



# Ethical Builders In The Age of AI

Kasia Chmielinski | October 2024

2010



Image: Firefly

What is **my responsibility**  
as a builder of  
AI systems?



HELLO!

# I'm Kasia

PRONOUNS: THEY / THEM

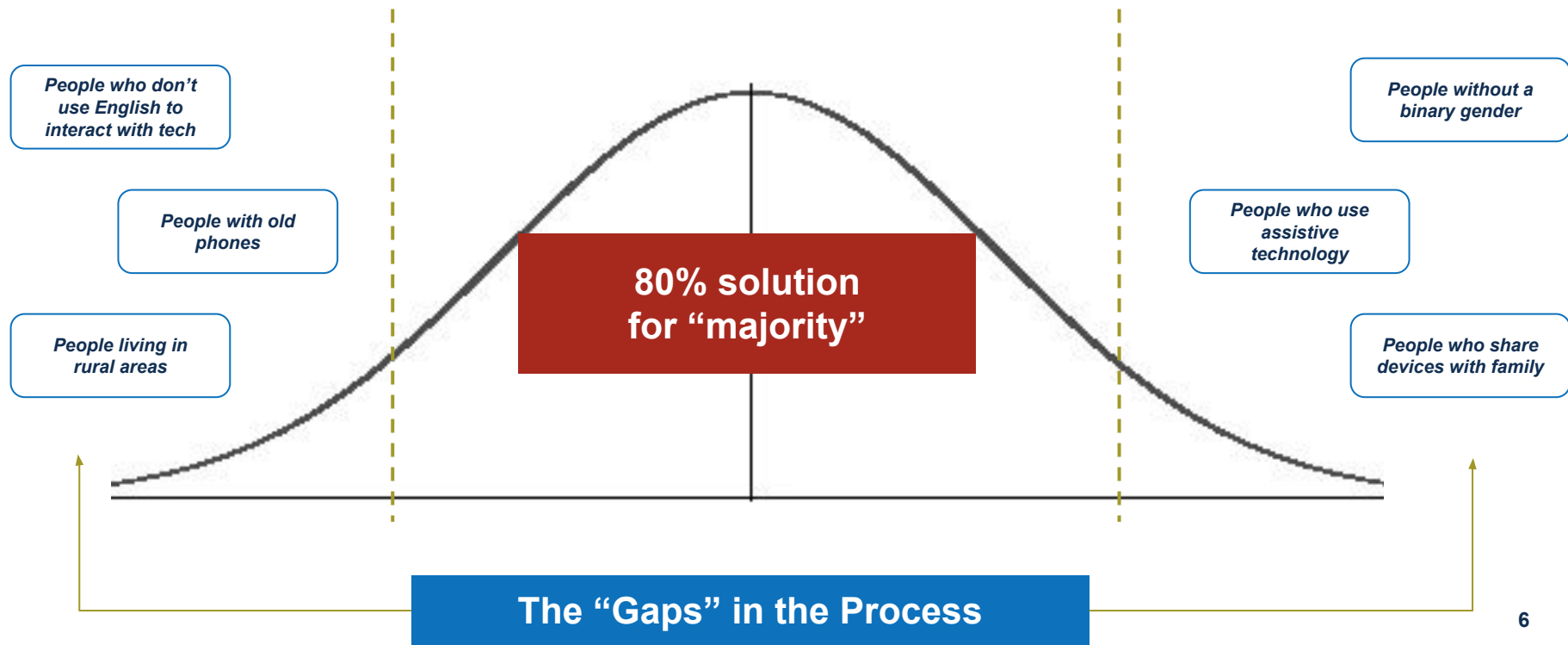
I'm a technologist currently focused on building **responsible data systems** across industry, academia, government, and non-profit domains.

I've worked at places like **Google, MIT, McKinsey**, the **US Digital Service** (White House), and now at the **United Nations**.

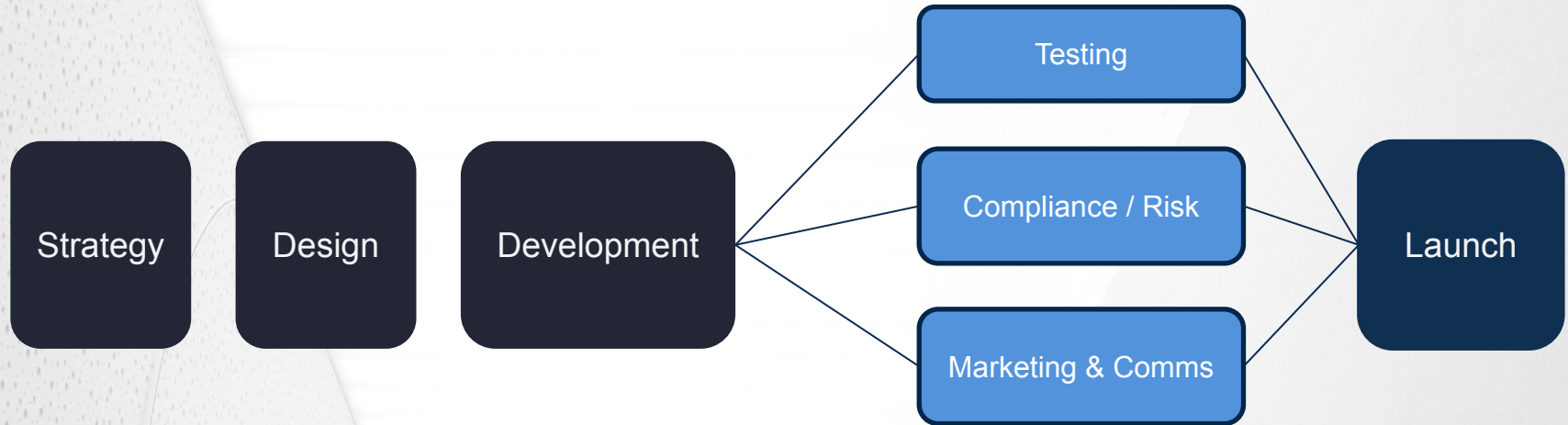
Why are there these  
gaps in the product  
development process?

