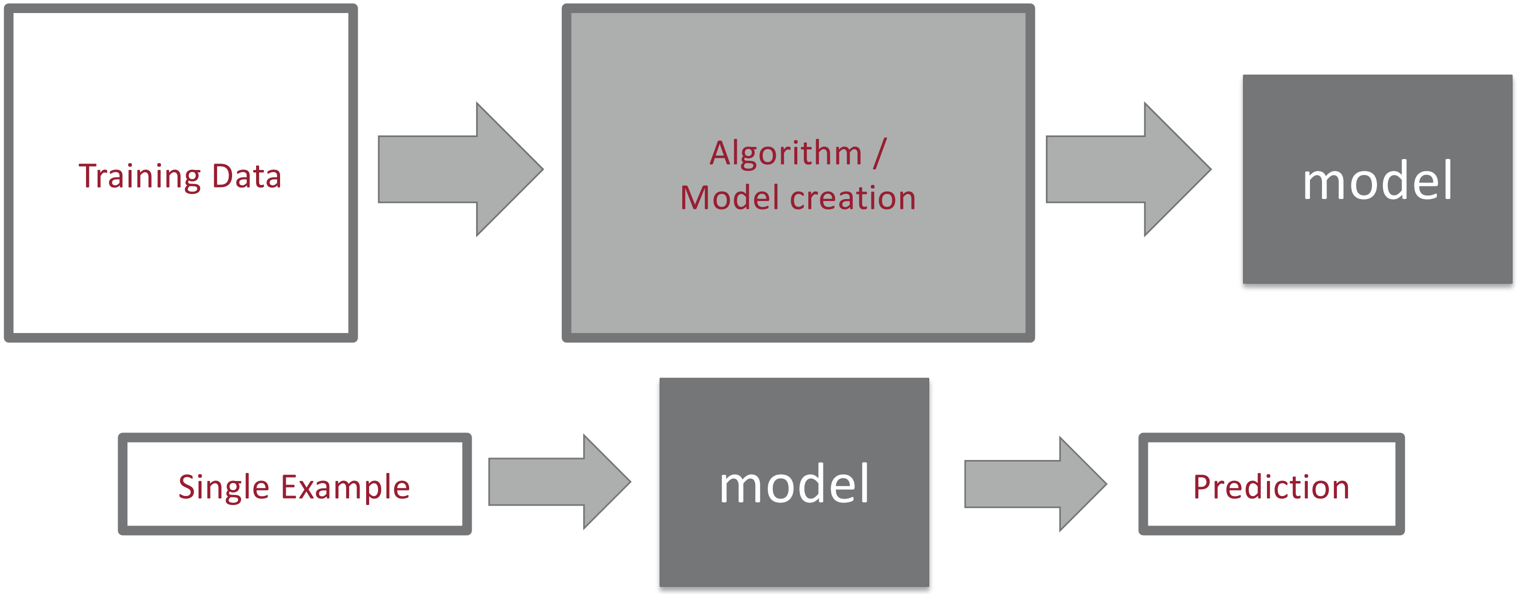
CS 360 Machine Learning

Sources of Error in an ML Pipeline and Governance

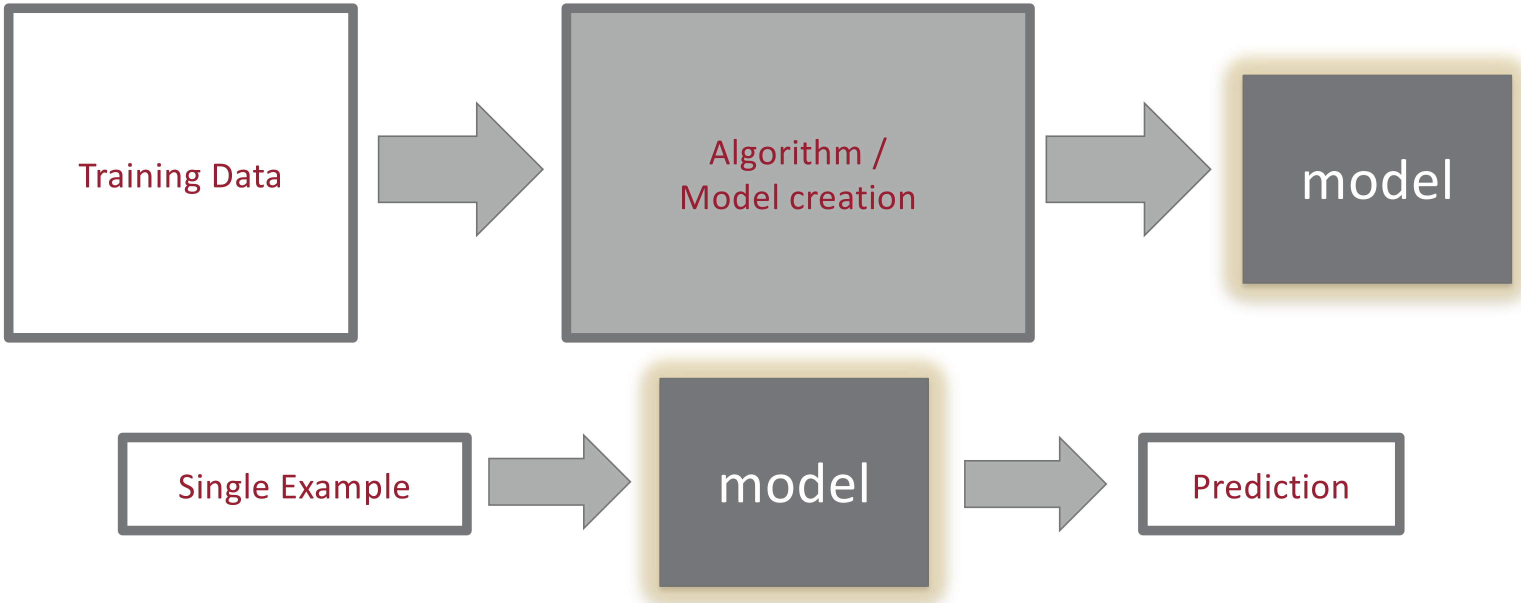
HAVERFORD
COLLEGE

DEPARTMENT OF COMPUTER SCIENCE

Machine Learning Pipeline

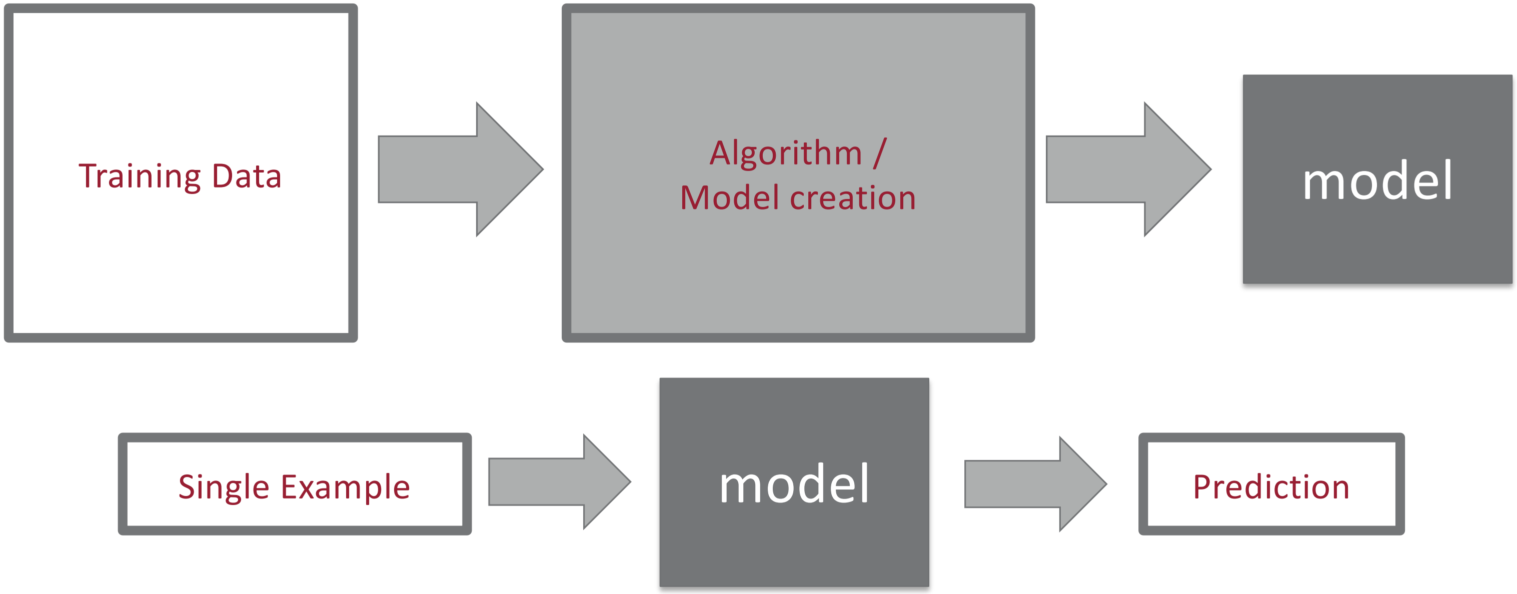


Error Measures



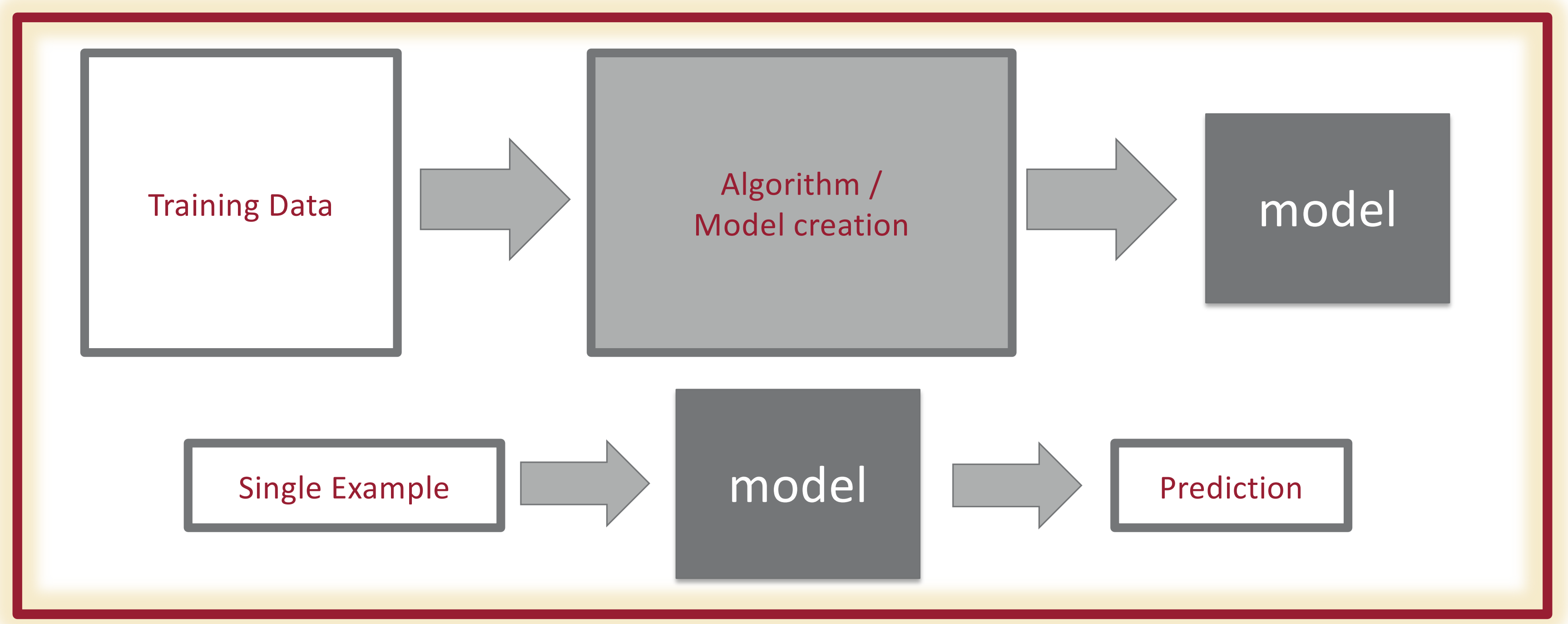
Accuracy and other traditional error measures focus on evaluating the model against the test data.

Sources of Error



In a real world problem, you've made assumptions throughout this pipeline – what if they're wrong?

