TECHIMPACT®

Understanding Ethical Considerations in Artificial Intelligence

Ryan Harrington

TECHIMPACT®

Ryan Harrington

Director, Data Lab

ryanh@techimpact.org

TECHIMPACT®

Hiring Practices

Regulation Navigation

Child Protection

If you were in a position where you could decide if the organization would use the program or not, then what would you decide to do? Why?

TECHIMPACT®

Scenario #1: Hiring Practices

An organization is overwhelmed by the number of applicants that it is receiving for the roles that it has posted.

In an effort to make it easier to hire candidates, a team of data scientists has built a machine learning model that will determine whether or not a candidate should be hired based upon their resume.

Some personally identifiable information, such as the name of the candidate were not considered as a feature of the model.

slido

Please download and install the
Slido app on all computers you use

Scenario #1: Hiring Practices

ⓘ Start presenting to display the poll results on this slide.

Scenario #2: Regulation Navigation

Your organization helps people navigate the complex web of regulations that govern their lives. In an effort to make it easier for your community to find the regulation that is most meaningful for them, your team has launched a chatbot to help guide people to the appropriate information seamlessly.

You limit the information that the chatbot is allowed to consider in its responses, focusing specifically on regulations for organizational operations.

slido

Please download and install the
Slido app on all computers you use

Scenario #2: Regulation Navigation

ⓘ Start presenting to display the poll results on this slide.

TECHIMPACT®

Scenario #3: Child Protection

Child welfare workers are asked to make thousands of decisions in any given year. This is an overwhelming number of decisions to make - often with imperfect information. A county's Department of Health and Human Services built a model that can aid decision making for its child welfare workers. Each case is given a score that indicates how risky it is based upon the likelihood of the child to be removed from the home within 2 years.

Incidents of potential neglect are reported to the county's child protection hotline. The reports go through a screening process where the algorithm calculates the child's potential risk and assigns a score. Child welfare workers then use their discretion to decide whether to investigate.

slido

Please download and install the
Slido app on all computers you use

Scenario #3: Child Protection

ⓘ Start presenting to display the poll results on this slide.

TECHIMPACT®

What is artificial intelligence?