*It's a direct result of how we  
build product.*

# We optimize for growth. This means building for the “majority” ... often at the expense of others



We also look for issues too late in the process - **reactively**, rather than proactively



... sometimes these checks are even considered “**anti-innovation**”

## Intelligencer

SELECT ALL | AUG. 10, 2018

### ‘Okay Google, Play ‘Dura.’: Voice Assistants Still Can’t Understand Bilingual Users

By Ximena N. Larkin



The result:

Products that don't work for everyone

BUSINESS

The New York Times

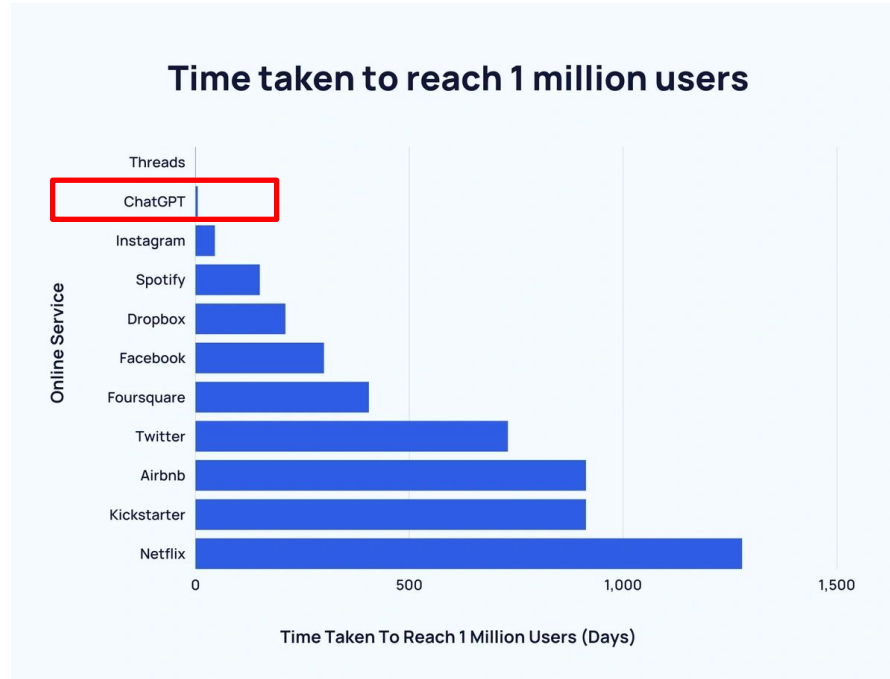
GIVE THE TIMES

### Is an Algorithm Less Racist Than a Loan Officer?

Digital mortgage platforms have the potential to reduce discrimination. But automated systems provide rich opportunities to perpetuate bias, too.



# These AI products are rapidly scaling to massive audiences



Source: [explodingtopics.com](https://explodingtopics.com)

*It falls to people like you and me to fill more product gaps, more quickly*



Image: Firefly

So: what can we do?



**1. We can refuse to build**  
*(and sometimes redirect)*

2017



Image: Firefly

*Problem Statement:*

***Decrease burden on farmers*** who are trying to hire agricultural workers under the H-2A migrant farmworker visa.



U.S. DIGITAL SERVICE

Proposed Solution:

***Build a mobile app*** that makes it easy for a farmer to find a farmworker who can work legally in the country.

Final Solution:

***Streamline the application process*** for farmers to hire workers under the H-2A migrant farmworker visa program by enabling data sharing between agencies.