Sources of Error



Training Data

Algorithm / Model creation

model

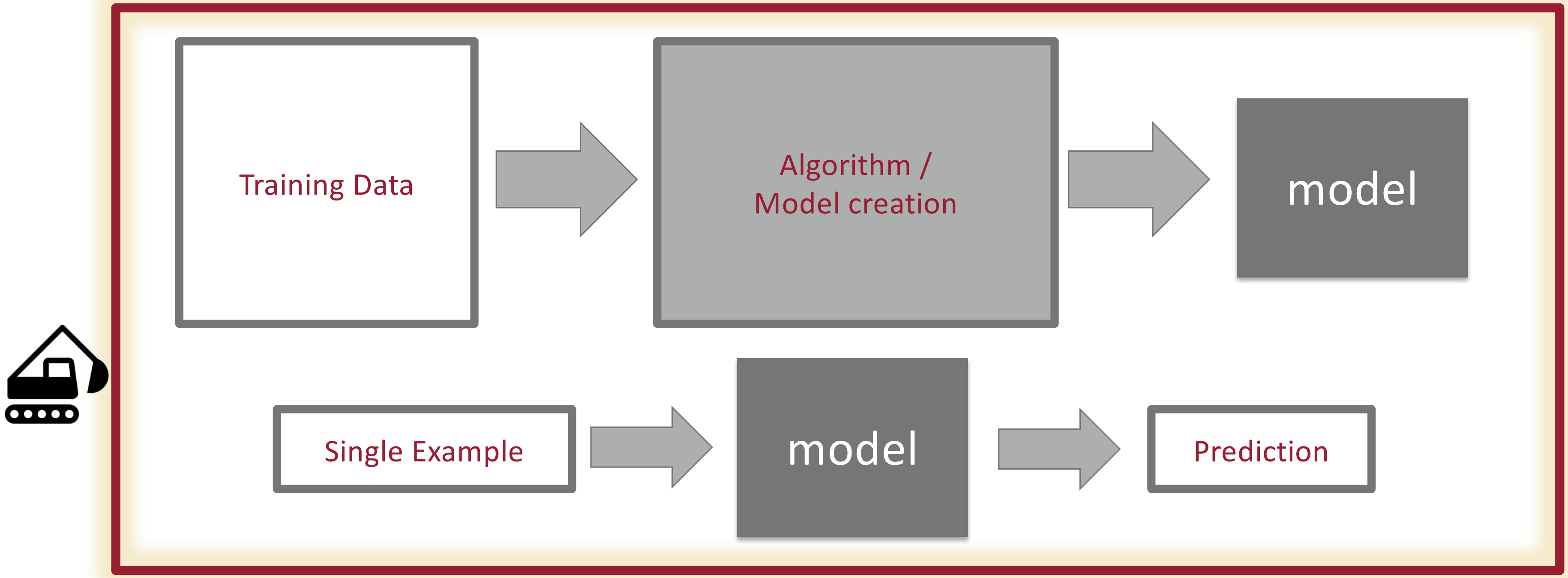
Single Example

model

Prediction

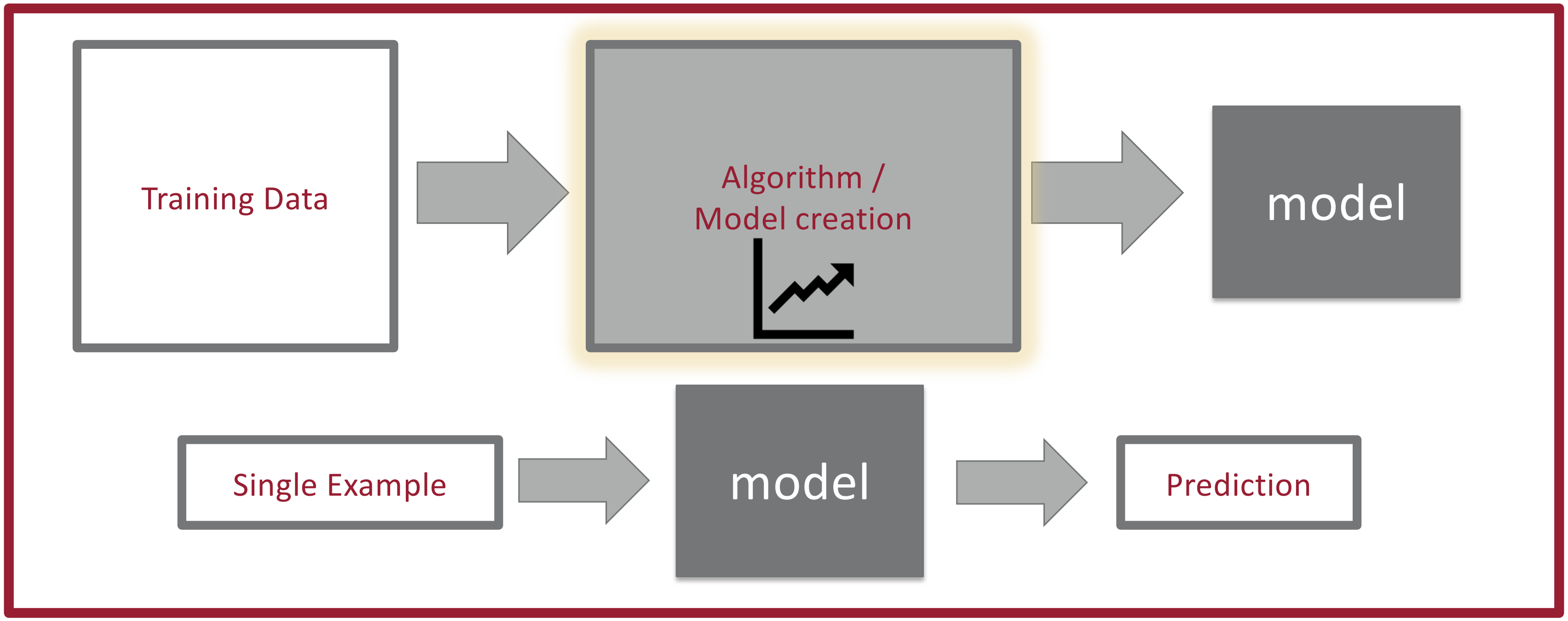
Assumption 0: the problem is appropriate to solve with ML

Sources of Error



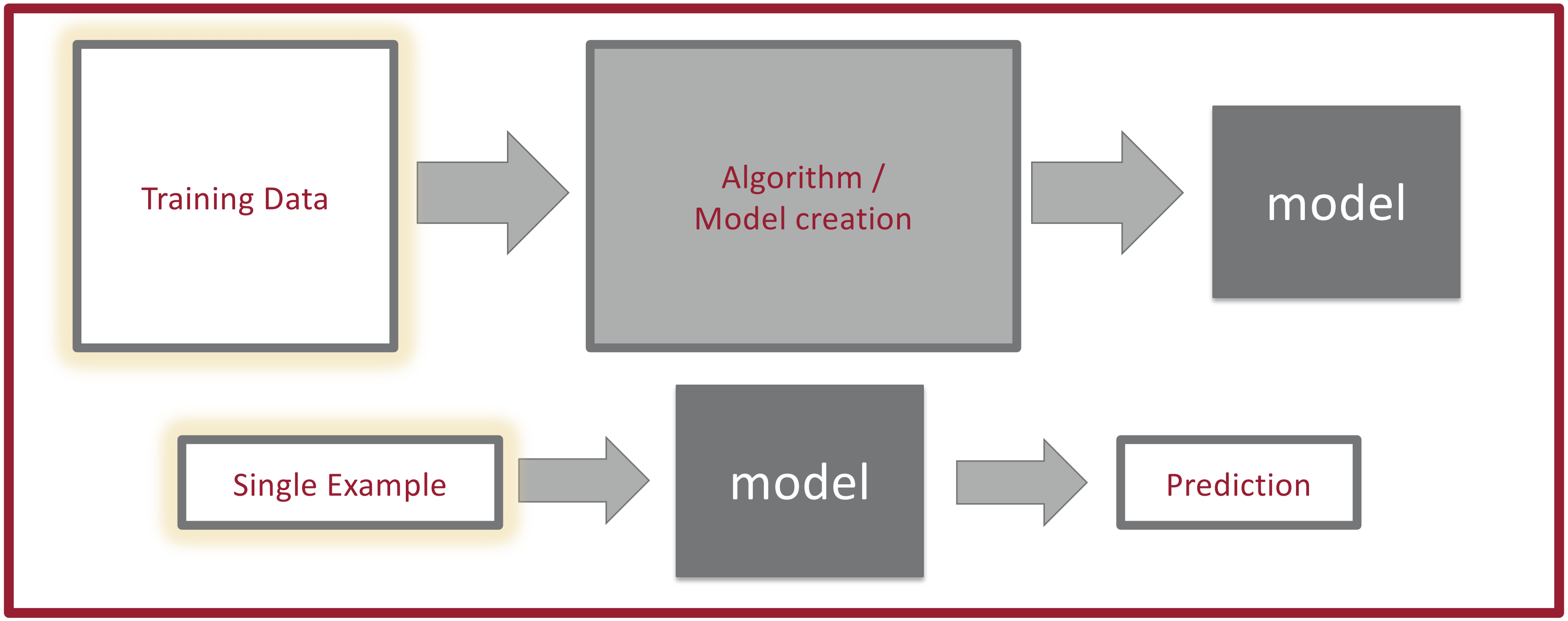
Assumption 1: the real world won't change or impact the ML pipeline

Sources of Error