TECHIMPACT®

Machine Learning

Robotics

Expert Systems

Vision

Planning

Speech

Natural Language Processing

TECHIMPACT®

Machine Learning

Robotics

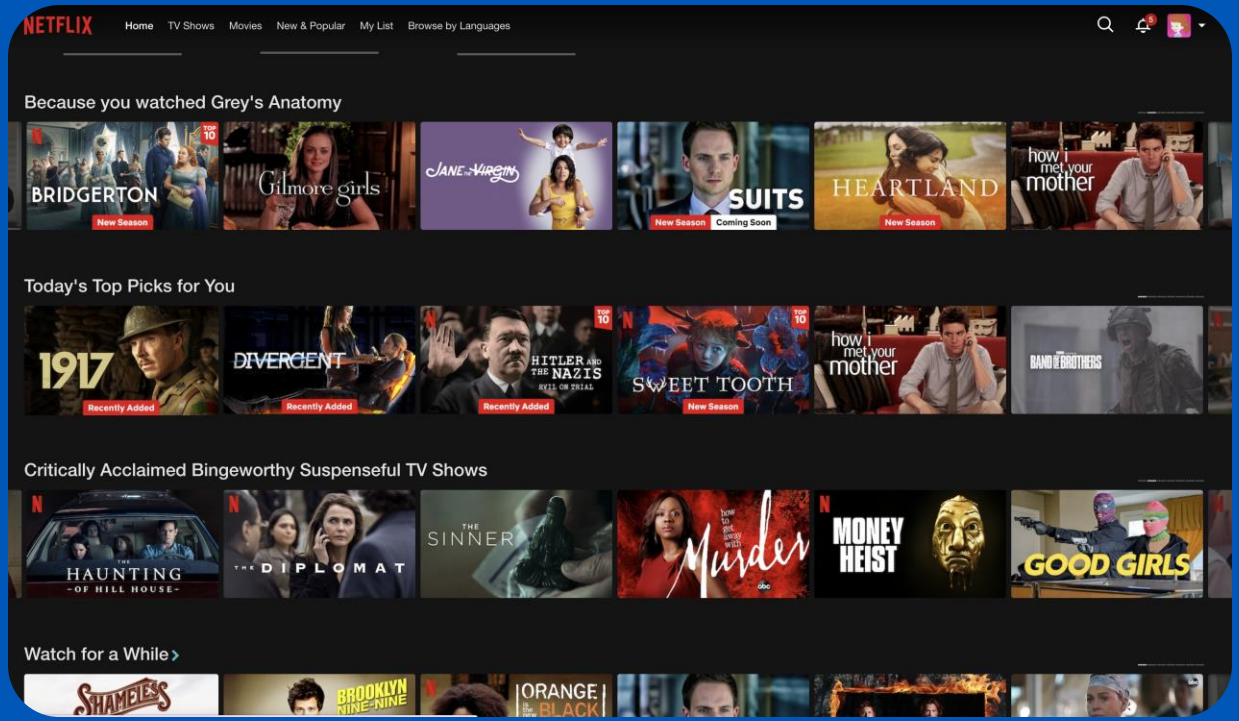
Expert Systems

Vision

Planning

Speech

Natural Language Processing



TECHIMPACT®

Machine Learning

Robotics

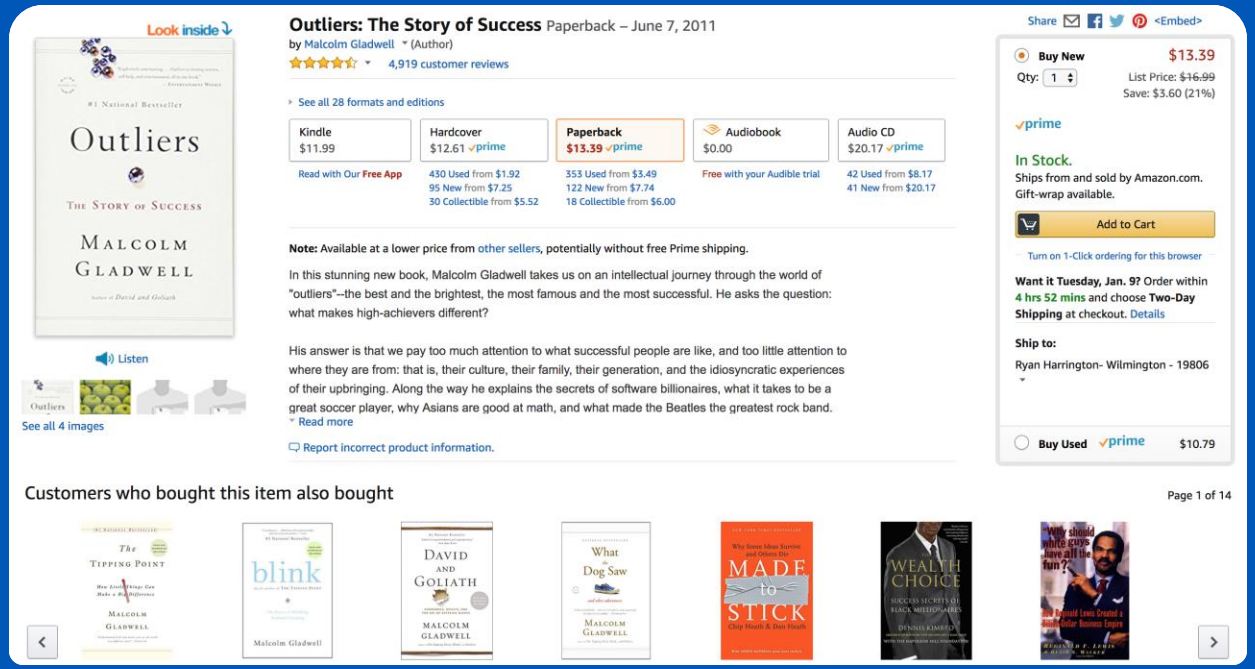
Expert Systems

Vision

Planning

Speech

Natural Language Processing



TECHIMPACT®

Machine Learning

Robotics

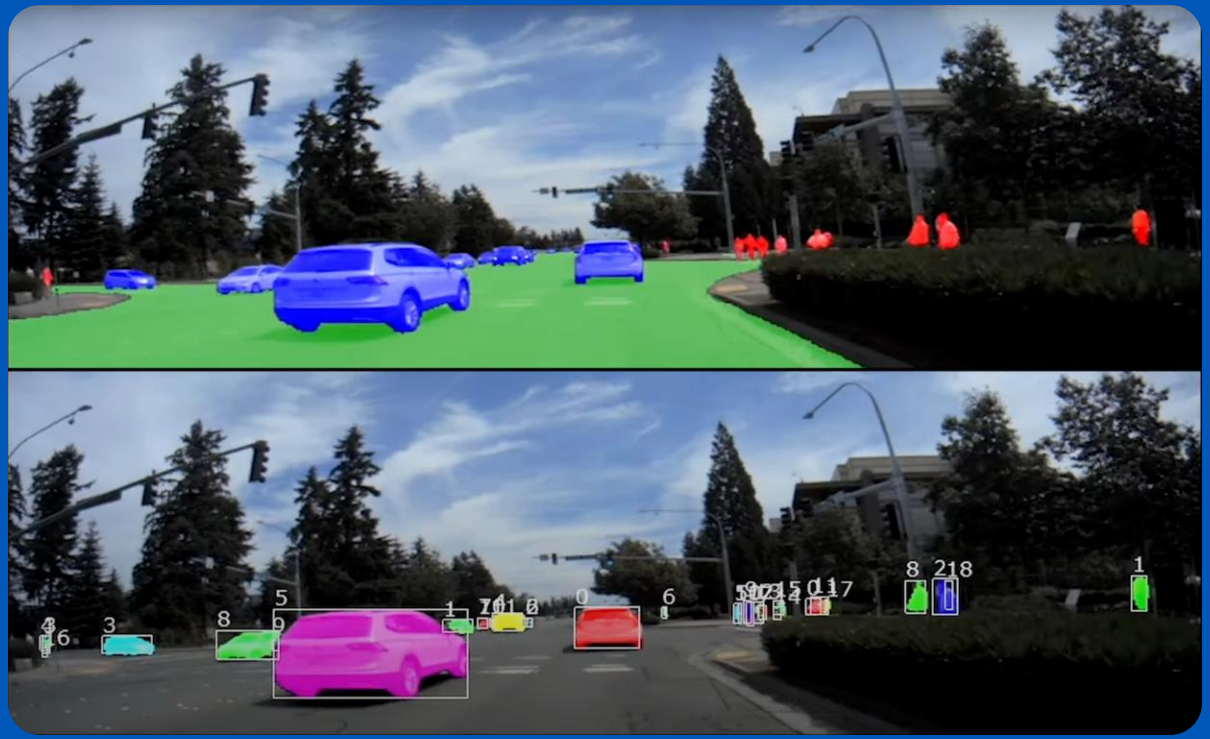
Expert Systems

Vision

Planning

Speech

Natural Language Processing



TECHIMPACT®

Machine Learning

Robotics

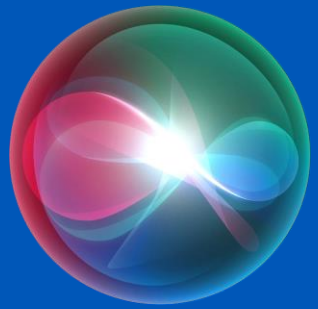
Expert Systems

Vision

Planning

Speech

Natural Language Processing



alexa

TECHIMPACT®

Machine Learning

Robotics

Expert Systems

Vision

Planning

Speech

Natural Language Processing

Artificial Intelligence

TECHIMPACT®

Artificial Intelligence

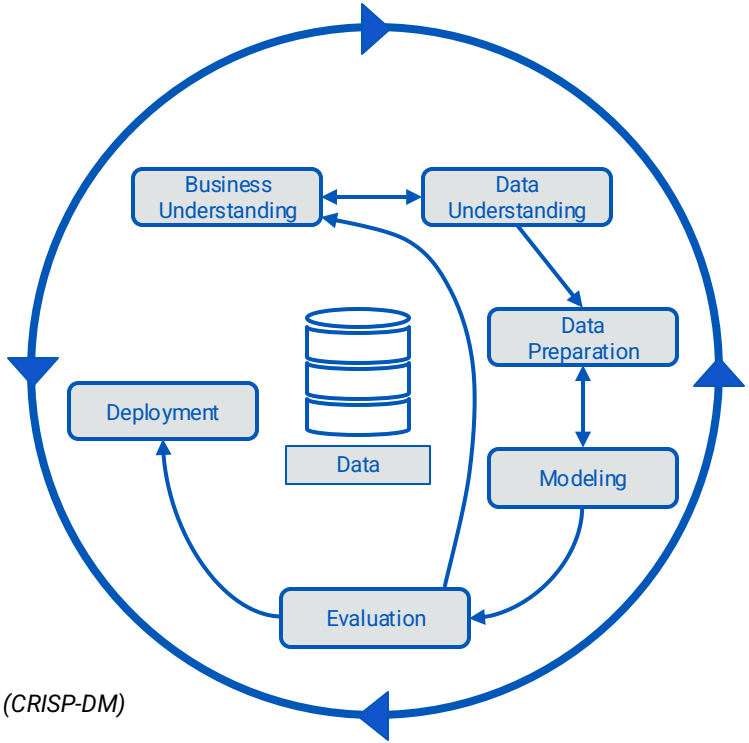
↓

Machine Learning

↓

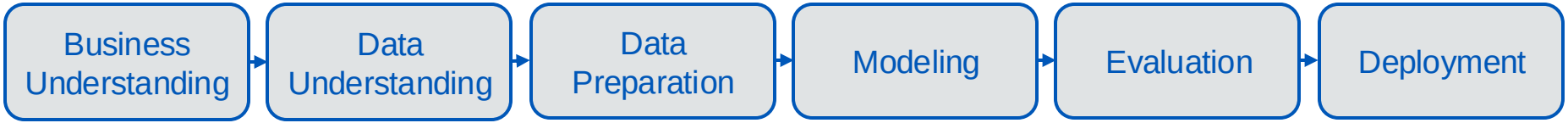
Data Science

The Typical Data Science Process



Cross Industry Standard Process for Data Mining (CRISP-DM)

