. (2016, October 25). Uber's Self-Driving Truck Makes Its First Delivery: 50,000 Beers. *Wired.* https://www.wired.com/2016/10/ubers-self-driving-truck-makes-first-delivery-50000-beers/

[40] American Trucking Association- Reports, Trends & Statistics. (n.d.). Retrieved May 02, 2017, from http://www.trucking.org/News_and_Information_Reports_Industry_Data.aspx

[41] Ramachandran, V. (2017, April 05). How Driverless Vehicles Could Harm Professional Drivers Of Color. *NPR Code Switch.* http://www.npr.org/sections/codeswitch/2017/04/05/522597987/how-driverless-vehicles-could-harm-professional-drivers-of-color

[42] US Bureau of Labor Statistics (May 2015). STEM crisis or STEM surplus? Yes and Yes. *Monthly Labor Review*. https://www.bls.gov/opub/mlr/2015/article/stem-crisis-or-stem-surplus-yes-and-yes.htm

[43] PBS Newshour (2014, Feb. 13). Second machine age' will require more human creativity. http://www.pbs.org/newshour/bb/second-machine-age-will-require-more-human-creativity/

[44] Brynjolfsson, E. (2014, September 07). What does the second machine age mean for our jobs? *World Economic Forum.* https://www.weforum.org/agenda/2014/09/video-second-machine-age-mean-jobs/

[45] Meister, J. C. (2014, Aug. 07). Make Sure Your Dream Company Can Find You. *Harvard Business Review*. https://hbr.org/2013/12/make-sure-your-dream-company-can-find-you

[46] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, (1). doi:10.1515/popets-2015-0007

[47] Brustein, J. (2016, Nov. 22). Studies Show Racial and Gender Discrimination Throughout the Gig Economy. *Bloomberg Technology*. https://www.bloomberg.com/news/articles/2016-11-22/studies-show-racial-and-gender-discrimination-throughout-the-gig-economy

[48] Murphy, L. (2016, Sept. 8). Airbnb's work to fight discrimination and build inclusion. http://blog.airbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf

[49] Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. Working Paper, *National Bureau of Economic Research*. doi:10.2139/ssrn.422902

[50] Executive Office of the President (2016, Oct). Preparing for the Future of AI. Washington DC 20502: Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[51] Joonko. (2016, Sept. 14). *How Joonko can ensure equal opportunities to succeed within your sales team.* https://www.youtube.com/watch?v=Zefg8vMjwW8

[52] Bousquet, C. (2017, May 5). 5 ways chatbots could transform government services. *Government Technology.* http://www.govtech.com/computing/5-Ways-Chatbots-Could-Transform-Government-Services.html

[53] Douglas, T. (2017, May 3). Los Angeles, Microsoft unveil Chip: New chatbot project centered on streamlining. *Government Technology.* http://www.govtech.com/computing/Los-Angeles-Microsoft-Unveil-Chip-New-Chatbot-Project-Centered-on-Streamlining.html

[54] Desouza, K. C., & Krishnamurthy, R. (2016, October 14). How can cognitive computing improve public services?. *Brookings Institution*. https://www.brookings.edu/blog/techtank/2016/10/13/how-can-cognitive-computing-improve-public-services/

[55] Microsoft. (June 2017). Gartner Insights: Government Cloud Trends. https://enterprise.microsoft.com/en-us/industries/government/state-and-local/

[56] https://www-01.ibm.com/common/ssi/cgi bin/ssialias?subtype=SP&infotype=PM&htmlfid=GVD03009USEN&attachment=GVD03009USEN.PDF

[57] Coffman, B. (2013, Jan. 01). Code Enforcement: Critical for a Successful Fire Prevention Program. *Fire Engineering.* http://www.fireengineering.com/articles/print/volume-166/issue-01/features/code-enforce-critic-success-fire-prevent-prog.html

[58] Roth, J. Chen, J. (2014, May 22). FDNY's Data Science Strategy for Risk Management. *Fire Department of New York.* http://nationaluasi.com/dru/2014%20Presentations/FDNY_FireCast_UASI_2014-5-22.pdf