Assumption 2: the chosen ML algorithm is appropriate to the real world context – does your model match the underlying phenomena and real-world societal understandings?

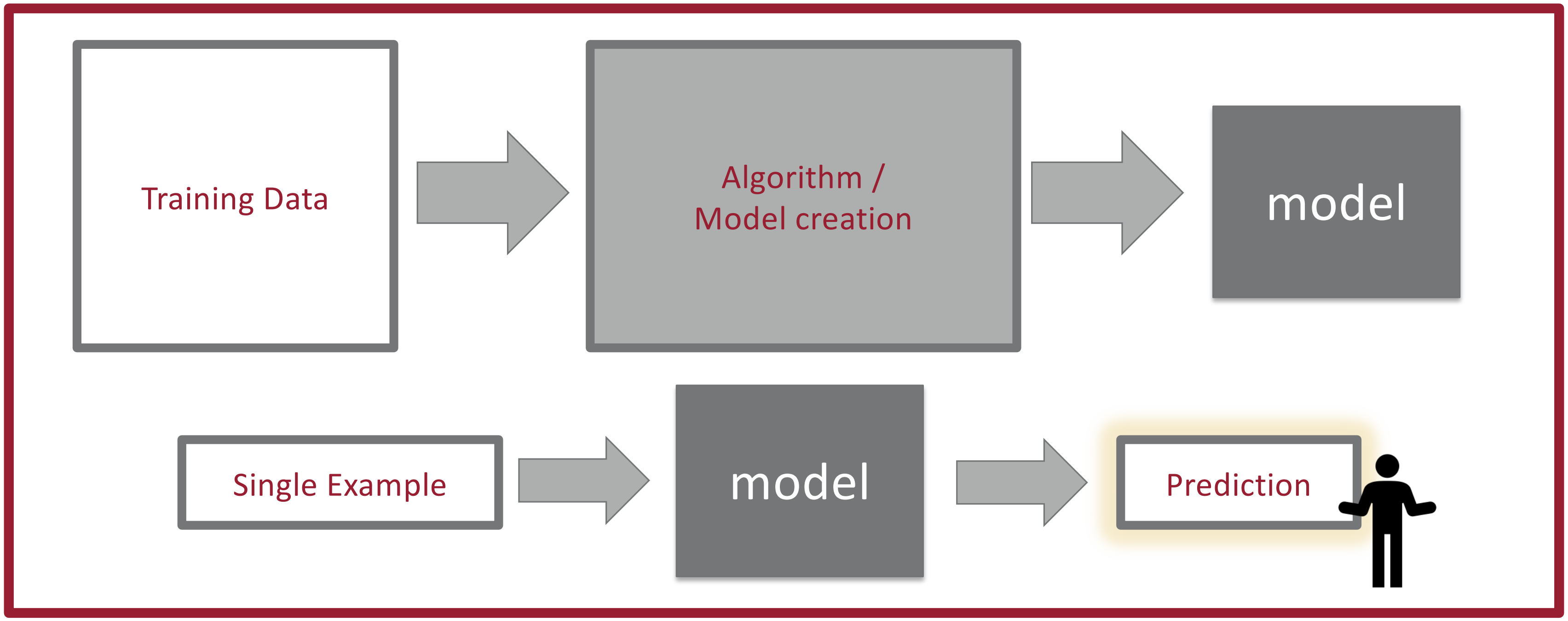
Sources of Error



Assumption 3: the developed pipeline and/or model can be applied in a new context

- Assumptions about the training data and/or example distributions may not hold!