The Typical Data Science Process



Cross Industry Standard Process for Data Mining (CRISP-DM)

Questions to Consider

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

How diverse and familiar with the development context are the group of people defining the problem?

Cross Industry Standard Process for Data Mining (CRISP-DM)

Questions to Consider

Business Understanding → **Data Understanding** → Data Preparation → Modeling → Evaluation → Deployment

What might need to be done to improve representativeness of data?

Cross Industry Standard Process for Data Mining (CRISP-DM)

Questions to Consider

Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

What are the protected attributes for this context or problem?

Cross Industry Standard Process for Data Mining (CRISP-DM)

TECHIMPACT®

Questions to Consider

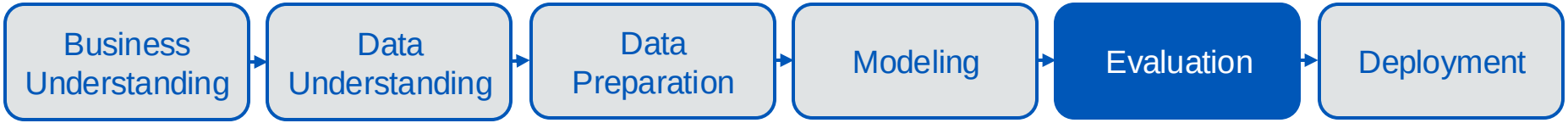
Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

What potential biases will the algorithm introduce?

How well can individual decisions or predictions be explained in human-friendly terms?

Cross Industry Standard Process for Data Mining (CRISP-DM)

Questions to Consider

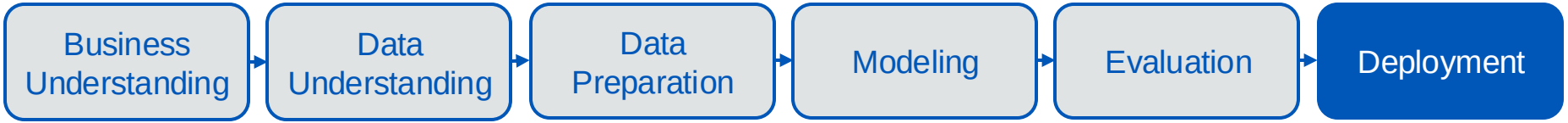


How can you best implement fairness?

What tradeoff between model accuracy and equity is appropriate for my context?

Cross Industry Standard Process for Data Mining (CRISP-DM)

Questions to Consider



Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

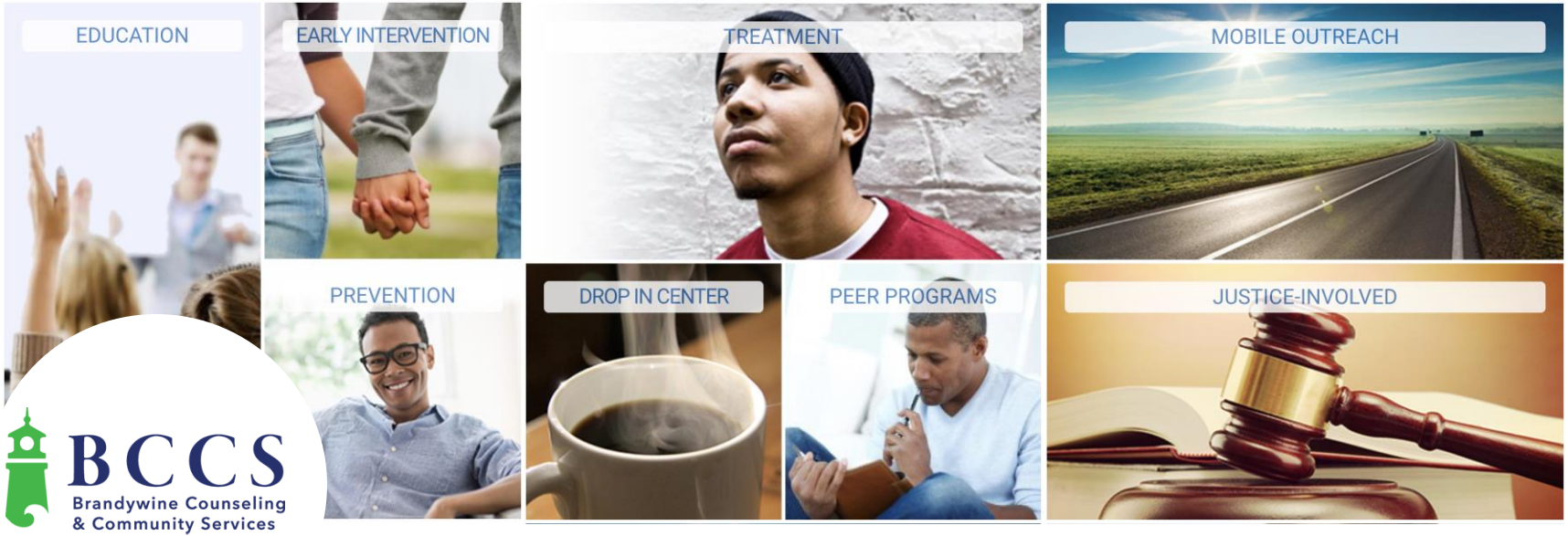
What mechanisms will be put into place to audit models over time and enhance accountability for model results?

Cross Industry Standard Process for Data Mining (CRISP-DM)

TECHIMPACT®

Case Study:
Opiate Counseling Attrition

TECHIMPACT®



BCCS
Brandywine Counseling
& Community Services

How can we identify which individuals involved in opiate counseling are most likely to not complete the program?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

How diverse and familiar with the development context are the group of people defining the problem?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment

How diverse and familiar with the development context are the group of people defining the problem?