# Refusing works best when:

- **Leadership is fully supportive** of granting you autonomy
- You are able to **provide another path** to success (redirect)
- There is **internal good will and mutual respect**

*Note: sometimes you may end up having to quit.*





## 2. We can build better



2020

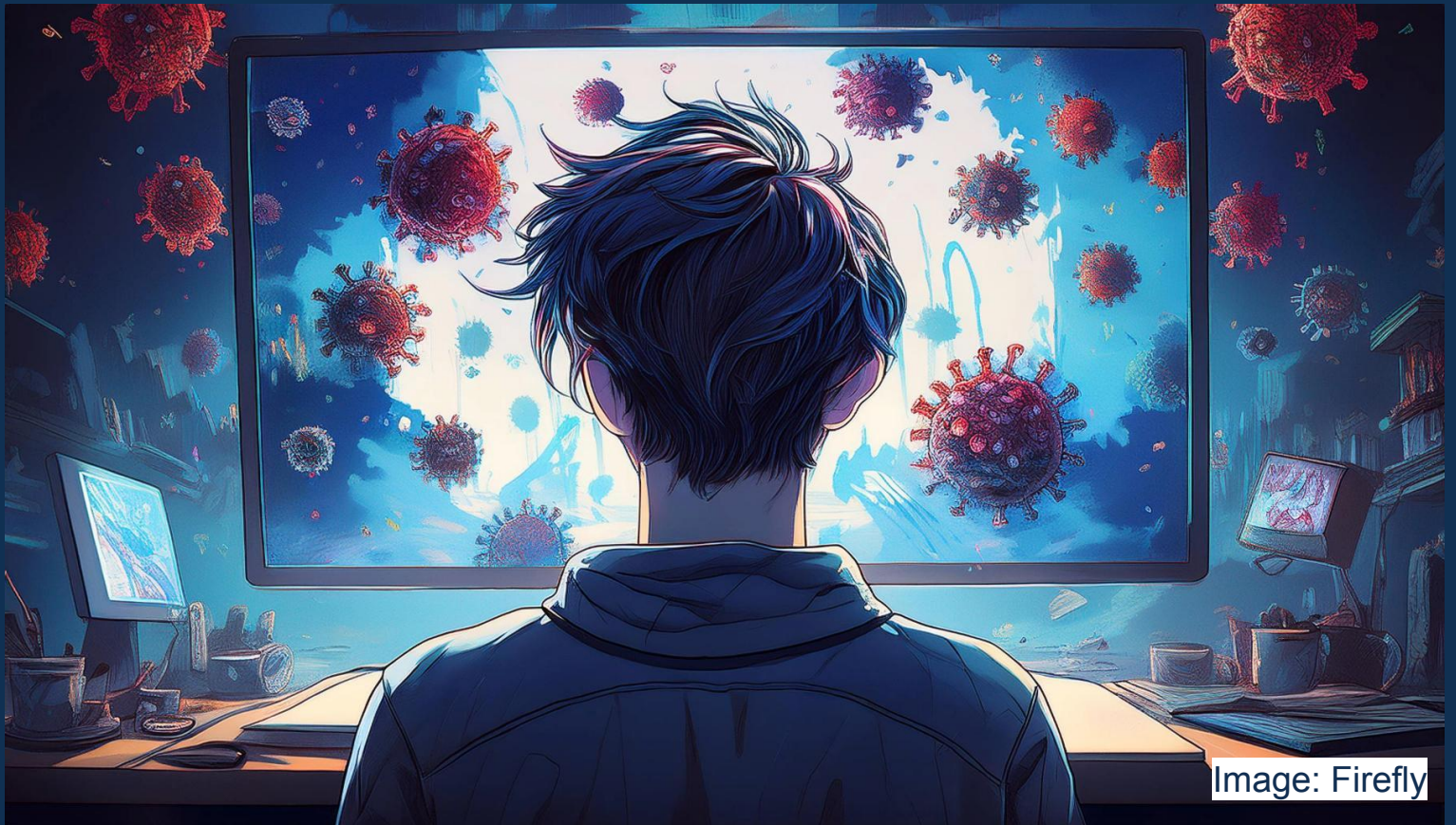
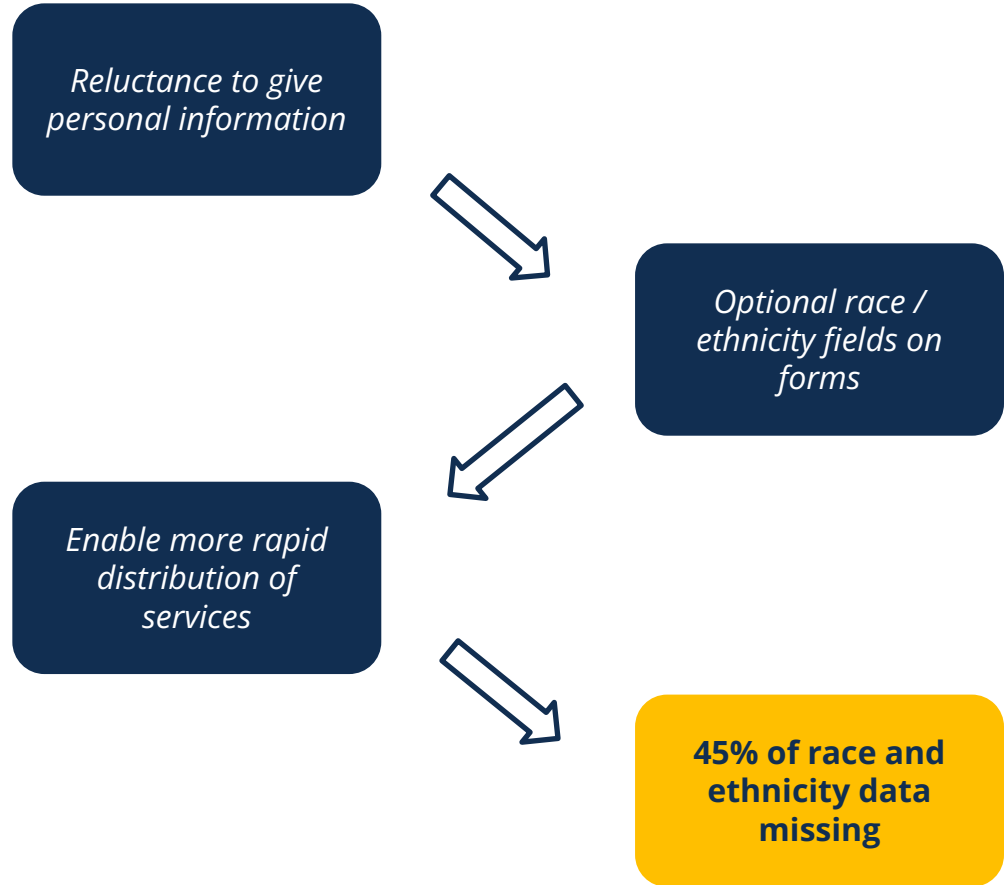