[59] Sadilek, A., Kautz, H., Diprete, L., Labus, B., Portman, E., Teitel, J., Silenzio, V. (2016). *Deploying nEmesis: Preventing Foodborn Illness by Data Mining Social Media.* Association for the Advancement of Artificial Intelligence (AAAI). http://www.cs.rochester.edu/~sadilek/publications/Sadilek,%20Kautz,%20et%20al.%20Deploying%20nEmesis.pdf

[60] Newitz, A. (2013, May 13). Our Algorithms Can Predict Future Disasters — Now What? *Wired.* https://www.wired.com/2013/05/newitz-disasters/

[61] Sneed, A. (2017, February 15). Can Artificial Intelligence Predict Earthquakes? *Scientific American.* https://www.scientificamerican.com/article/can-artificial-intelligence-predict-earthquakes/

[62] Spector, J. (2015, November 18). An Earthquake Response System That's Faster Than 911. *The Atlantic Citylab.* http://www.citylab.com/cityfixer/2015/11/an-earthquake-response-system-thats-faster-than-911/416427/

[63] IBM. (June 2017). IBM Intelligent Operations Center for Emergency Management. https://www.ibm.com/us-en/marketplace/emergency-management

[64] Palmer, S. (n.d.). Crowdbreaks tracks disease trends through social media. https://www.huck.psu.edu/content/crowdbreaks-tracks-disease-trends-through-social-media

[65] Augur, H. (2016, January 18). How Big Data Is Quietly Fighting Diseases and Illnesses. Dataconomy. http://dataconomy.com/2016/01/how-big-data-is-quietly-fighting-diseases-and-illnesses/

[66] Pierson, D. (2015, June 20). Banjo's Ability to Track Events in Real Time Gives Clients Competitive Edge. *Los Angeles Times*. http://www.latimes.com/business/la-fi-0621-cutting-edge-banjo-20150621-story.html

[67] Crawford, K. & Whittaker, M. (2016, Sep 22). The AI Now: Report The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term. https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf

[68] David Zaharchuk, IBM Global Industry Research, personal communication, March 2017.

[69] Crawford, K. (2013, May 10). Think Again: Big Data. *Foreign Policy.* http://foreignpolicy.com/2013/05/10/think-again-big-data/

[70] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017, March 27). Fairness in Criminal Justice Risk Assessments: The State of the Art. https://arxiv.org/abs/1703.09207v1

[71] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. https://arxiv.org/abs/1701.08230

[72] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society,3*(1), 205395171562251. doi:10.1177/2053951715622512.

[73] U.S. Department of Education. (2016, Aug. 16). FACT SHEET: ED Launches Initiative for Low-Income Students to Access New Generation of Higher Education Providers. https://www.ed.gov/news/press-releases/fact-sheet-ed-launches-initiative-low-income-students-access-new-generation-higher-education-providers

[74] Girls Who Code. https://girlswhocode.com/

[75] American Academy of Actuaries (May 2010). Risk Assessment and Risk Adjustment. https://www.actuary.org/pdf/health/Risk_Adjustment_Issue_Brief_Final_5-26-10.pdf

[76] Wachter, S., Mittelstadt, B., & Floridi, L. (2016, Dec.). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. https://www.researchgate.net/publication/312597416_Why_a_right_to_explanation_of_automated_decision-making_does_not_exist_in_the_General_Data_Protection_Regulation

[77] Eggers, W. Schatsky, D. Viechniki, P. (2017). AI-augmented government using cognitive technologies to redesign public sector work. Deloitte University Press. https://dupress.deloitte.com/content/dam/dup-us-en/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf

[78] Egan, P. (2017, March 06). Software vendor says it's not to blame in false jobless fraud findings. *Detroit Free Press*. http://www.freep.com/story/news/local/michigan/2017/03/06/software-vendor-jobless-fraud/98777050/

[79] Egan, P. (2017, January 10). 'No remedy' for unemployed falsely accused by Michigan's fraud system. *Detroit Free Press*. Retrieved May 03, 2017, from http://www.freep.com/story/news/local/michigan/2017/01/09/unempoyment-insurance-claims-fraud/96338462/

[80] Rintanen, J. & Grastien, A. (2007). Diagnosability testing with satisfiability algorithms. Proceedings of 20th International Joint Conference on Artificial Intelligence (IJCAI'2007), 532–537. https://users.ics.aalto.fi/rintanen/papers/RintanenGrastien07.pdf

CITRIS AND THE BANATAO INSTITUTE

Microsoft