Embedded EthiCS: Designing Responsible AI

CS 182 | Fall 2022

Hello!

I am Michael Pope

Graduate Fellow

Embedded EthiCS @ Harvard

Agenda

- What does it mean to be responsible?
- **Case Study 1:** Allegheny Family Screening Tool
- **Case Study 2:** Hello Baby
- Applying ethical lenses for responsible design

Embedded EthiCS

Identify ethical & social issues

Reason through ethical & social issues

Communicate reasoned positions

Design ethically & socially responsible systems

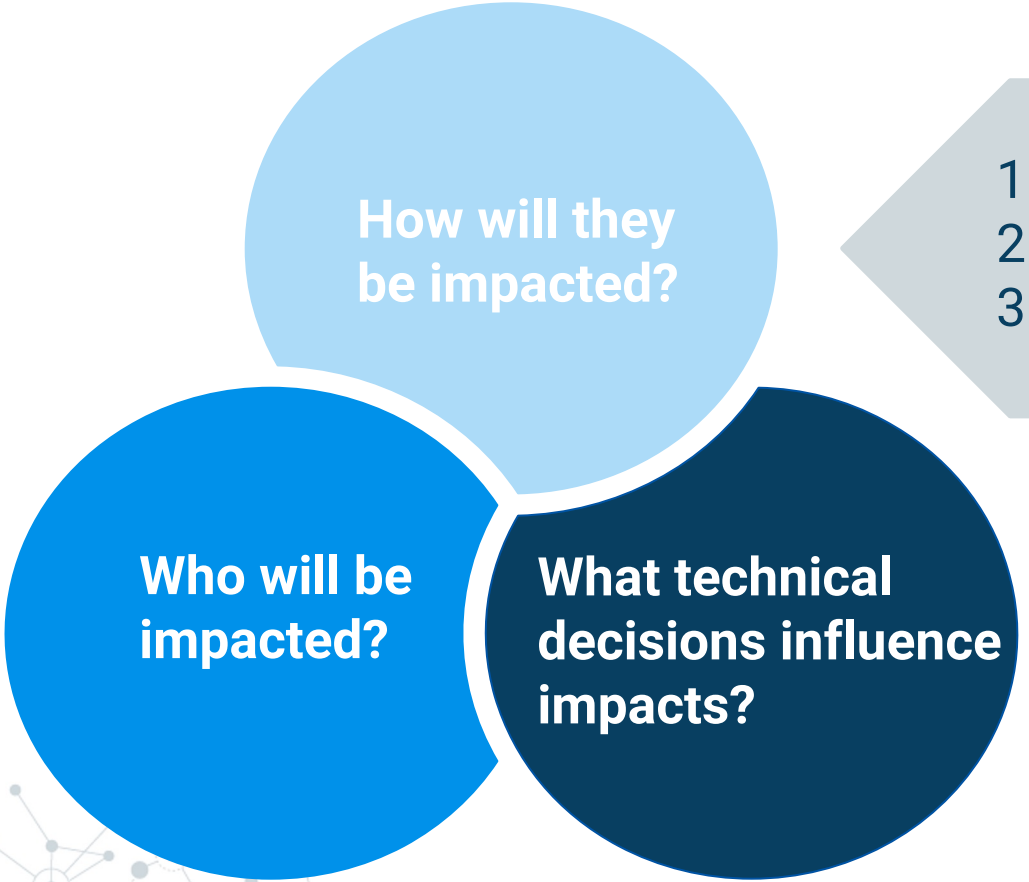
Being Responsible

How can we think about AI systems responsibly?

“Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider **impacts** of their work, consistently supporting the **public good**.”

– *Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct*

Designing Responsibly



1. Benefits/harms
2. Respect
3. Justice

1. Data
2. Design
3. Deployment

Case Study 1:
The Allegheny Family Screening Tool (AFST)

The AFST is a predictive risk modeling tool that predicts the likelihood of future child maltreatment.