Partnered with treatment leaders at BCCS to define and iterate upon the problem

Incorporated voices of leaders and case workers into understanding of the problem

Did not include voices of individuals undergoing treatment into user research

TECHIMPACT®

Business Understanding → **Data Understanding** → Data Preparation → Modeling → Evaluation → Deployment

What might need to be done to improve representativeness of data?

TECHIMPACT®

Business Understanding → **Data Understanding** → Data Preparation → Modeling → Evaluation → Deployment

What might need to be done to improve representativeness of data?

Utilized all (relevant) historical data from BCCS databases

No major considerations for this

TECHIMPACT®

Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

What are the protected attributes for this context or problem?

TECHIMPACT®

Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

What are the protected attributes for this context or problem?

Primarily considerate of:

- Race
- Gender

Benchmark historical attrition rates based upon protected classes

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

What potential biases will the algorithm introduce?

How well can individual decisions or predictions be explained in human-friendly terms?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

What potential biases will the algorithm introduce?

Could exacerbate any existing discrepancies in treatment for individuals that the model does not identify

How well can individual decisions or predictions be explained in human-friendly terms?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

What potential biases will the algorithm introduce?	Could exacerbate any existing discrepancies in treatment for individuals that the model does not identify
How well can individual decisions or predictions be explained in human-friendly terms?	Ultimately selected an algorithm with a medium level of interpretability due to non-linear interactions

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → **Evaluation** → Deployment

How can you best implement fairness?

What tradeoff between model accuracy and equity is appropriate for my context?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → **Evaluation** → Deployment

How can you best implement fairness?

Ensuring similar accuracy metrics across subgroups of vulnerable populations in comparison with selected privileged subgroups

What tradeoff between model accuracy and equity is appropriate for my context?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → **Evaluation** → Deployment

How can you best implement fairness?	Ensuring similar accuracy metrics across subgroups of vulnerable populations in comparison with selected privileged subgroups
What tradeoff between model accuracy and equity is appropriate for my context?	Determined that accuracy was more important than fairness for this use case

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → **Deployment**

Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

What mechanisms will be put into place to audit models over time and enhance accountability for model results?

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → **Deployment**

Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

What mechanisms will be put into place to audit models over time and enhance accountability for model results?

Deploying a humans in the loop methodology, which allows case managers to see if a patient is at risk and act upon it

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → **Deployment**

Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

Deploying a humans in the loop methodology, which allows case managers to see if a patient is at risk and act upon it

What mechanisms will be put into place to audit models over time and enhance accountability for model results?

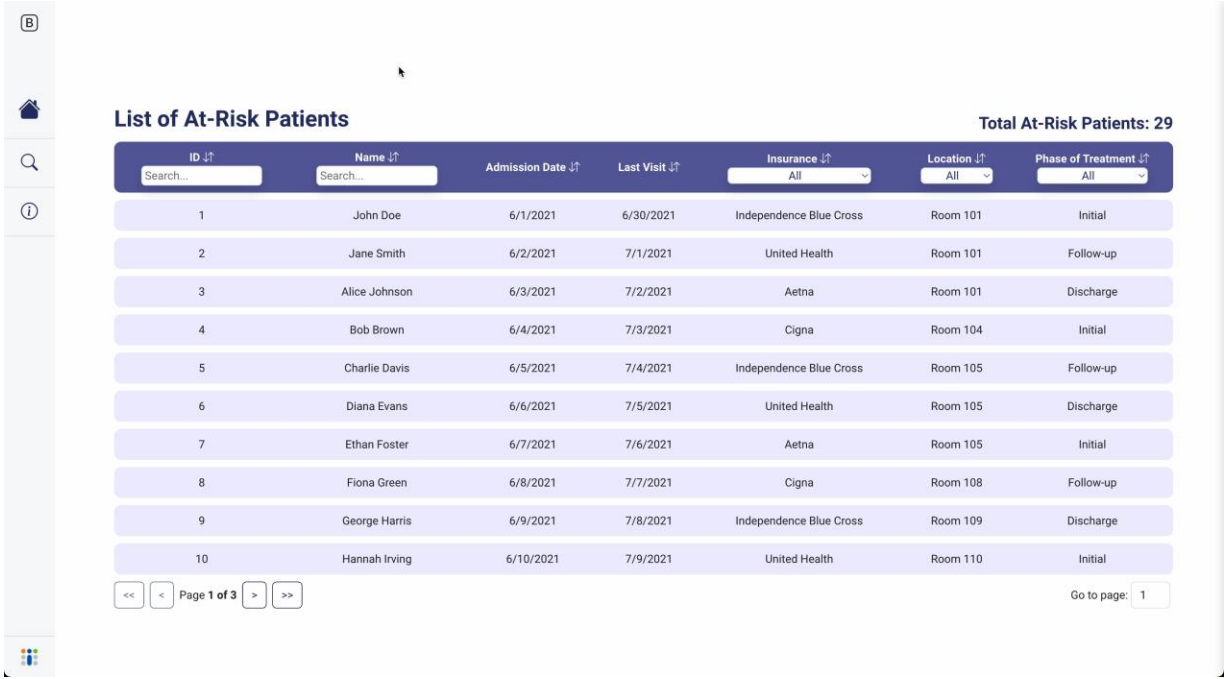
Developing a model evaluation framework which includes regular monitoring based upon accuracy and identified ethical risks

TECHIMPACT®

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → **Deployment**

Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

What mechanisms will be put into place to audit models over time and enhance accountability for model results?



List of At-Risk Patients

Total At-Risk Patients: 29

ID	Name	Admission Date	Last Visit	Insurance	Location	Phase of Treatment
Search...	Search...			All	All	All
1	John Doe	6/1/2021	6/30/2021	Independence Blue Cross	Room 101	Initial
2	Jane Smith	6/2/2021	7/1/2021	United Health	Room 101	Follow-up
3	Alice Johnson	6/3/2021	7/2/2021	Aetna	Room 101	Discharge
4	Bob Brown	6/4/2021	7/3/2021	Cigna	Room 104	Initial
5	Charlie Davis	6/5/2021	7/4/2021	Independence Blue Cross	Room 105	Follow-up
6	Diana Evans	6/6/2021	7/5/2021	United Health	Room 105	Discharge
7	Ethan Foster	6/7/2021	7/6/2021	Aetna	Room 105	Initial
8	Fiona Green	6/8/2021	7/7/2021	Cigna	Room 108	Follow-up
9	George Harris	6/9/2021	7/8/2021	Independence Blue Cross	Room 109	Discharge
10	Hannah Irving	6/10/2021	7/9/2021	United Health	Room 110	Initial

<< < Page 1 of 3 > >>

Go to page: 1

TECHIMPACT®

Hiring Practices

Regulation Navigation

Child Protection

Questions to Consider

If you were in a position where you could decide if the organization would use the program or not, then what would you decide to do? Why?

TECHIMPACT®

Questions to Consider

If you were in a position where you could decide if the organization would use the program or not, then what would you decide to do? Why?