Sources of Error



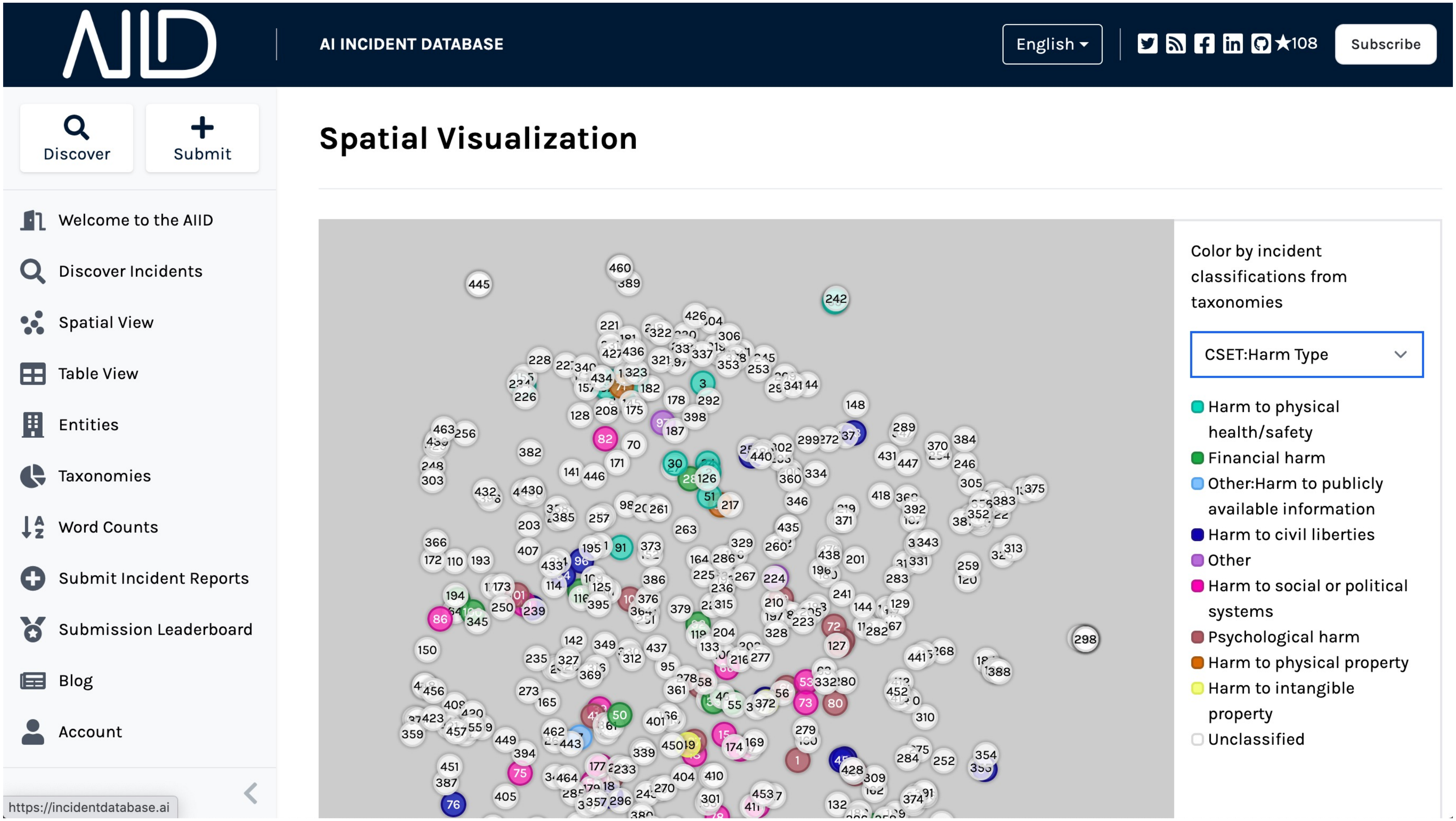
Assumption 4: the resulting prediction will be applied correctly and in the appropriate context – what real-world considerations might you have forgotten?

Scenarios

1. What real-world harm occurred?

2. Why did that happen – what was the technical or sociotechnical error?

3. What could have been done to prevent it?

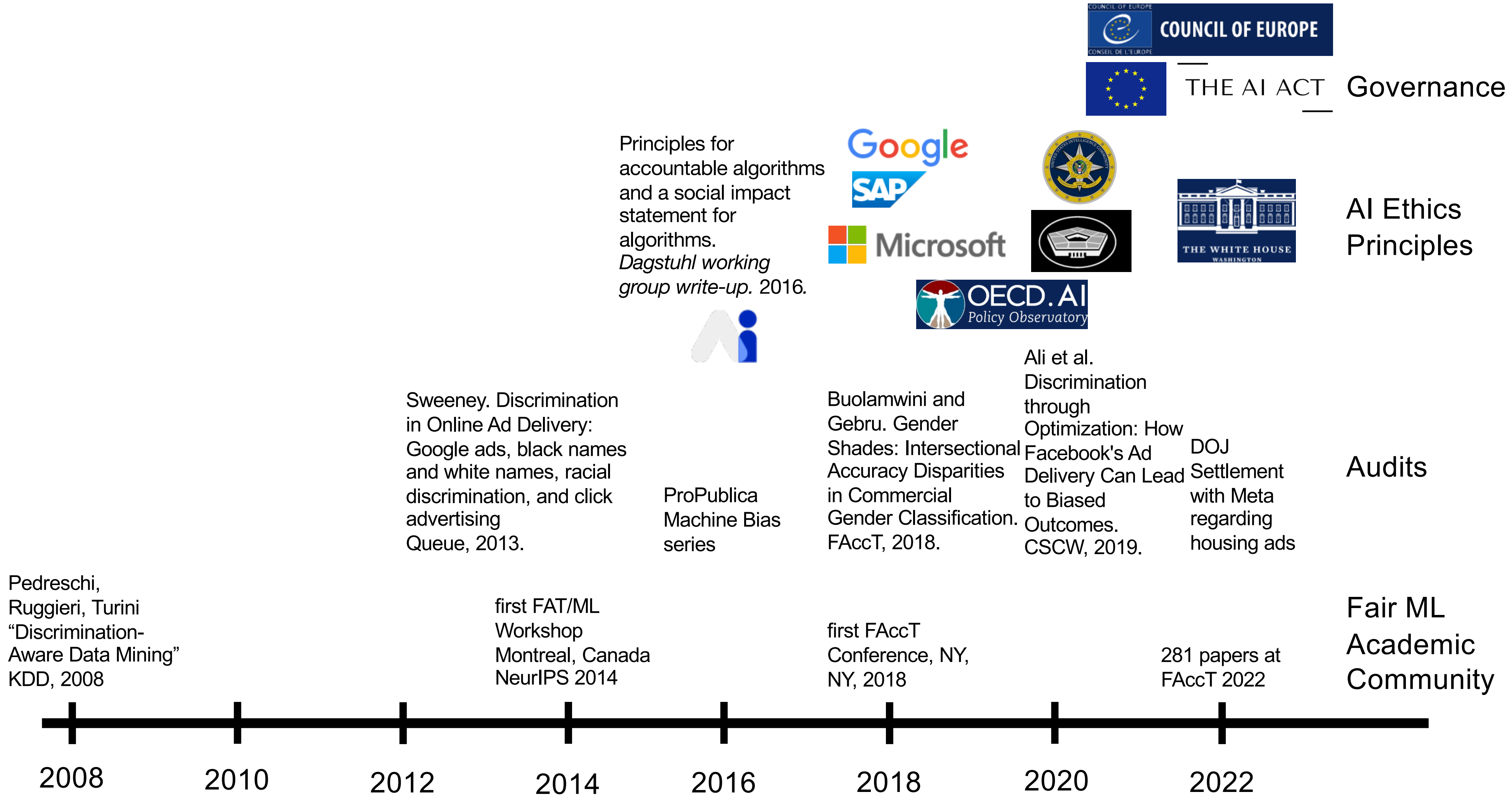


Spatial Visualization

Color by incident classifications from taxonomies

CSET:Harm Type

- ■ Harm to physical health/safety
- ■ Financial harm
- ■ Other:Harm to publicly available information
- ■ Harm to civil liberties
- ■ Other
- ■ Harm to social or political systems
- ■ Psychological harm
- ■ Harm to physical property
- ■ Harm to intangible property
- □ Unclassified



Governance

AI Ethics Principles

Principles for accountable algorithms and a social impact statement for algorithms. *Dagstuhl working group write-up.* 2016.

