

# Applications of CAST to Understand and Mitigate Risks in AI Systems

SEPTEMBER, 2025

Matt Walsh | Senior Data Scientist (mmwalsh@cert.org)  
David Schulker | Senior Data Scientist (dschulker@cert.org)



# Document Markings

Copyright 2025 Carnegie Mellon University.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at [permission@sei.cmu.edu](mailto:permission@sei.cmu.edu).

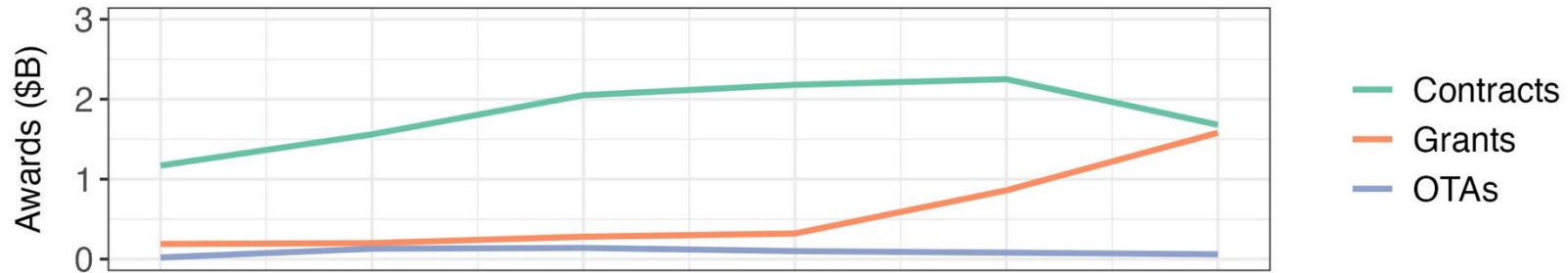
DM25-0667

Applications of CAST to Mitigate Risks in AI Systems

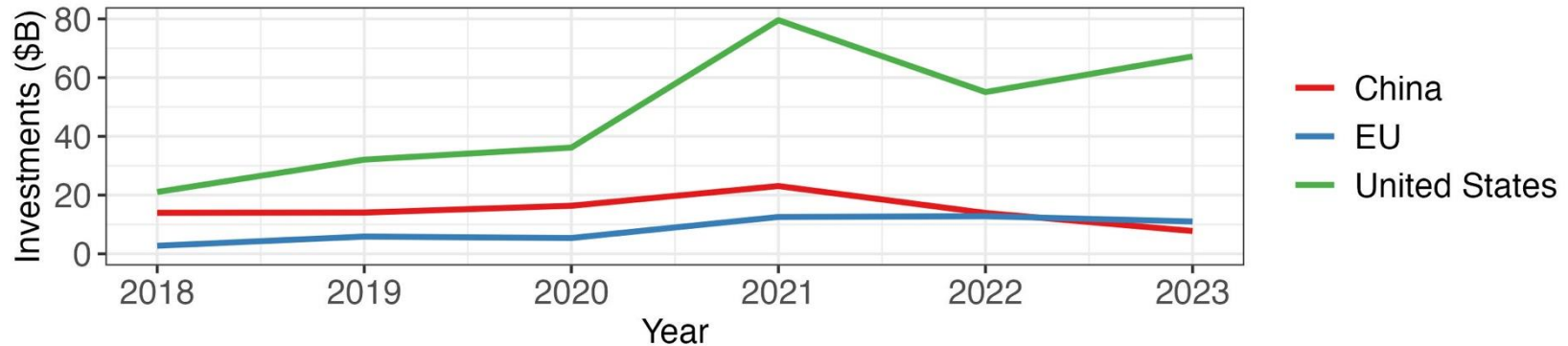
# Background

# Government and industry are investing heavily in AI

## U.S. Government Investments in AI by Funding Vehicle

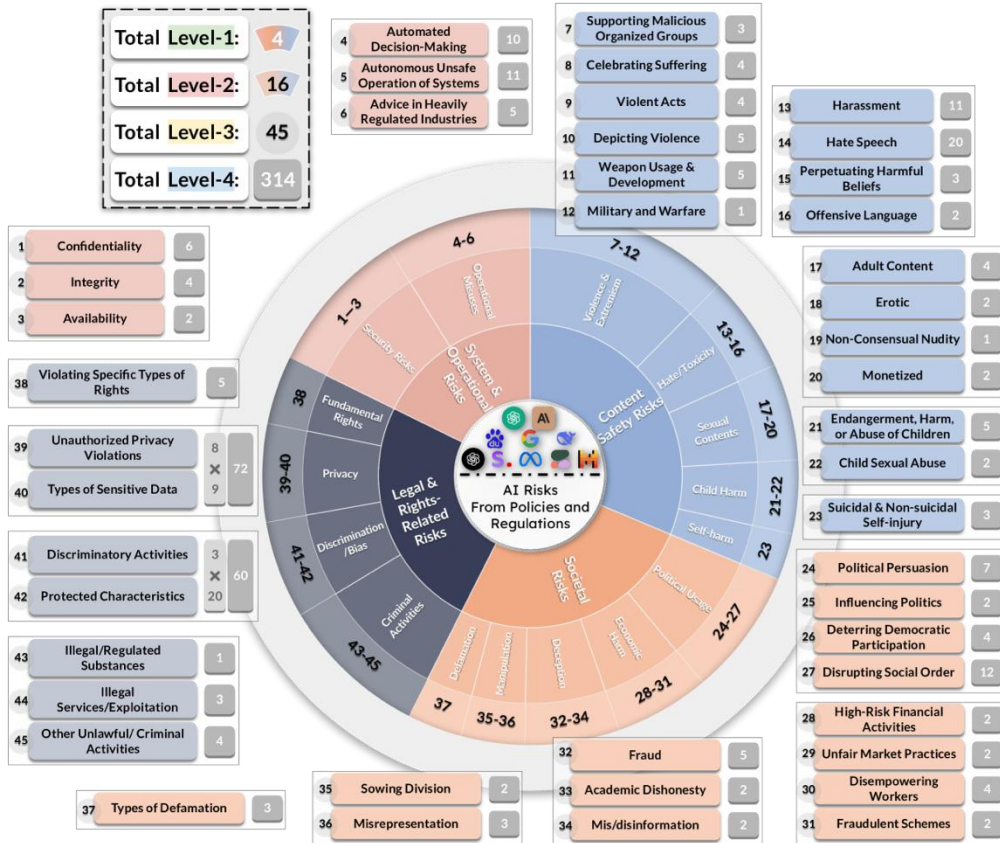


## Private Investments in AI by Region



HAI, 2024

# AI has tremendous potential, but it also presents risks



Risk Types	Definition
System and Operational	Misuse or attack of system
Content and Safety	Generation of harmful content, intentional or not
Societal	Weaponization of system to cause political or economic harm
Legal and Rights Related	Rights violations during system development or use

AIR, 2024

# Goals of our research

- **Investigate evidence for AI incidents**
- **Analyze associated losses**
- **Apply system theory to enhance AI safety**

Applications of CAST to Mitigate Risks in AI Systems

# Quantitative Analysis of AI Incidents

# There has been a proliferation of AI incident databases

**AIID**



AI incidents and controversies

**MITRE ATLAS**



Attacks of AI Systems

**AIAAIC**



AI, algorithms, and automation

# We focused our analysis on AIID

## AIID



AI incidents and controversies

“The AIID is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. Like similar databases in aviation and computer security, the AIID aims to learn from experience so we can prevent or mitigate bad outcomes.”

AIID, 2025

CAST may be an effective tool to learn from these experiences

# This is what an AIID incident looks like

**AI INCIDENT DATABASE**

English | Sign Up

## Incident 124: Algorithmic Health Risk Scores Underestimated Black Patients' Needs

**Description**

Welcome to the AIID

**Description:** Optum's algorithm deployed by a large academic hospital was revealed by researchers to have under-predicted the health needs of black patients, effectively de-prioritizing them in extra care programs relative to white patients with the same health burden.

**Tools**

Notify Me of Updates | + New Report | + New Response | Discover | Citation Info | View History

**Tags**

**Entities** [View all entities](#)

Alleged: [Optum](#) developed an AI system deployed by [unnamed large academic hospital](#), which harmed [Black patients](#).