How the AFST works at the DHS:

1. Potential incident of concern

2. Call placed to CYF about possible child abuse or neglect

3. A Screener takes the call, collects information, and decides to:

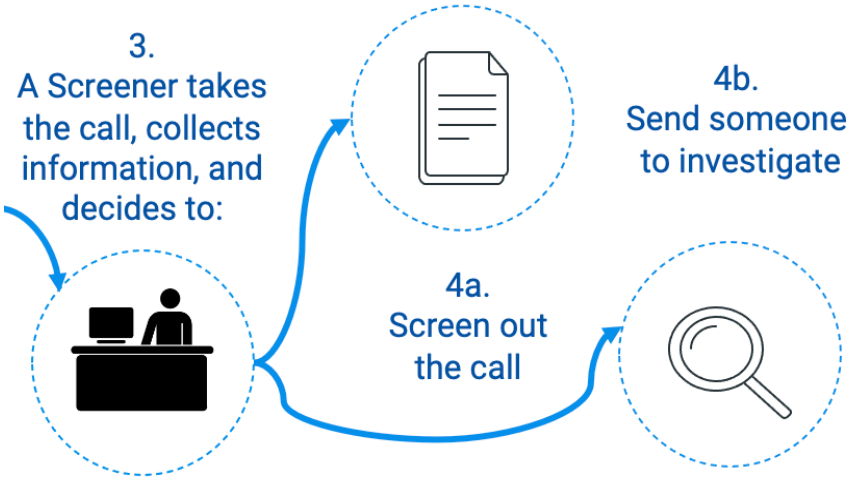
4a. Screen out the call

4b. Send someone to investigate

Case Study 1: AFST

The AFST aims to **improve** human assessment in call screening

- Limited resources and overwhelming workload
- Inability to access and integrate all available information
- Address idiosyncrasies and biased decisions



Case Study 1: AFST Data

Data Sources:

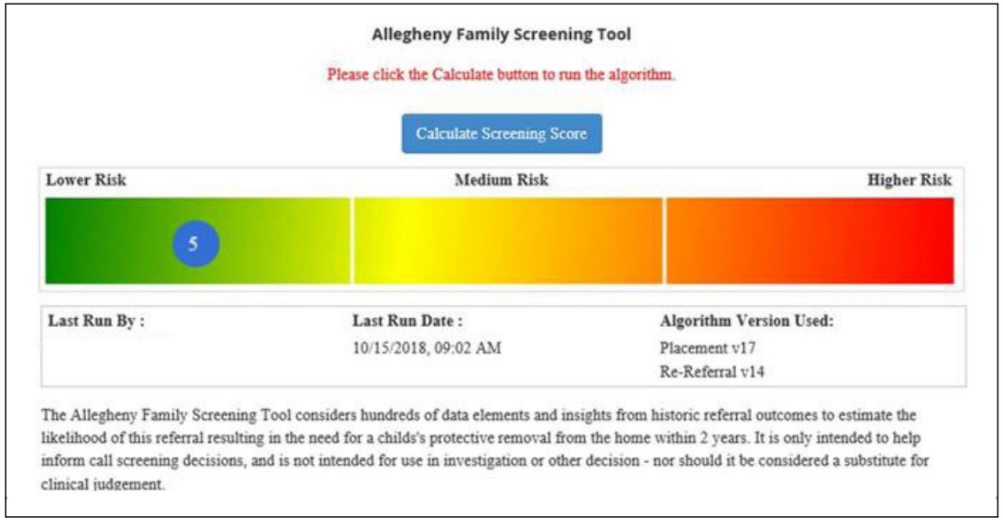
- Child welfare records
- Jail records
- Juvenile probation records
- Public benefit records
- Behavioral health records

Case Study 1: AFST

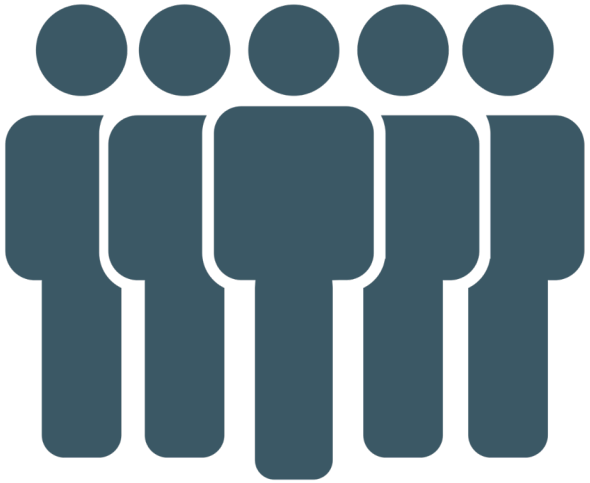
Output based on two predictions:

1. Whether the child would be re-referred within two years

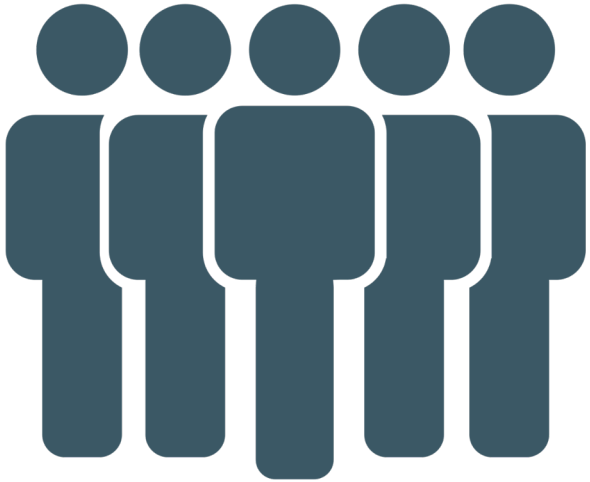
2. Whether the child will be removed from their home within two years



Who will be impacted by the AFST?



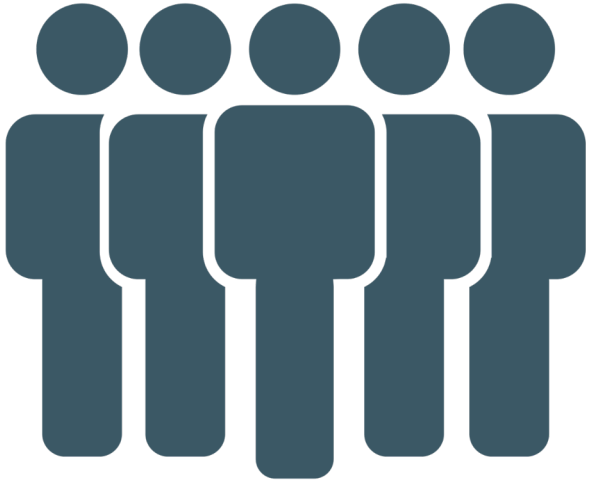
Who will be impacted by the AFST?



1. Benefits/harms

What are the potential consequences of this system for each stakeholder?

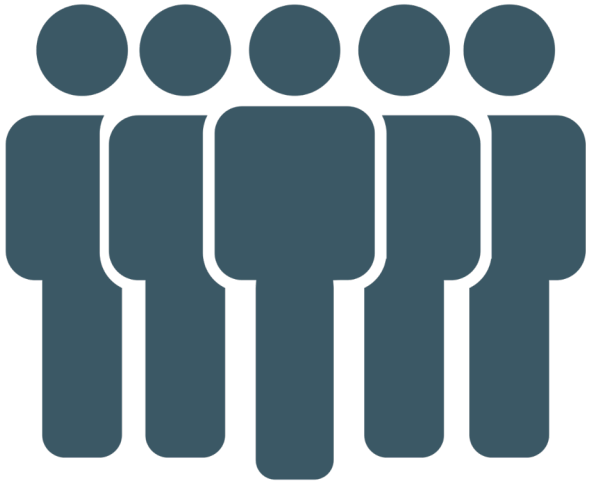
Who will be impacted by the AFST?



1. **Benefits/harms**
2. **Respect**

How does this system show respect for each stakeholder's autonomy (think: transparency, consent, control, etc.)?

Who will be impacted by the AFST?



1. **Benefits/harms**
2. **Respect**
3. **Justice**

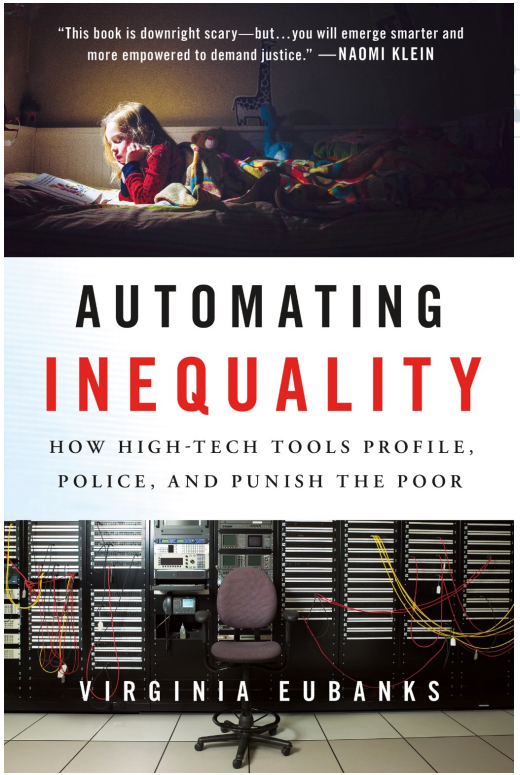
Does this process treat each stakeholder fairly? Does this process lead to fair outcomes?