How diverse and familiar with the development context are the group of people defining the problem?

What might need to be done to improve representativeness of data?

What are the protected attributes for this context or problem?

Given the equity of outcomes in practice, representativeness, and explainability, how will model predictions be used in practice?

What potential biases will the algorithm introduce?

How can you best implement fairness?

How well can individual decisions or predictions be explained in human-friendly terms?

What tradeoff between model accuracy and equity is appropriate for my context?

What mechanisms will be put into place to audit models over time and enhance accountability for model results?

TECHIMPACT®

Questions to Consider

If you were in a position where you could decide if the organization would use the program or not, then what would you decide to do? Why?

de-data-lab.github.io/
ai-ethics-scenarios

TECHIMPACT®

Scenario #1: Hiring Practices

An organization is overwhelmed by the number of applicants that it is receiving for the roles that it has posted.

In an effort to make it easier to hire candidates, a team of data scientists has built a machine learning model that will determine whether or not a candidate should be hired based upon their resume.

Some personally identifiable information, such as the name of the candidate were not considered as a feature of the model.

TECHIMPACT®

Scenario #1: Hiring Practices

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Source

TECHIMPACT®

Scenario #1: Hiring Practices

That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. **Most came from men, a reflection of male dominance across the tech industry.**

Source

TECHIMPACT®

Scenario #1: Hiring Practices

In effect, Amazon's system taught itself that male candidates were preferable. **It penalized resumes that included the word "women's," as in "women's chess club captain."** And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Source

TECHIMPACT®

Scenario #1: Hiring Practices

Amazon edited the programs to make them neutral to these particular terms. But that was **no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory,** the people said.

Source

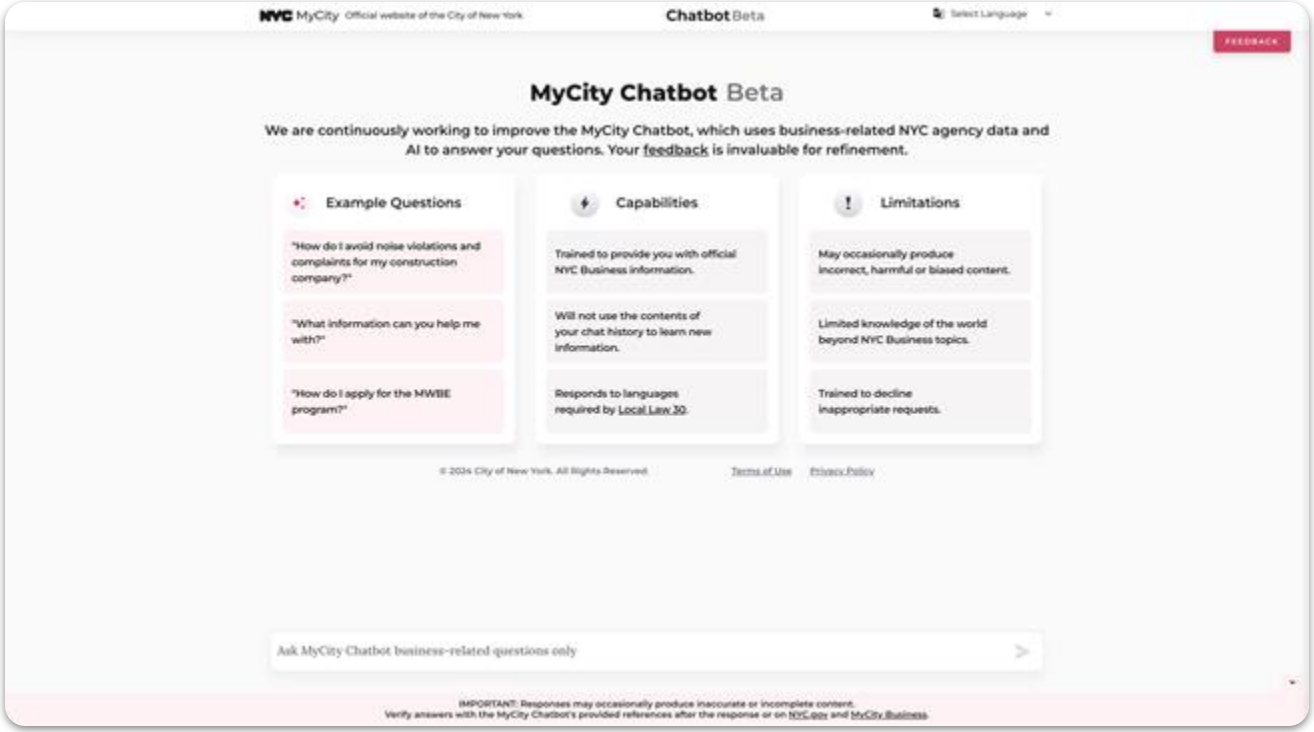
Scenario #2: Regulation Navigation

Your organization helps people navigate the complex web of regulations that govern their lives. In an effort to make it easier for your community to find the regulation that is most meaningful for them, your team has launched a chatbot to help guide people to the appropriate information seamlessly.

You limit the information that the chatbot is allowed to consider in its responses, focusing specifically on regulations for organizational operations.