**Metadata**

**Incident Stats**

Incident ID	124
Report Count	7
Incident Date	2019-10-24
Editors	Sean McGregor, Khoa Lam
Applied Taxonomies	<a href="#">CSETVI</a> , <a href="#">GMF</a> , <a href="#">MIT</a>

**Taxonomy**

**CSETVI Taxonomy Classifications** [Taxonomy Details](#)

**Similar Incidents**

By textual similarity

Did our AI mess up? Flag the unrelated incidents

**COMPAS Algorithm Performs Poorly in Crime Recidivism Prediction**

May 2016 - 22 reports

**Kidney Testing Method Allegedly Underestimated Risk of Black Patients**

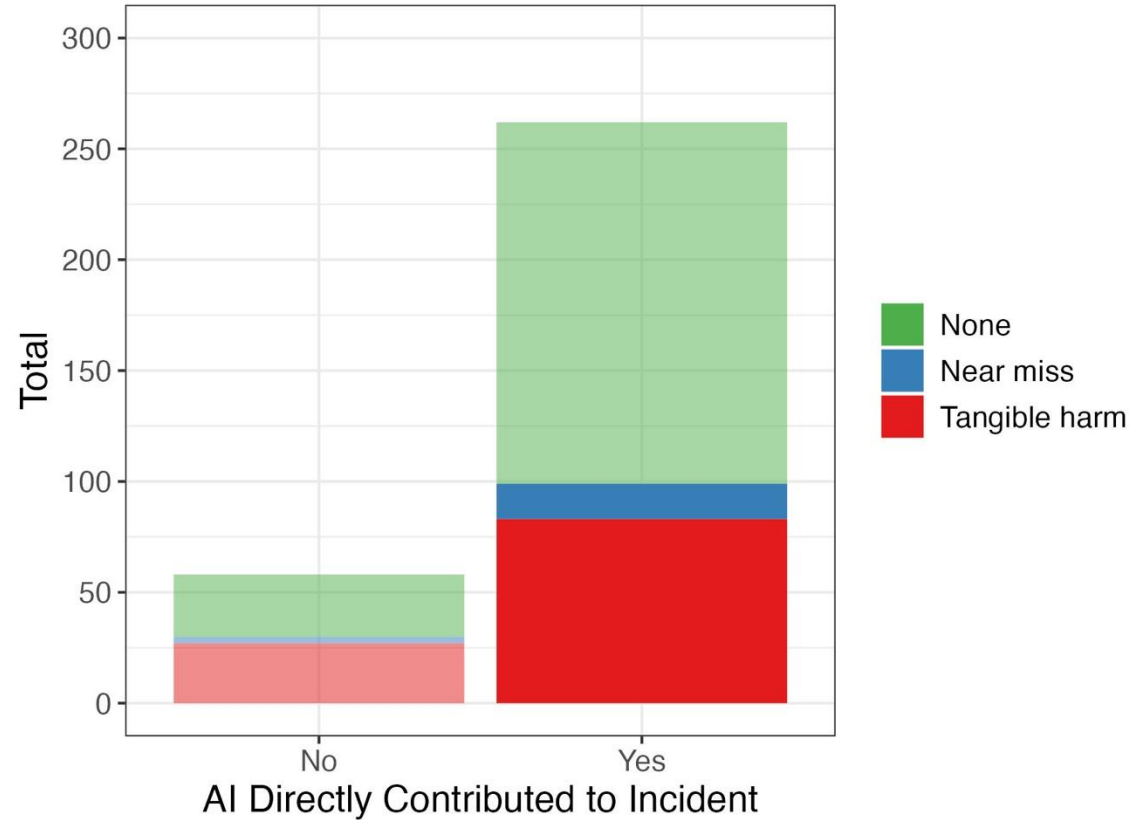
Mar 1999 - 3 reports

# CSET taxonomy has been applied to a subset of incidents

Category	Responses	Definitions	
Tangible Harm	Tangible harm	A specific entity experiences tangible harm (injury, loss, or damage).	← Losses
	Near miss	A specific entity experiences <u>imminent risk</u> of tangible harm (injury, loss, or damage).	← Hazards
	None	A specific entity experiences <u>non-imminent risk</u> of tangible harm (injury, loss, or damage).	← UCAs
AI System	Yes	Technologies and processes in which AI <u>plays</u> a meaningful role	
	No	Technologies and processes in which AI <u>does not play</u> a meaningful role	
Physical Objects	Yes	AI system(s) is embedded in hardware that <u>can directly interact</u> with, affect, and change the physical world	
	No	AI system(s) is embedded in hardware that <u>cannot directly interact</u> with, affect, and change the physical world	