Who will be impacted by the AFST?

1. Benefits/harms
2. Respect
3. Justice

Case Study 1:
Evaluating the AFST

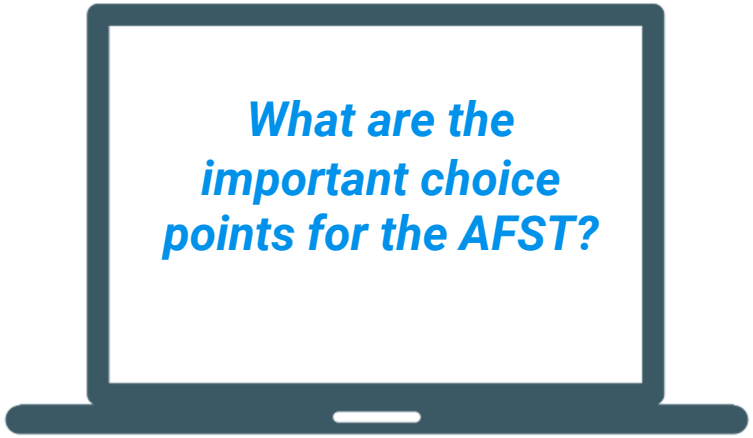
"Once the big blue button is clicked and the AFST runs, it manifests a thousand invisible human choices. But it does so under a cloak of evidence-based objectivity and infallibility."



What technical choices influence these impacts?

- **Data**
  - Which data sources am I using?
- **Design**
  - How am I defining my objectives?
  - How did I build the model?
- **Deployment**
  - How will users interact with the system?

What technical choices influence these impacts?



- **Data**
  - Which data sources am I using?
- **Design**
  - How am I defining my objectives?
  - How did I build the model?
- **Deployment**
  - How will users interact with the system?

Taking Responsibility: AFST Version 2

Data: No longer uses public benefits records as a data source

Design: No longer predicts re-referrals, only out-of-home placement

Deployment: Defaults to automatically screening out low predicted-risk calls and screening in high predicted-risk calls, with the option for human screeners to override

Case Study 2: Hello Baby (HB)

The Hello Baby program aims to engage and support new parents in Allegheny County. The program uses a PRM to place parents and children in three support tiers according to predicted risk of maltreatment.

How Hello Baby works:

1. Child Born

2. After an opt-out period, a PRM places families into one of three support tiers.

Priority Tier

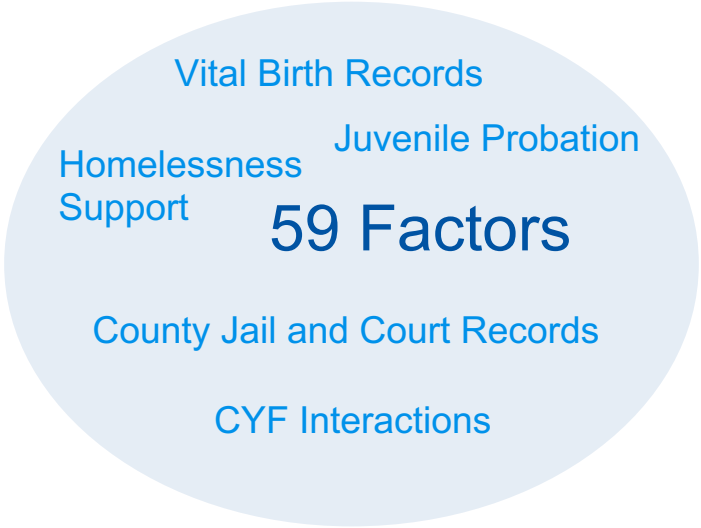
Family Support Tier

Universal Tier

Case Study 2: Hello Baby

Universal (non-means-tested) Data:

"This is a universal program," explained Emily Putnam-Hornstein. "In the [child services] model the county was being forced to make a decision after an allegation had been received; in this case we're talking about more proactively using data … so we wanted that to be built around universally available data."