Image: Firefly

*Problem Statement:*

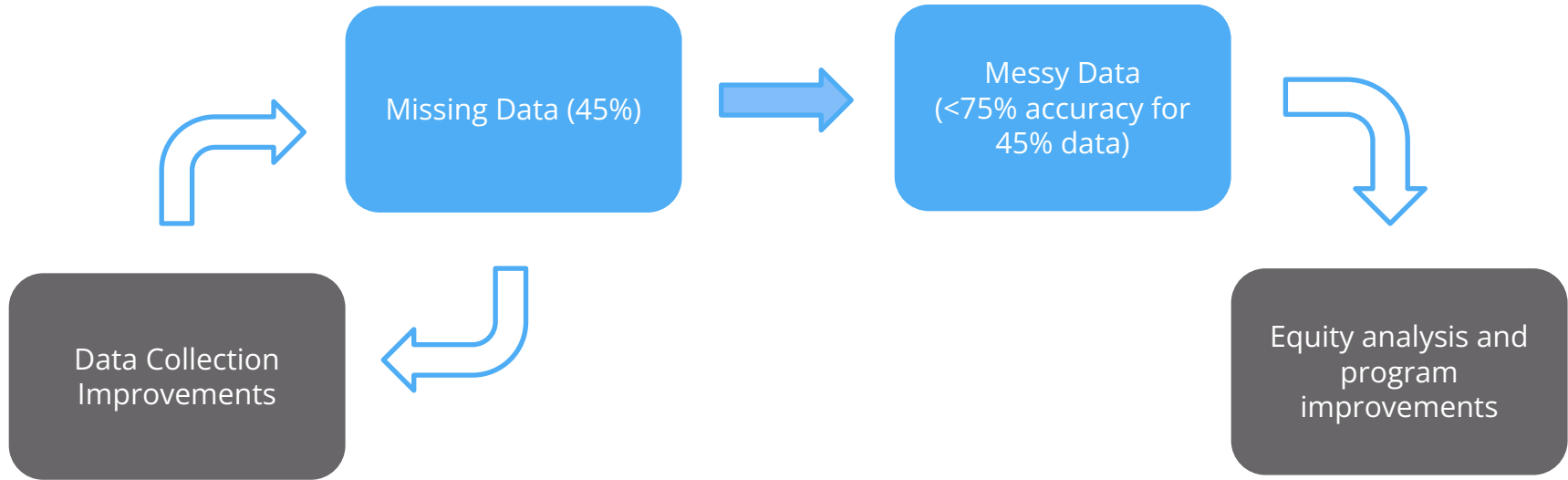
***Identify potential racial equity issues in the ongoing distribution of the Covid-19 vaccine in the United States***

McKinsey  
& Company



# The tradeoff: missing data for messy data

The **Bayesian Improved Surname Geocoding (BISG)** technique predicts race and ethnicity with **<75% accuracy** (varies by race: 43.1% of Black adults are misclassified).