# We focused on AI tangible harms and near misses

Category	Responses
Tangible Harm	Tangible harm
	Near miss
	None
AI System	Yes
	No
Physical Objects	Yes
	No

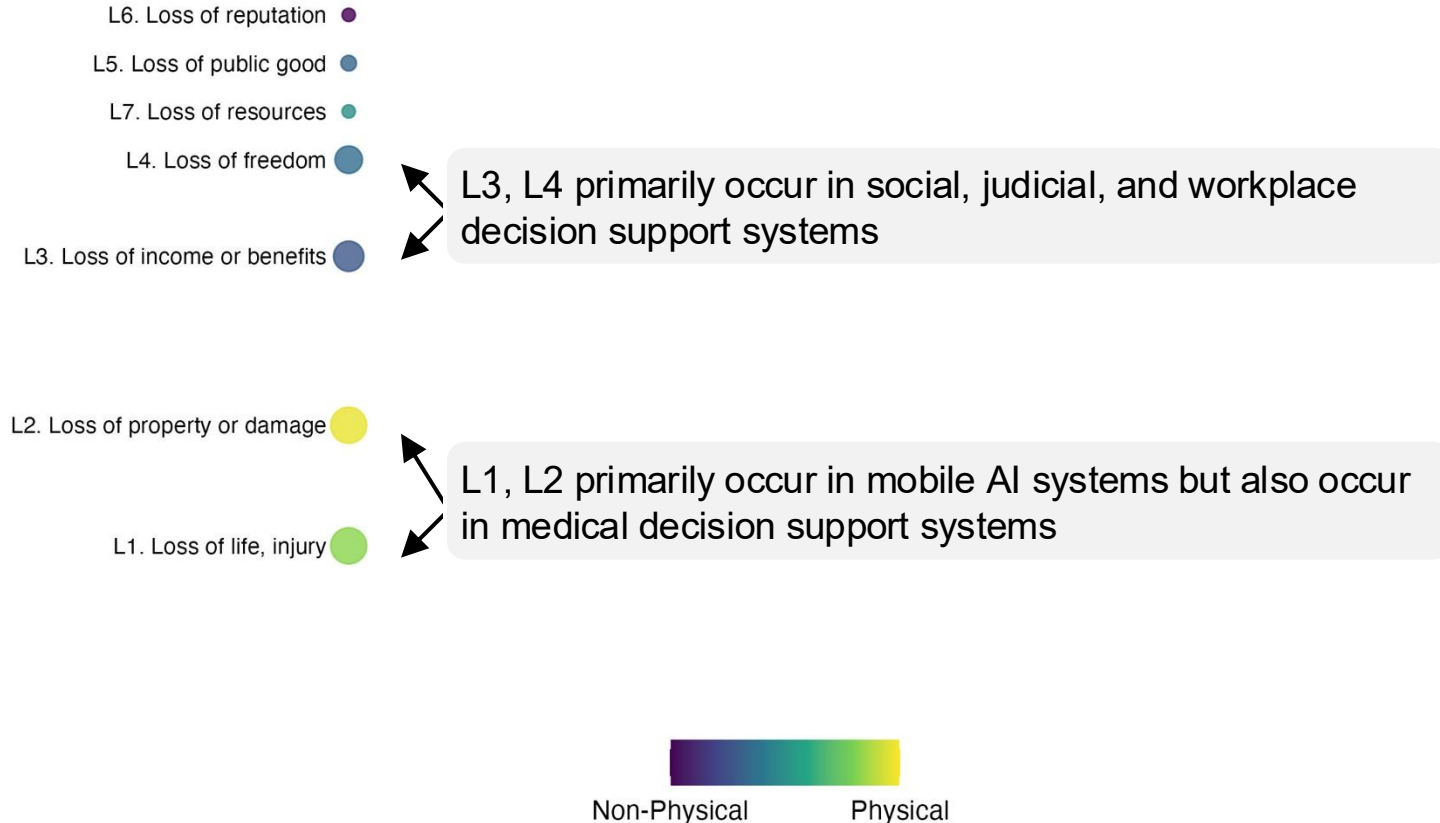


# Our analysis of 100 incidents revealed 7 losses