Case Study 2: Hello Baby

The model predicts whether, within three years, a child will be removed from the home.

92.4% accuracy

For a sample of 13,155 births, the model placed 662 into Priority Tier—representing 54% of children removed from the home.

Vital Birth Records

Juvenile Probation

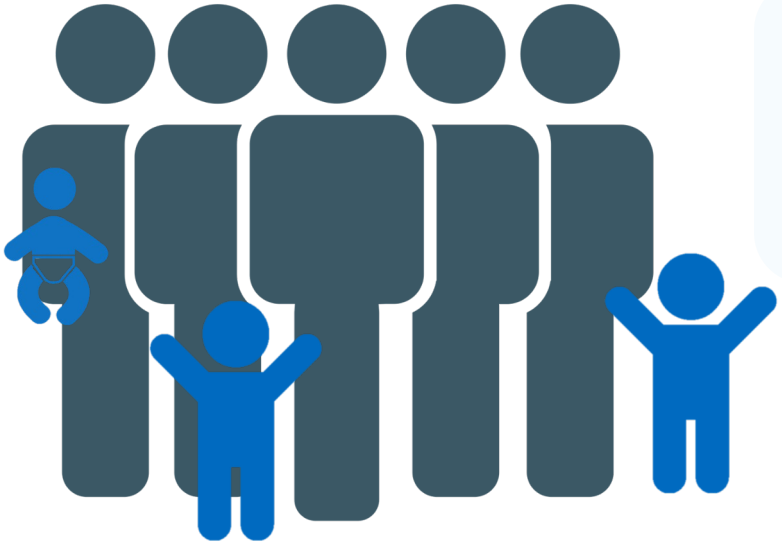
Homelessness Support

59 Factors

County Jail and Court Records

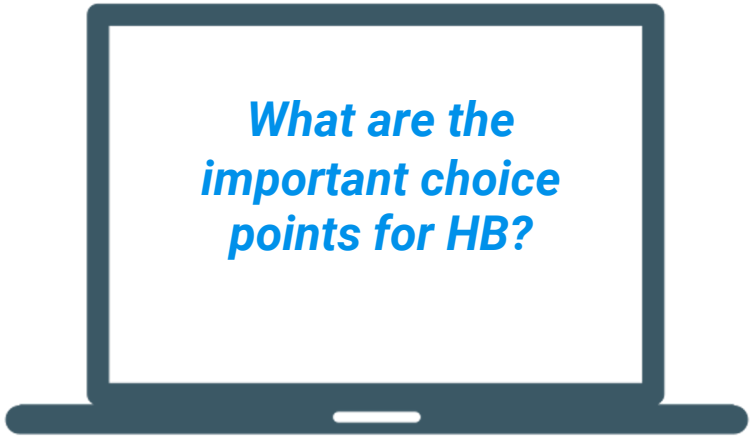
CYF Interactions

Who will be impacted by Hello Baby?



1. Benefits/harms
2. Respect
3. Justice

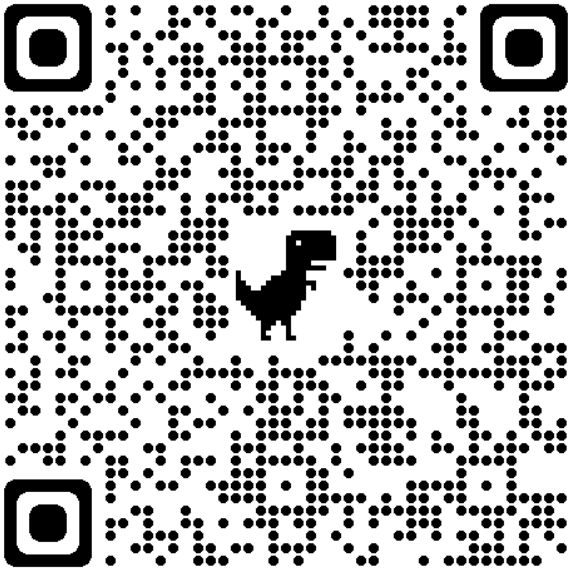
What technical choices influence these impacts?



- **Data**
  - Which data sources am I using?
- **Design**
  - How am I defining my objectives?
  - How did I build the model?
- **Deployment**
  - How will users interact with the system?

Stepping Back: When and How to Use AI Systems Responsibly and Effectively

Thanks!



https://tinyurl.com/CS182F22