Audits

Sweeney. Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising Queue, 2013.

ProPublica Machine Bias series

Buolamwini and Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAccT, 2018.

Ali et al. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. CSCW, 2019.

DOJ Settlement with Meta regarding housing ads

Fair ML Academic Community

Pedreschi, Ruggieri, Turini "Discrimination-Aware Data Mining" KDD, 2008

first FAT/ML Workshop Montreal, Canada NeurIPS 2014

first FAccT Conference, NY, NY, 2018

281 papers at FAccT 2022

An Incomplete History of Responsible AI

Blueprint for an AI Bill of Rights

THE WHITE HOUSE

Safe and Effective Systems

You should be protected from unsafe or ineffective systems.

Algorithmic Discrimination Protections

You should not face discrimination by algorithms and systems should be used and designed in an equitable way.

Data Privacy

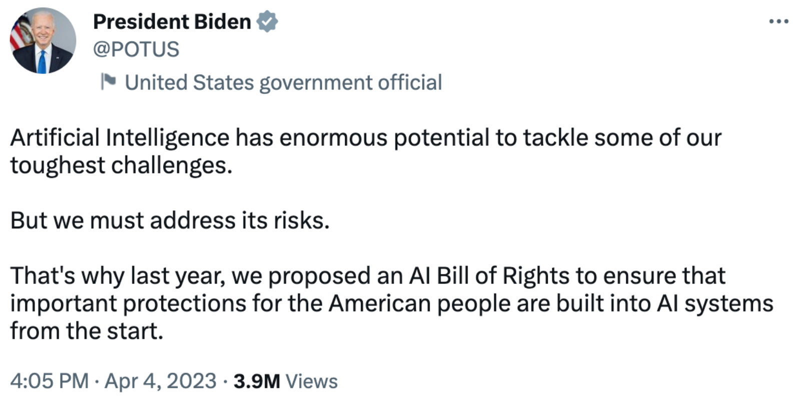
You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used.

Notice and Explanation

You should know when an automated system is being used and understand how and why it contributes to outcomes that impact you.

Human Alternatives, Consideration, and Fallback

You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.



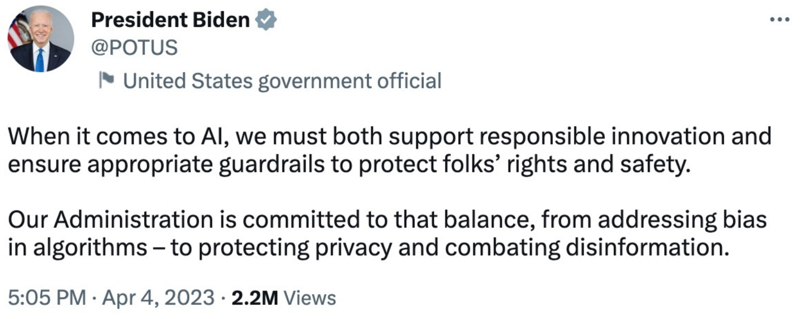
President Biden
@POTUS
United States government official

Artificial Intelligence has enormous potential to tackle some of our toughest challenges.

But we must address its risks.

That's why last year, we proposed an AI Bill of Rights to ensure that important protections for the American people are built into AI systems from the start.

4:05 PM · Apr 4, 2023 · **3.9M** Views



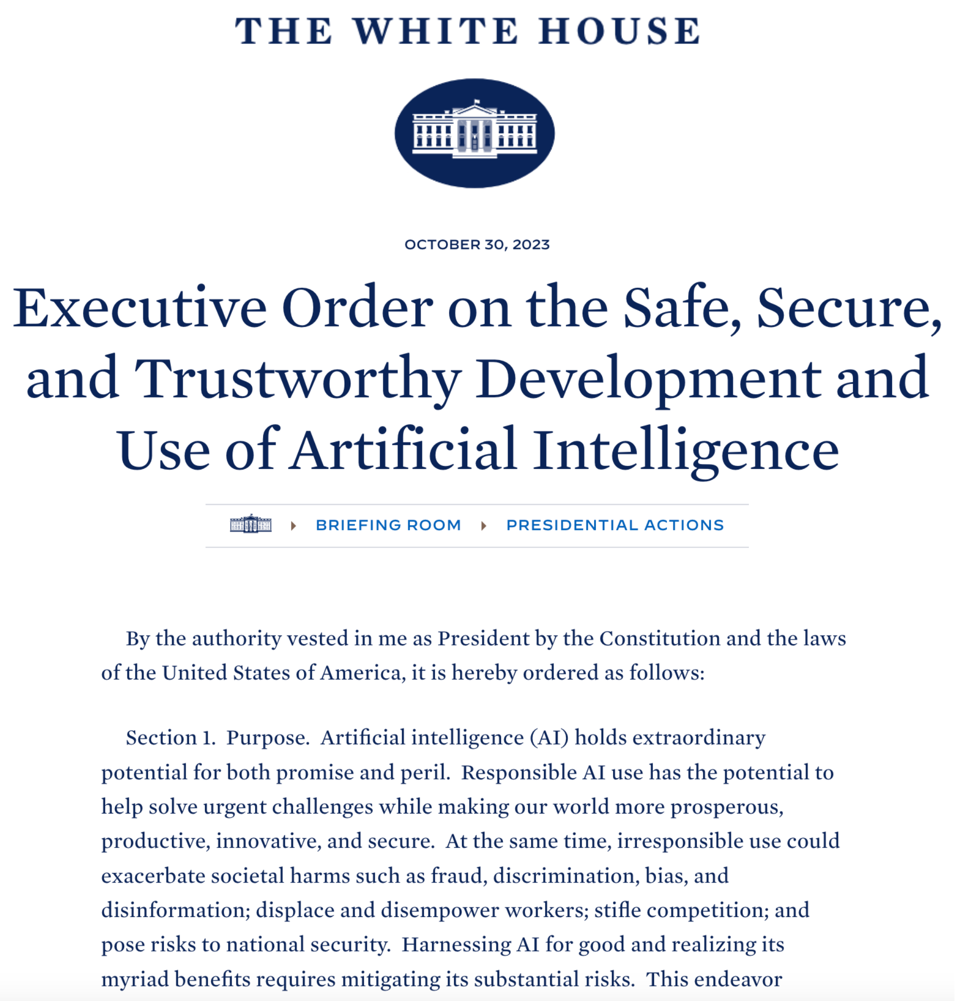
President Biden
@POTUS
United States government official

When it comes to AI, we must both support responsible innovation and ensure appropriate guardrails to protect folks' rights and safety.

Our Administration is committed to that balance, from addressing bias in algorithms – to protecting privacy and combating disinformation.