TECHIMPACT®

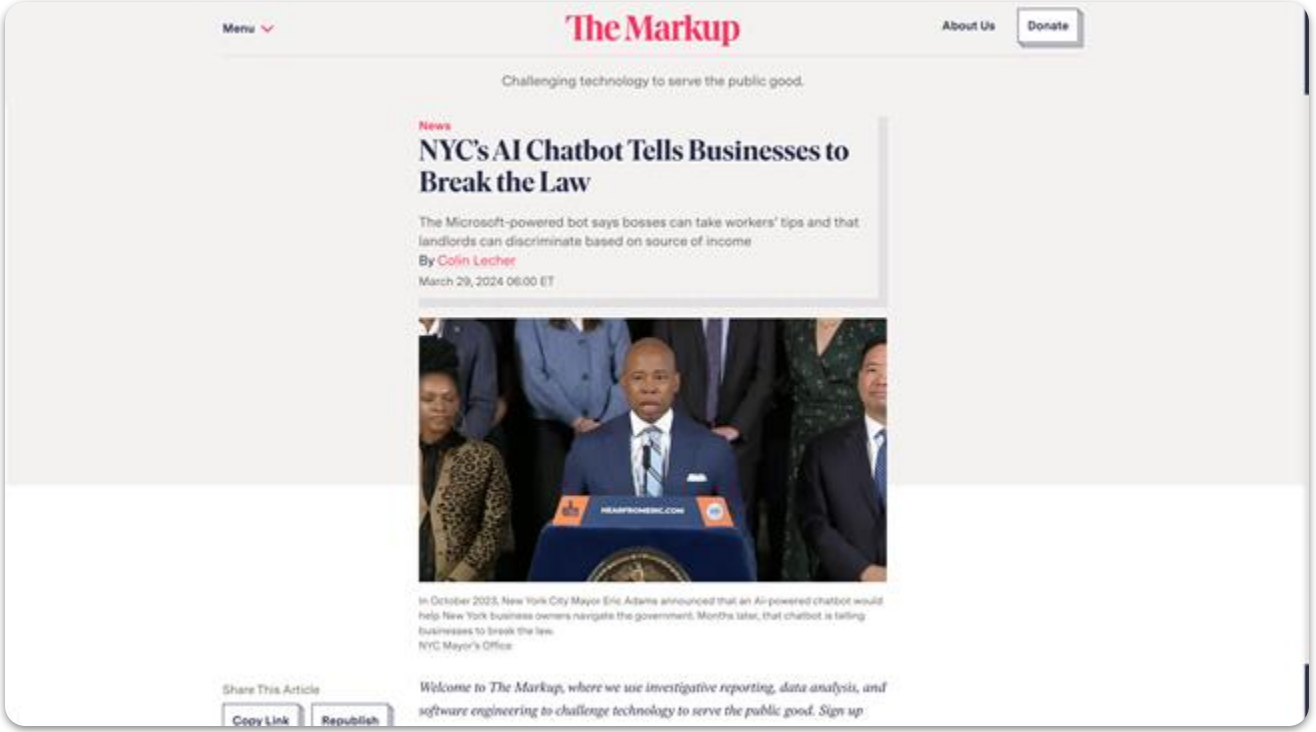
Scenario #2: Regulation Navigation



Source

TECHIMPACT®

Scenario #2: Regulation Navigation



Menu

The Markup

About Us

Donate

Challenging technology to serve the public good.

News

NYC's AI Chatbot Tells Businesses to Break the Law

The Microsoft-powered bot says bosses can take workers' tips and that landlords can discriminate based on source of income

By Colin Lecher

March 29, 2024 06:00 ET

In October 2023, New York City Mayor Eric Adams announced that an AI-powered chatbot would help New York business owners navigate the government. Months later, that chatbot is telling businesses to break the law.

NYC Mayor's Office

Share This Article

Copy Link

Republish

Welcome to The Markup, where we use investigative reporting, data analysis, and software engineering to challenge technology to serve the public good. Sign up

Source

TECHIMPACT®

Scenario #2: Regulation Navigation

Five months after launch, it's clear that while the bot appears authoritative, the information it provides on housing policy, worker rights, and rules for entrepreneurs is often incomplete and in worst-case scenarios **"dangerously inaccurate,"** as one local housing policy expert told The Markup.

Source

TECHIMPACT®

Scenario #2: Regulation Navigation

Question Submitted	NYC Chatbot Answer	Reality
Are buildings required to accept section 8 vouchers?	“No, buildings are not required to accept Section 8 vouchers.”	Landlords cannot discriminate by source of income, with a minor exception for small buildings where the landlord or their family lives.

Source

TECHIMPACT®

Scenario #2: Regulation Navigation

Question Submitted	NYC Chatbot Answer	Reality
Are buildings required to accept section 8 vouchers?	“No, buildings are not required to accept Section 8 vouchers.”	Landlords cannot discriminate by source of income, with a minor exception for small buildings where the landlord or their family lives.
Do landlords have to accept tenants on rental assistance?	“No, landlords are not required to accept tenants on rental assistance.”	Landlords cannot discriminate by source of income, with a minor exception for small buildings where the landlord or their family lives.

Source

TECHIMPACT®

Scenario #2: Regulation Navigation

Question Submitted	NYC Chatbot Answer	Reality
Are buildings required to accept section 8 vouchers?	“No, buildings are not required to accept Section 8 vouchers.”	Landlords cannot discriminate by source of income, with a minor exception for small buildings where the landlord or their family lives.
Do landlords have to accept tenants on rental assistance?	“No, landlords are not required to accept tenants on rental assistance.”	Landlords cannot discriminate by source of income, with a minor exception for small buildings where the landlord or their family lives.
Can I make my store cashless?	“Yes, you can make your store cashless in New York City. There are no specific requirements that prohibit businesses from operating as cashless establishments in the city.”	Stores have been required to accept cash as payment since 2020.