Think about  
**balancing tradeoffs**  
based on which problems  
you're trying to solve

We also need to think  
**holistically about the product**  
when it comes to identifying  
and mitigating issues

# Bias can arise at any point in the development lifecycle

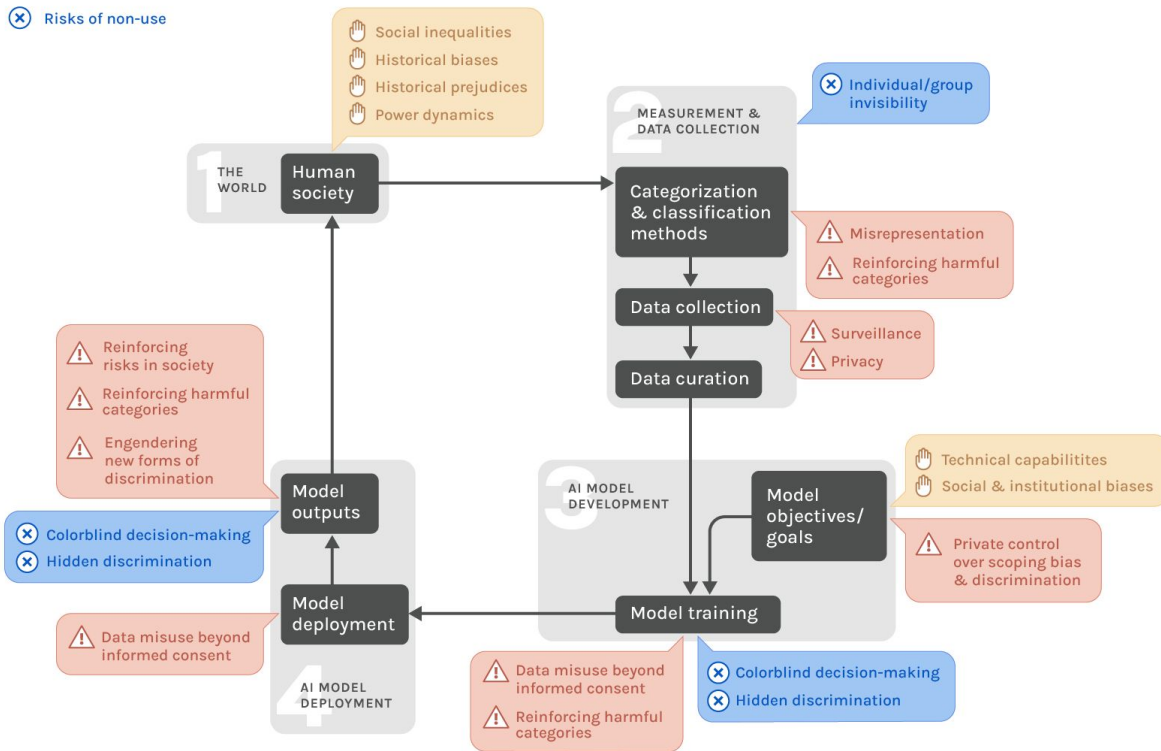
Within the full system, there are many sites of risk for bias and therefore many sites of intervention

## Risks in Algorithmic Decision-Making Systems

👤 Socio-political & contextual influences

⚠️ Risks of use

⊗ Risks of non-use



# When you are building systems...

- Ask whether this is the **right problem to address** with AI
- Ask about the **training data**
- Ask about how the model was **tested**
- Ask about the **success criteria** for the model
- Ask about how the model will be **monitored and updated**
- Ask about criteria for **decommission**





**3. We can  
create new solutions**



# Today

## *'Thousands of Dollars for Something I Didn't Do'*

Because of a bad facial recognition match and other hidden technology, Randal Reid spent nearly a week in jail, falsely accused of stealing purses in a state he said he had never even visited.



## **Suicide Risk Prediction Models Could Perpetuate Racial Disparities**

Two suicide risk prediction models are less accurate for some minority groups, which could exacerbate ethnic and racial disparities.

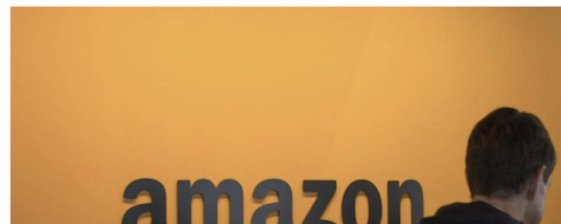


MONEYBOX

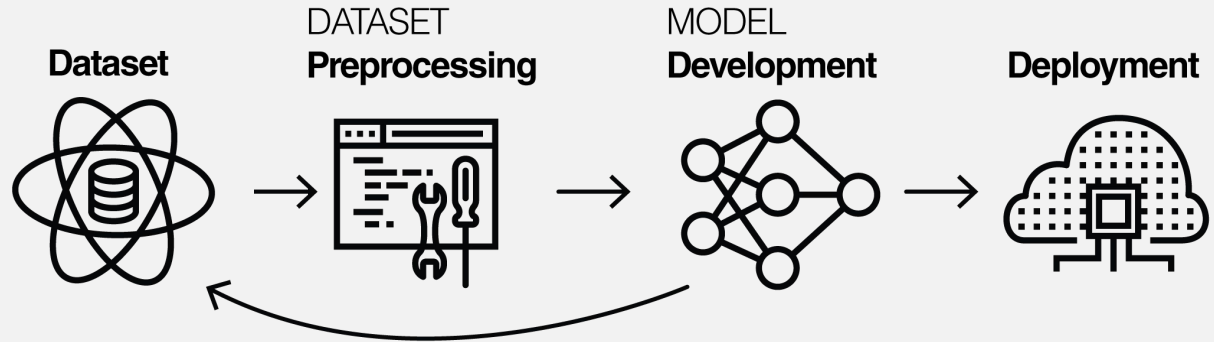
## **Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.**