5:05 PM · Apr 4, 2023 · **2.2M** Views

http://www.whitehouse.gov/ostp/ai-bill-of-rights



THE WHITE HOUSE

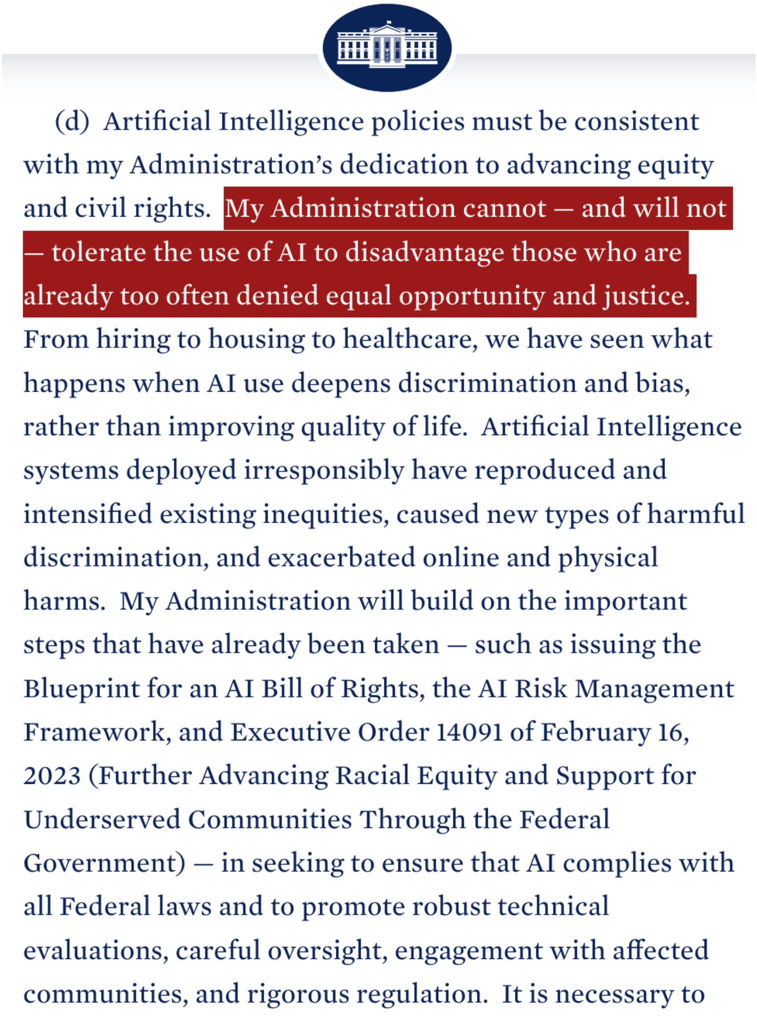
OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

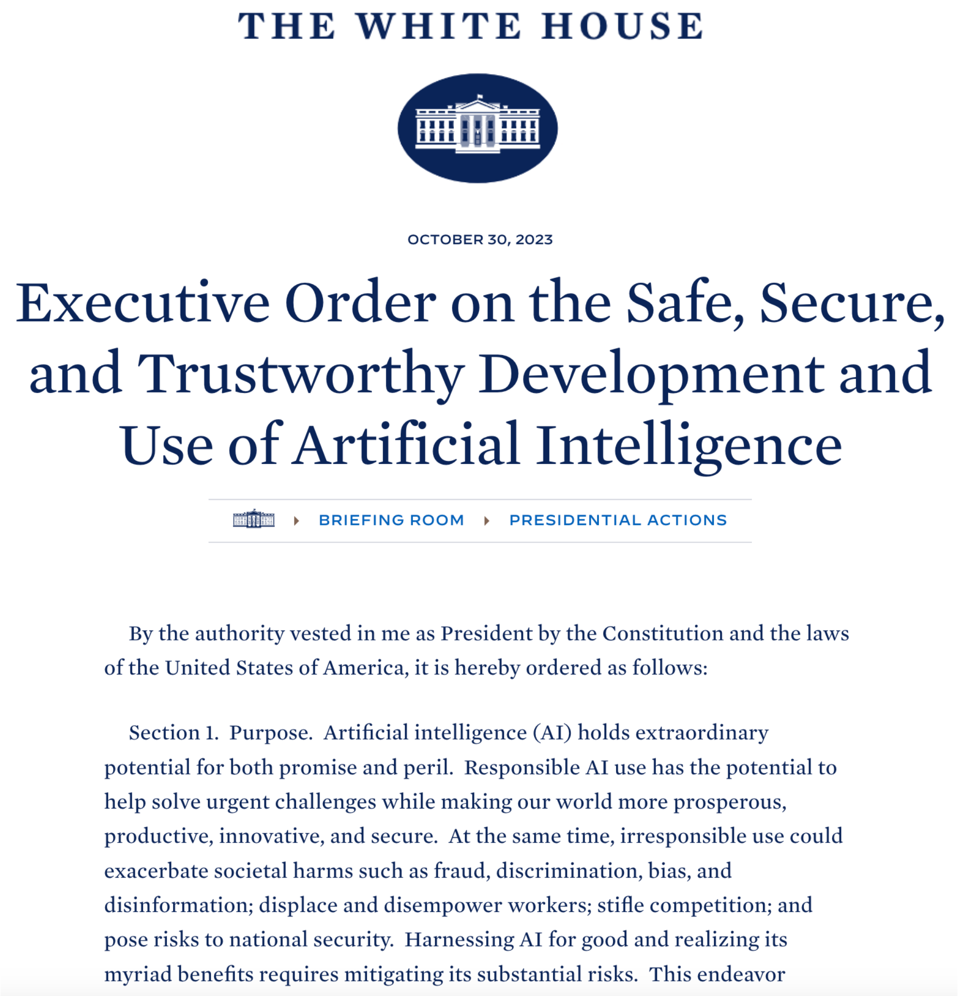
By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/



(d) Artificial Intelligence policies must be consistent with my Administration's dedication to advancing equity and civil rights. My Administration cannot — and will not — tolerate the use of AI to disadvantage those who are already too often denied equal opportunity and justice. From hiring to housing to healthcare, we have seen what happens when AI use deepens discrimination and bias, rather than improving quality of life. Artificial Intelligence systems deployed irresponsibly have reproduced and intensified existing inequities, caused new types of harmful discrimination, and exacerbated online and physical harms. My Administration will build on the important steps that have already been taken — such as issuing the Blueprint for an AI Bill of Rights, the AI Risk Management Framework, and Executive Order 14091 of February 16, 2023 (Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government) — in seeking to ensure that AI complies with all Federal laws and to promote robust technical evaluations, careful oversight, engagement with affected communities, and rigorous regulation. It is necessary to



THE WHITE HOUSE

OCTOBER 30, 2023

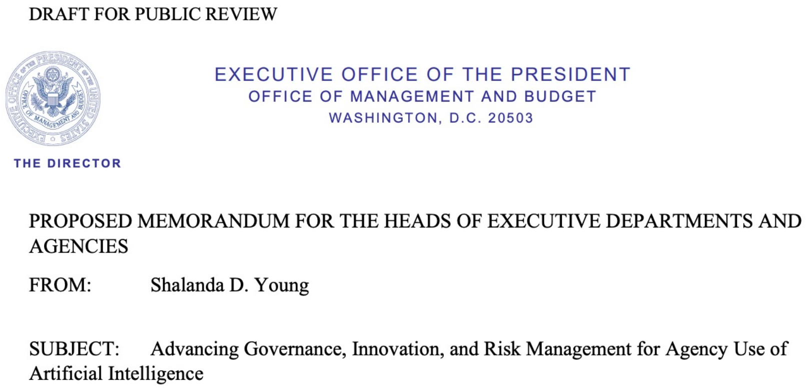
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