Source

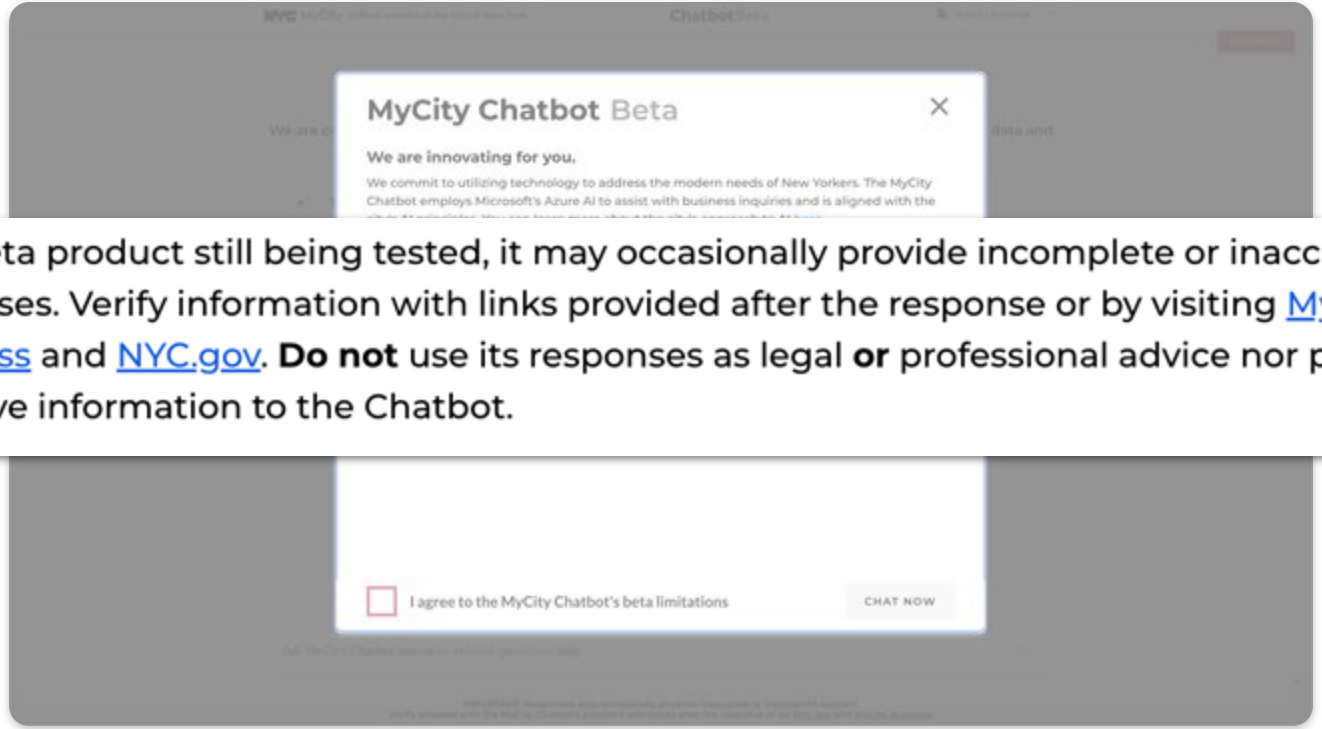
TECHIMPACT®

Scenario #2: Regulation Navigation

Source

TECHIMPACT®

Scenario #2: Regulation Navigation



As a beta product still being tested, it may occasionally provide incomplete or inaccurate responses. Verify information with links provided after the response or by visiting MyCity Business and NYC.gov. **Do not** use its responses as legal **or** professional advice nor provide sensitive information to the Chatbot.

Source

TECHIMPACT®

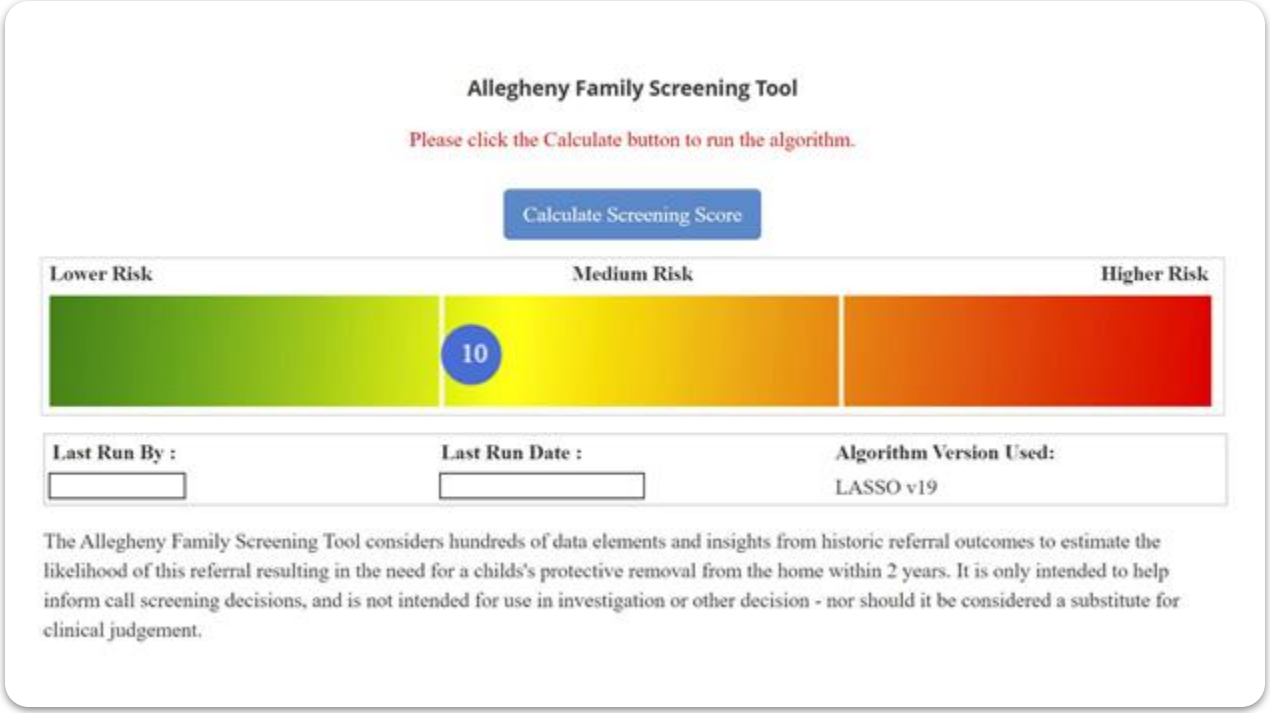
Scenario #3: Child Protection

Child welfare workers are asked to make thousands of decisions in any given year. This is an overwhelming number of decisions to make - often with imperfect information. A county's Department of Health and Human Services built a model that can aid decision making for its child welfare workers. Each case is given a score that indicates how risky it is based upon the likelihood of the child to be removed from the home within 2 years.

Incidents of potential neglect are reported to the county's child protection hotline. The reports go through a screening process where the algorithm calculates the child's potential risk and assigns a score. Child welfare workers then use their discretion to decide whether to investigate.

TECHIMPACT®

Scenario #3: Child Protection



Source

TECHIMPACT®

Scenario #3: Child Protection

An algorithm that screens for child neglect raises concerns

By SALLY HO and GARANCE BURKE
April 29, 2022

Inside a cavernous stone fortress in downtown Pittsburgh, attorney Robin Frank defends parents at one of their lowest points – when they risk losing their children.

The job is never easy, but in the past she knew what she was up against when squaring off against child protective services in family court. Now, she worries she's fighting something she can't see: an opaque algorithm whose statistical calculations help social workers decide which families should be investigated in the first place.

Source

TECHIMPACT®

Scenario #3: Child Protection

According to new research from a Carnegie Mellon University team obtained exclusively by AP, Allegheny's algorithm in its first years of operation showed a **pattern of flagging a disproportionate number of Black children for a "mandatory" neglect investigation, when compared with white children.** The independent researchers, who received data from the county, also found that social workers disagreed with the risk scores the algorithm produced about one-third of the time.

Source

TECHIMPACT®

Scenario #3: Child Protection

If the tool had acted on its own to screen in a comparable rate of calls, **it would have recommended that two-thirds of Black children be investigated, compared with about half of all other children reported,** according to another study published last month and co-authored by a researcher who audited the county's algorithm.

Source

TECHIMPACT®

Scenario #3: Child Protection

Given the high stakes – skipping a report of neglect could end with a child's death but scrutinizing a family's life could set them up for separation – **the county and developers have suggested their tool can help "course correct" and make the agency's work more thorough and efficient by weeding out meritless reports** so that social workers can focus on children who truly need protection.

The developers have described using such tools as a moral imperative, saying child welfare officials should use whatever they have at their disposal to make sure children aren't neglected.

Source

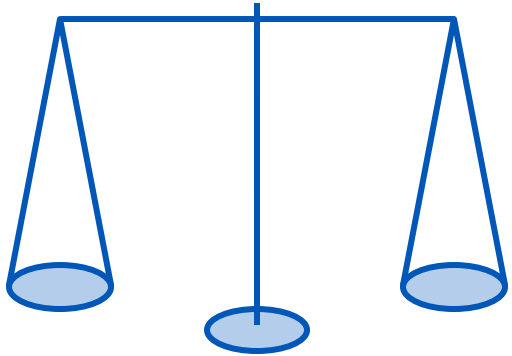
TECHIMPACT®

So, what now?