By JORDAN WEISSMANN

OCT 10, 2018 • 4:52 P



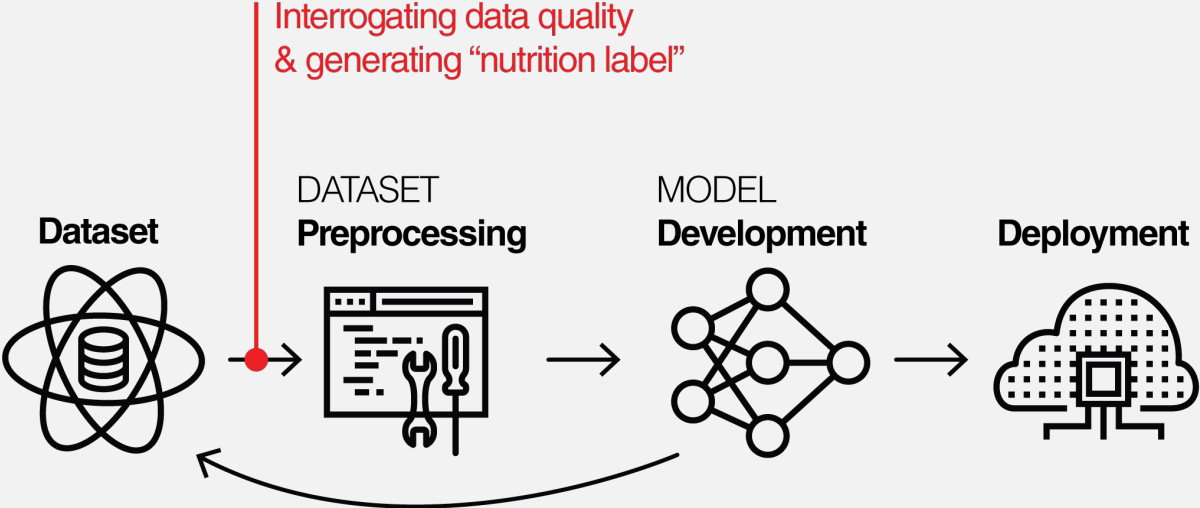
*Systems built on problematic data will exhibit those same issues, especially for historically marginalized people*



## AI Model Pipeline

# Shouldn't we interrogate for issues **before we build?**

*Systems built on problematic data will exhibit those same issues, especially for historically marginalized people*



**AI Model Pipeline**

# The Dataset Nutrition Label

A standardized documentation tool that tells you what's in a dataset and whether it's healthy for your model

**Dataset Fact Sheet**

**Metadata**

**Probabilistic Modeling**

**Missing Units**

**Dataset Nutrition Label**  
**2020 SIIM-ISIC Melanoma Classification Challenge Dataset**

**About**

**Alert Count by Category**

**Alert Count by Potential Harm**

**DATA NUTRITION PROJECT**

**75% COMPLETENESS**

**Studies of Human Cognition with Neural Language Models**






**Description**

**How to use it?**

**Alerts:** Should not be used, Intended use, This is not suitable for, Known uses





*The label includes at-a-glance information about key critical aspects of the dataset as well as usage information and known risks by category.*

### At a Glance ▼

				
About humans	Upstream sources	Technical review	Ethical review	Update frequency
Yes	Yes	No	No	No

### How to use it?

---

 <b>Intended Use</b> <span style="float: right;">▼</span> Domain. Disaster preparedness / social studies... <a href="#">read more</a>	 <b>Restrictions on Use</b> <span style="float: right;">▼</span> no... <a href="#">read more</a>
 <b>Known Uses</b> <span style="float: right;">▼</span> Academic papers, local government usage, disaster modeling... <a href="#">read more</a>	 <b>Do Not Use</b> <span style="float: right;">▼</span> Domain. Identifying industrial sites for projects (public bad) - path of least resistance ... <a href="#">read more</a>

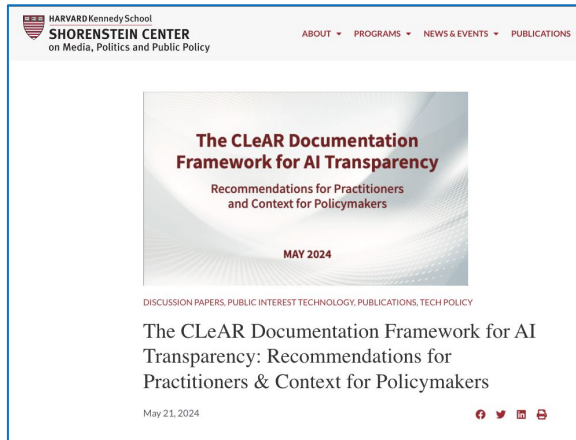
# We have worked with organizations internationally to build the “Label” into their data standard methodology