AI.GOV



DRAFT FOR PUBLIC REVIEW

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

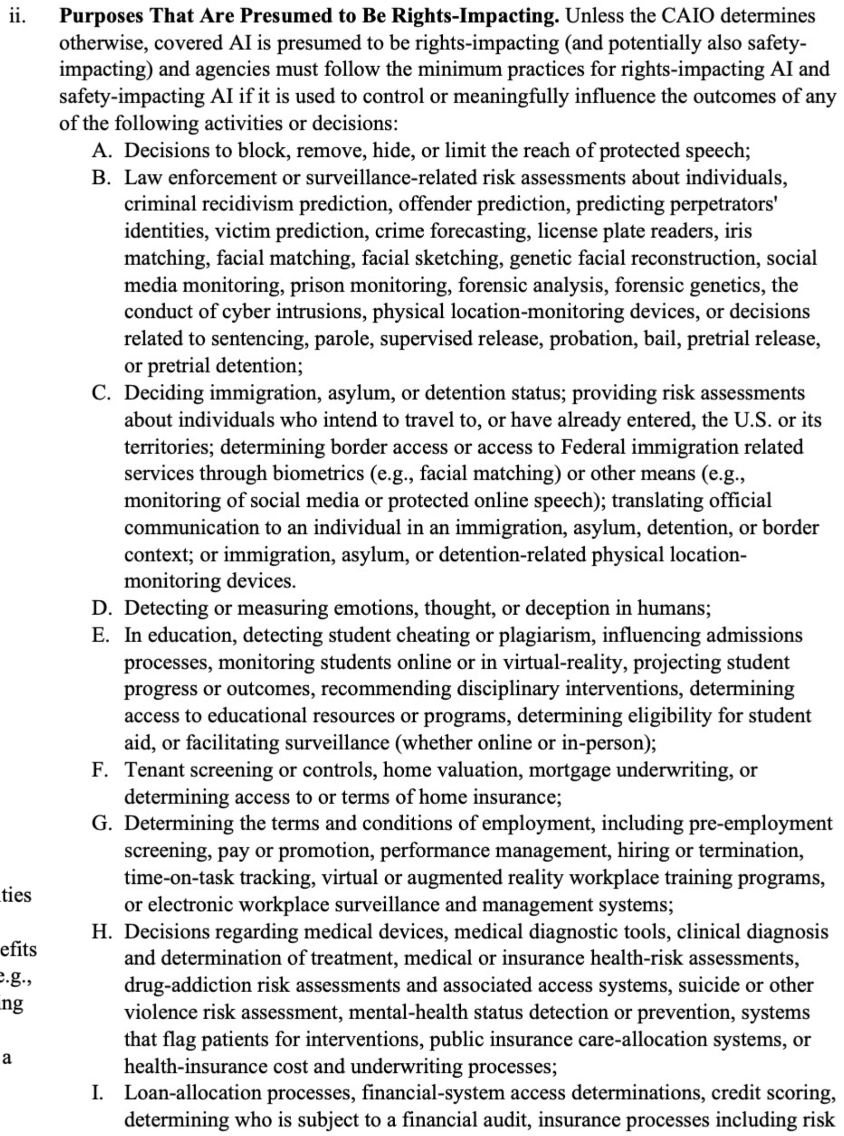
PROPOSED MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Shalanda D. Young

SUBJECT: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence

b. Determining Which Artificial Intelligence Is Presumed to Be Safety-Impacting or Rights-Impacting

All AI within the scope of this section that matches the definitions of "safety-impacting AI" or "rights-impacting AI" as defined in Section 6 must follow the minimum practices in Section 5(c) by the appropriate deadline. Agencies must review each use of AI that they are developing or using to determine whether it matches the definition of safety-impacting or rights-impacting.



ii. **Purposes That Are Presumed to Be Rights-Impacting.** Unless the CAIO determines otherwise, covered AI is presumed to be rights-impacting (and potentially also safety-impacting) and agencies must follow the minimum practices for rights-impacting AI and safety-impacting AI if it is used to control or meaningfully influence the outcomes of any of the following activities or decisions:

A. Decisions to block, remove, hide, or limit the reach of protected speech;

B. Law enforcement or surveillance-related risk assessments about individuals, criminal recidivism prediction, offender prediction, predicting perpetrators' identities, victim prediction, crime forecasting, license plate readers, iris matching, facial matching, facial sketching, genetic facial reconstruction, social media monitoring, prison monitoring, forensic analysis, forensic genetics, the conduct of cyber intrusions, physical location-monitoring devices, or decisions related to sentencing, parole, supervised release, probation, bail, pretrial release, or pretrial detention;

C. Deciding immigration, asylum, or detention status; providing risk assessments about individuals who intend to travel to, or have already entered, the U.S. or its territories; determining border access or access to Federal immigration related services through biometrics (e.g., facial matching) or other means (e.g., monitoring of social media or protected online speech); translating official communication to an individual in an immigration, asylum, detention, or border context; or immigration, asylum, or detention-related physical location-monitoring devices.

D. Detecting or measuring emotions, thought, or deception in humans;

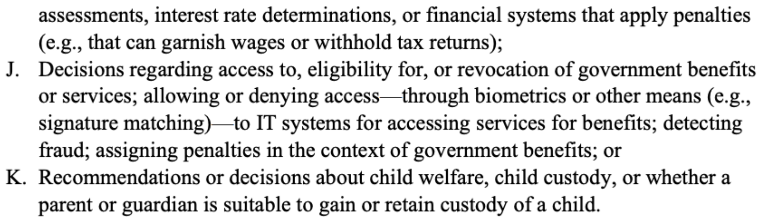
E. In education, detecting student cheating or plagiarism, influencing admissions processes, monitoring students online or in virtual-reality, projecting student progress or outcomes, recommending disciplinary interventions, determining access to educational resources or programs, determining eligibility for student aid, or facilitating surveillance (whether online or in-person);

F. Tenant screening or controls, home valuation, mortgage underwriting, or determining access to or terms of home insurance;

G. Determining the terms and conditions of employment, including pre-employment screening, pay or promotion, performance management, hiring or termination, time-on-task tracking, virtual or augmented reality workplace training programs, or electronic workplace surveillance and management systems;

H. Decisions regarding medical devices, medical diagnostic tools, clinical diagnosis and determination of treatment, medical or insurance health-risk assessments, drug-addiction risk assessments and associated access systems, suicide or other violence risk assessment, mental-health status detection or prevention, systems that flag patients for interventions, public insurance care-allocation systems, or health-insurance cost and underwriting processes;

I. Loan-allocation processes, financial-system access determinations, credit scoring, determining who is subject to a financial audit, insurance processes including risk

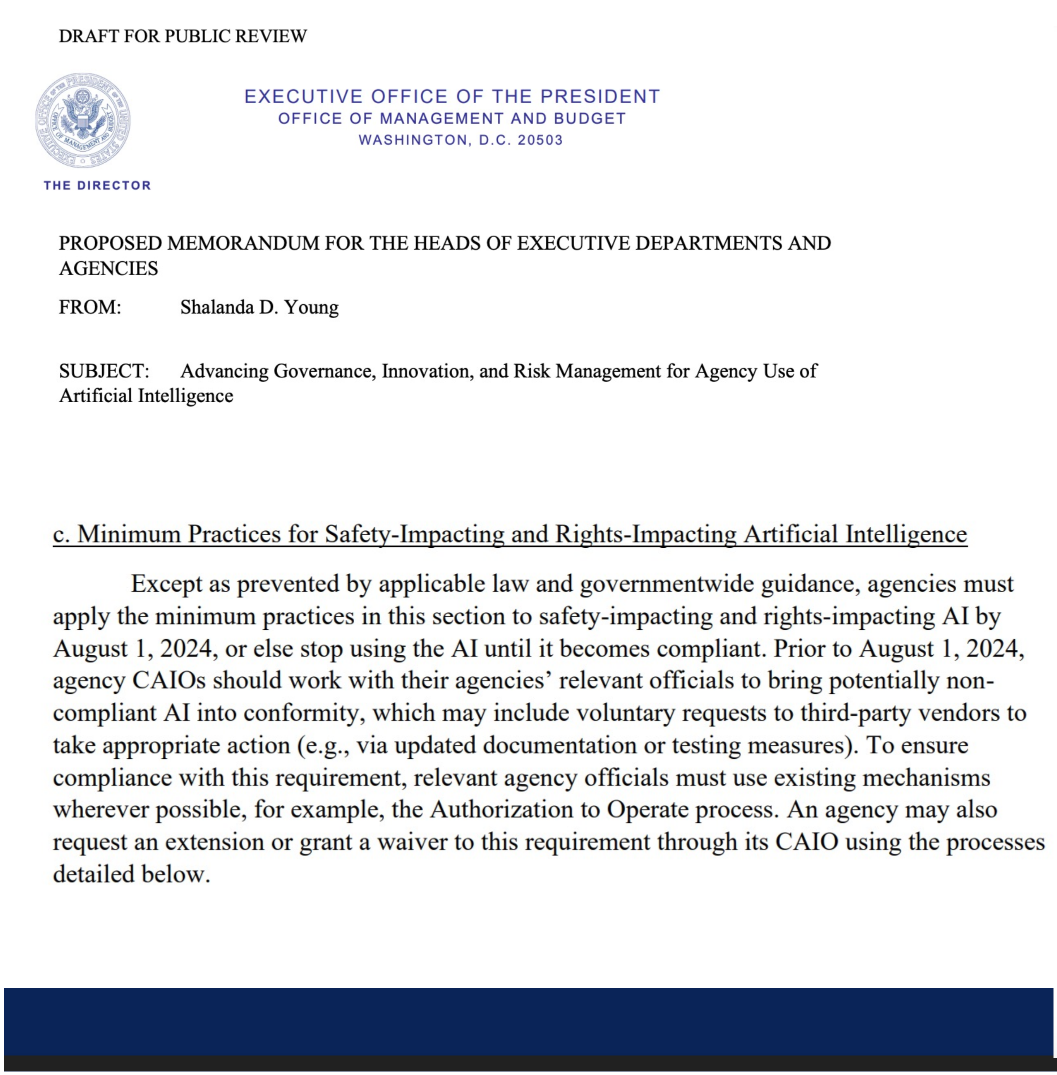


assessments, interest rate determinations, or financial systems that apply penalties (e.g., that can garnish wages or withhold tax returns);

J. Decisions regarding access to, eligibility for, or revocation of government benefits or services; allowing or denying access—through biometrics or other means (e.g., signature matching)—to IT systems for accessing services for benefits; detecting fraud; assigning penalties in the context of government benefits; or

K. Recommendations or decisions about child welfare, child custody, or whether a parent or guardian is suitable to gain or retain custody of a child.