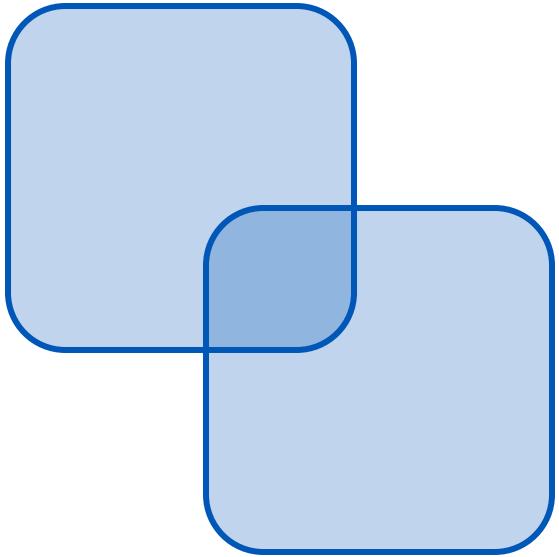
TECHIMPACT®

G	Garbage
I	In,
G	Garbage
O	Out

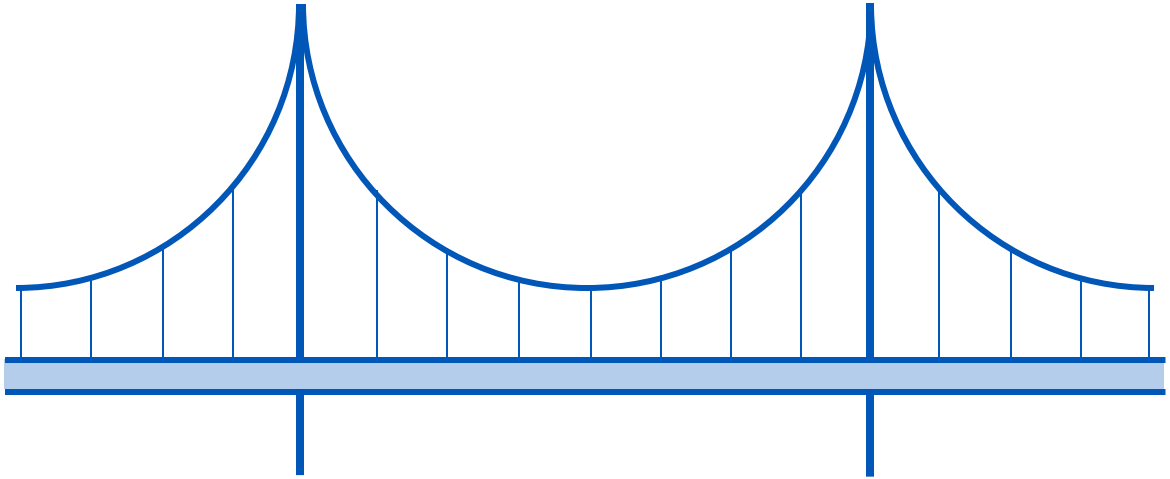
Fairness



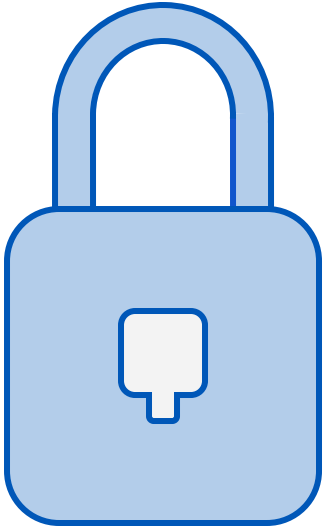
Transparency



Robustness



Privacy



TECHIMPACT®

Understanding Ethical Considerations in Artificial Intelligence

Ryan Harrington
ryanh@techimpact.org

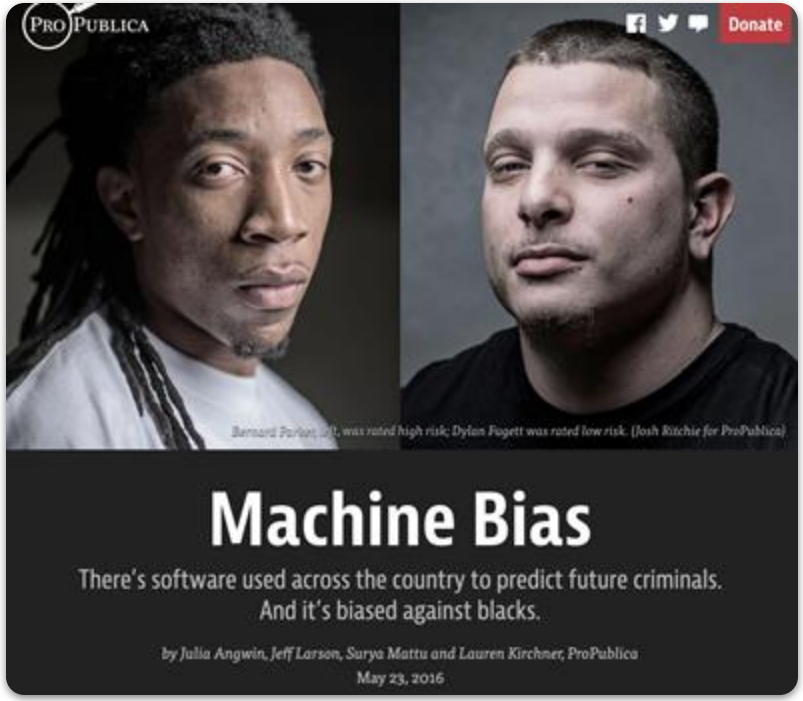
Scenario #4: Recidivism

Judges make decisions about the sentences that criminals receive. One of the factors that judges consider during sentencing is the likelihood of the person to re-offend (recidivism).

Courtrooms have adopted tools designed to eliminate bias in sentencing through the use of artificial intelligence. The history of the criminal can be input into the model. It will then output the likelihood of the person to re-offend.

Demographic information about the criminal is not included in the model.

Scenario #4: Recidivism



Source

TECHIMPACT®

Scenario #4: Recidivism

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use.

“Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice,” he said, adding, “they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”

Source

TECHIMPACT®

Scenario #4: Recidivism

The score proved remarkably unreliable in forecasting violent crime: **Only 20 percent of the people predicted to commit violent crimes actually went on to do so.**

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

Source

TECHIMPACT®

Scenario #4: Recidivism

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to **falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.**

- White defendants were mislabeled as low risk more often than black defendants.

Source

TECHIMPACT®

Scenario #4: Recidivism

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. **In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.**

Source