Microsoft Research



# We also publish, engage and educate around the importance of transparency in AI systems



**CLeAR Documentation Framework Report**  
Harvard Shorenstein Center, 2024

## *Recent discussions with policymakers:*

- *New York City - LOADinG Act*
- *State of California - AB 2013*
- *State of Nebraska - LB 954 and LB 1294*
- *CEN, CENELEC, and European Commission's Joint Research Centre (JRC)*
- *Federal Trade Commission comments*



**Johannes Kepler University (Austria):**  
Workshop - Finding Hidden Bias in Datasets

*“... While I know that the primary mission of the DNP is to improve the understanding, searching, and consumption of datasets by users of datasets, it has also been key to improving my dataset design moving forward.”*

*— Dataset Partner*





There will always be dangerous gaps in technology that **humans need to fill**

But we always have opportunities to **refuse, redirect, build better,**  
or **create new solutions**

**To make technology  
that works for everyone**

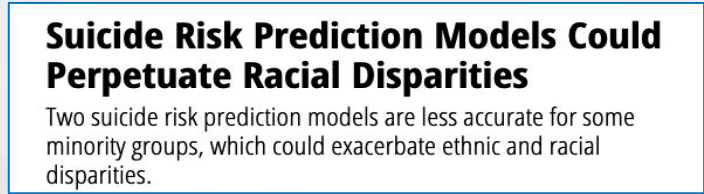
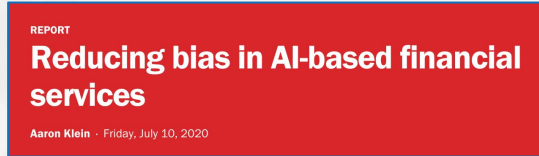


Thank you!

Kasia Chmielinski | October 2024

# Many kinds of harms

- **Discrimination** against specific individuals or groups
- **Distortion** of information
- **Exploitation** of sensitive information without consent
- **Misjudgment** (e.g. make incorrect predictions or classifications)



# Frustratingly, much of the focus is on existential future risks **rather than actual harms**

☰ **CNN BUSINESS** Markets Tech Media Calculators Videos

## AI pioneer quits Google to warn about the technology's 'dangers'

By Jennifer Korn  
Updated 6:15 AM EDT, Wed May 3, 2023

FINANCIAL TIMES

## OpenAI chief says new rules are needed to guard against AI risks

Sam Altman, co-founder of start-up behind ChatGPT, issues warning in first appearance before US Congress

WILL BEDINGFIELD CULTURE 08.05.2023 12:00 PM

## Hollywood's Screenwriters Are Right to Fear AI

The Writers Guild of America's demands for guardrails on artificial intelligence are a smart move—and the stakes are higher than ever.

## *Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'*

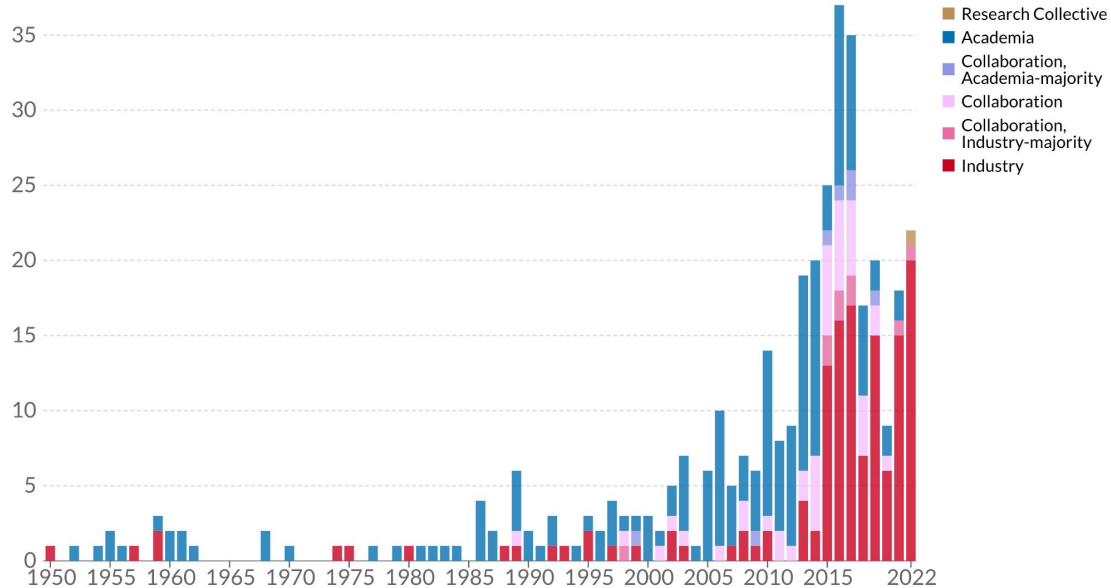
More than 1,000 tech leaders, researchers and others signed an open letter urging a moratorium on the development of the most powerful artificial intelligence systems.

# Yes, and... we need to focus on the **consolidation of power and control** in the hands of the few

## Affiliation of research teams building notable AI systems

Our World  
in Data

Systems are defined as "notable" by the authors based on several criteria, such as advancing the state of the art or being of historical importance.