AI.GOV



DRAFT FOR PUBLIC REVIEW

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

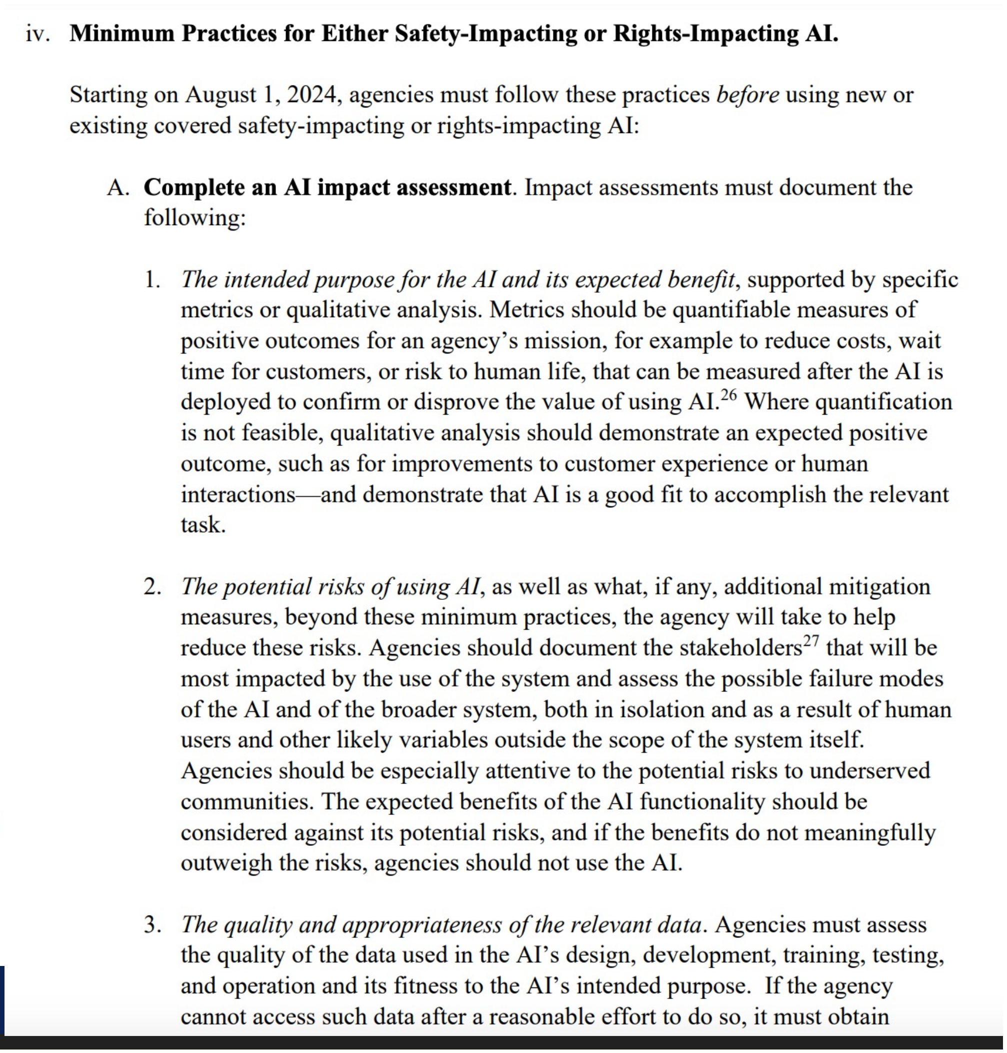
PROPOSED MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Shalanda D. Young

SUBJECT: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence

c. Minimum Practices for Safety-Impacting and Rights-Impacting Artificial Intelligence

Except as prevented by applicable law and governmentwide guidance, agencies must apply the minimum practices in this section to safety-impacting and rights-impacting AI by August 1, 2024, or else stop using the AI until it becomes compliant. Prior to August 1, 2024, agency CAIOs should work with their agencies' relevant officials to bring potentially non-compliant AI into conformity, which may include voluntary requests to third-party vendors to take appropriate action (e.g., via updated documentation or testing measures). To ensure compliance with this requirement, relevant agency officials must use existing mechanisms wherever possible, for example, the Authorization to Operate process. An agency may also request an extension or grant a waiver to this requirement through its CAIO using the processes detailed below.



iv. **Minimum Practices for Either Safety-Impacting or Rights-Impacting AI.**

Starting on August 1, 2024, agencies must follow these practices *before* using new or existing covered safety-impacting or rights-impacting AI:

A. **Complete an AI impact assessment**. Impact assessments must document the following:

1. *The intended purpose for the AI and its expected benefit*, supported by specific metrics or qualitative analysis. Metrics should be quantifiable measures of positive outcomes for an agency's mission, for example to reduce costs, wait time for customers, or risk to human life, that can be measured after the AI is deployed to confirm or disprove the value of using AI.[26] Where quantification is not feasible, qualitative analysis should demonstrate an expected positive outcome, such as for improvements to customer experience or human interactions—and demonstrate that AI is a good fit to accomplish the relevant task.

2. *The potential risks of using AI*, as well as what, if any, additional mitigation measures, beyond these minimum practices, the agency will take to help reduce these risks. Agencies should document the stakeholders[27] that will be most impacted by the use of the system and assess the possible failure modes of the AI and of the broader system, both in isolation and as a result of human users and other likely variables outside the scope of the system itself. Agencies should be especially attentive to the potential risks to underserved communities. The expected benefits of the AI functionality should be considered against its potential risks, and if the benefits do not meaningfully outweigh the risks, agencies should not use the AI.

3. *The quality and appropriateness of the relevant data*. Agencies must assess the quality of the data used in the AI's design, development, training, testing, and operation and its fitness to the AI's intended purpose. If the agency cannot access such data after a reasonable effort to do so, it must obtain

Minimum Practices

Safety- and Rights-impacting AI

Before use:

- ❑ Complete an AI impact assessment
  - ❑ The intended purpose for the AI and its expected benefit
  - ❑ The potential risks of using AI
  - ❑ The quality and appropriateness of the relevant data
- ❑ Test the AI for performance in a real-world context
- ❑ Independently evaluate the AI

Ongoing requirements:

- ❑ Conduct ongoing monitoring and establish thresholds for periodic human review
- ❑ Mitigate emerging risks to rights and safety
- ❑ Ensure adequate human training and assessment
- ❑ Provide appropriate human consideration as part of decisions that pose a high risk to rights or safety
- ❑ Provide public notice and plain-language documentation through the AI use case inventory.

Rights-impacting AI

- ❑ Take steps to ensure that the AI will advance equity, dignity, and fairness
  - ❑ Proactively identifying and removing factors contributing to algorithmic discrimination or bias
  - ❑ Assessing and mitigating disparate impacts
  - ❑ Using representative data
- ❑ Consult and incorporate feedback from affected groups.

- ❑ Conduct ongoing monitoring and mitigation for AI-enabled discrimination
- ❑ Notify negatively affected individuals
- ❑ Maintain human consideration and remedy processes
- ❑ Maintain options to opt-out where practicable