Source: Sevilla et al. (2023)

OurWorldInData.org/artificial-intelligence • CC BY

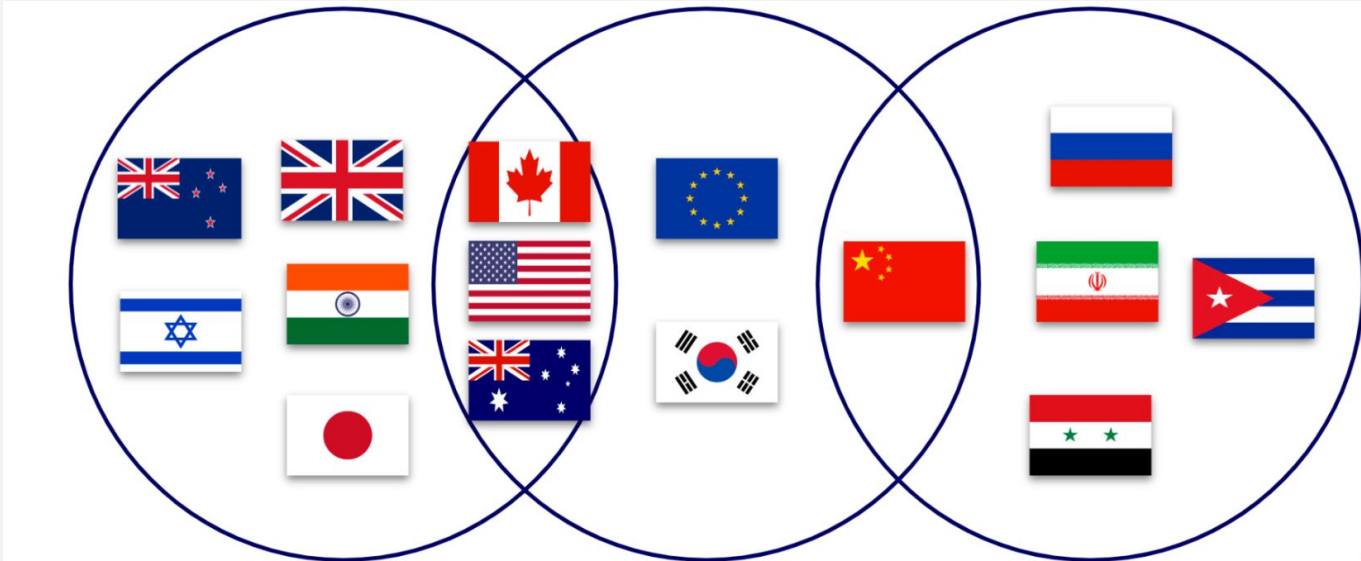
Note: A research collective is a group of AI researchers who are not organized under an academic or industry affiliation; e.g., [EleutherAI](#).

# And although regulation is coming, it is **regionalized** and **reactive**

Relying on **existing** laws  
and regulations

Introducing AI-specific  
**legislative frameworks**

**Banning specific services**  
(e.g. ChatGPT)



**ARTIFICIAL INTELLIGENCE**



**WHAT'S THERE TO WORRY ABOUT?!**



***"Hey, the rich guy is doing sketchy stuff to get even richer!"***



**SEE, NOBODY CARES**



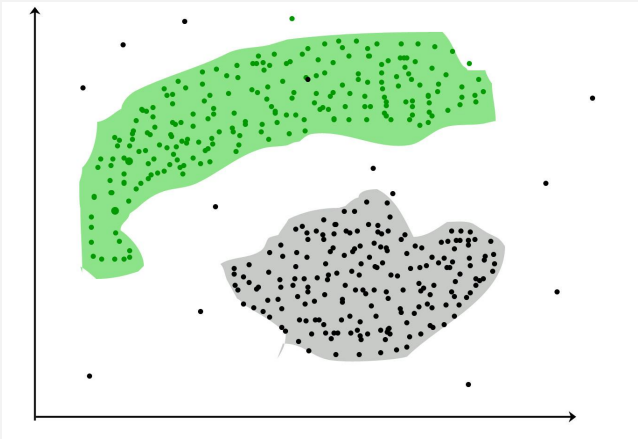
How do we build technology  
that works better and  
for more people?

2017



Image:  
Firefly

It is not possible to remove all biases in the system,



Profile 1 =  
Approved

bi·as

/ˈbiːəs/

prejudice in **favor of** or **against one thing**,  
person, or group compared with another ...

... the goal is to understand and control them

# Even our definitions of **fairness** cannot be satisfied simultaneously!

METRIC	DESCRIPTION	Example: Medical Testing
<b>Demographic parity</b>	All groups have <b>equal probability</b> of being assigned by the model to the positive class	Medical tests should show positive rates equally for people of all sexes
<b>Predictive equality</b>	<b>Equal false positive rates</b> between groups	Medical tests should not over-predict positives
<b>Equal opportunity</b>	<b>Equal false negative rates</b> between groups	Medical tests should not over-predict negatives