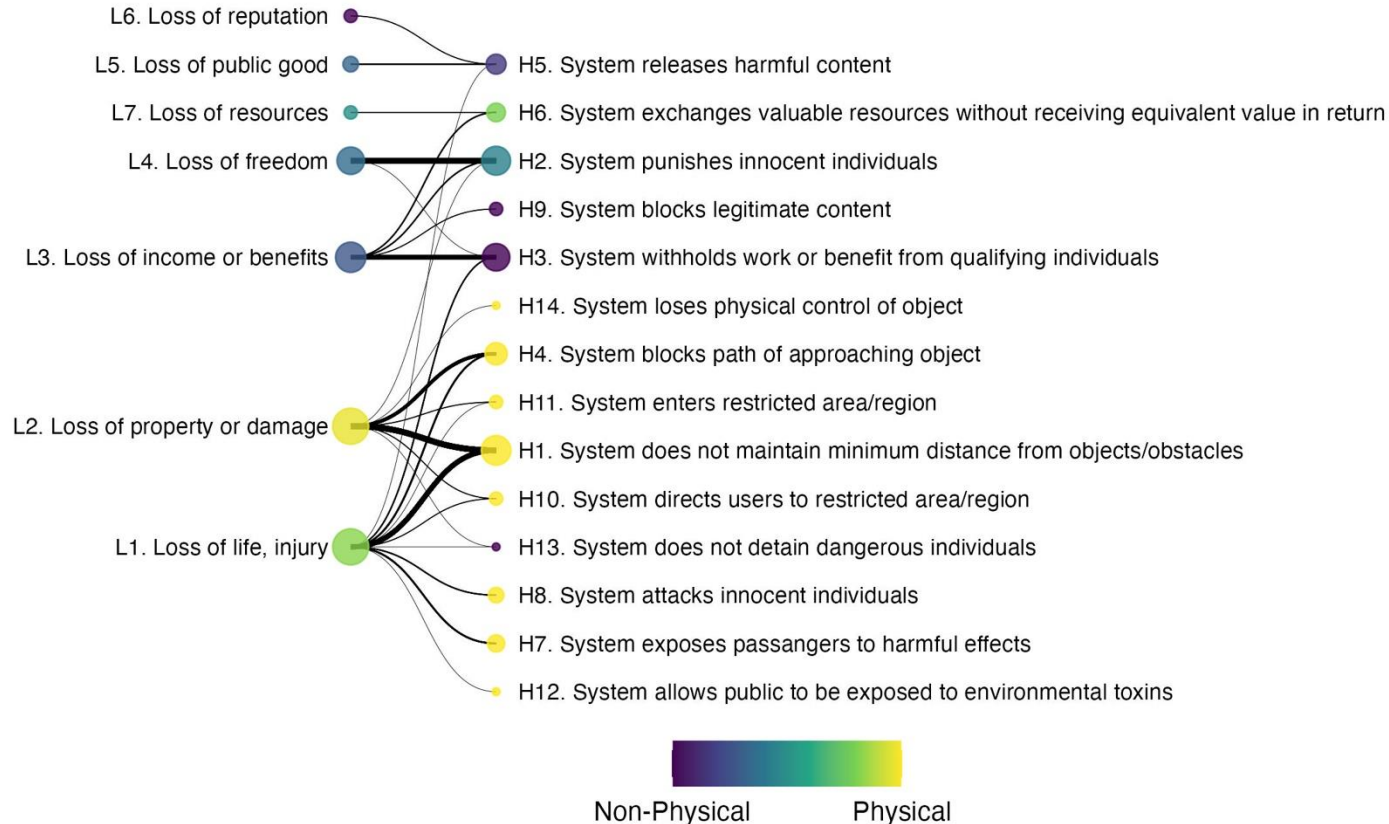
- L6. Loss of reputation ●
- L5. Loss of public good ●
- L7. Loss of resources ●
- L4. Loss of freedom ●
- L3. Loss of income or benefits ●
- L2. Loss of property or damage ●
- L1. Loss of life, injury ●



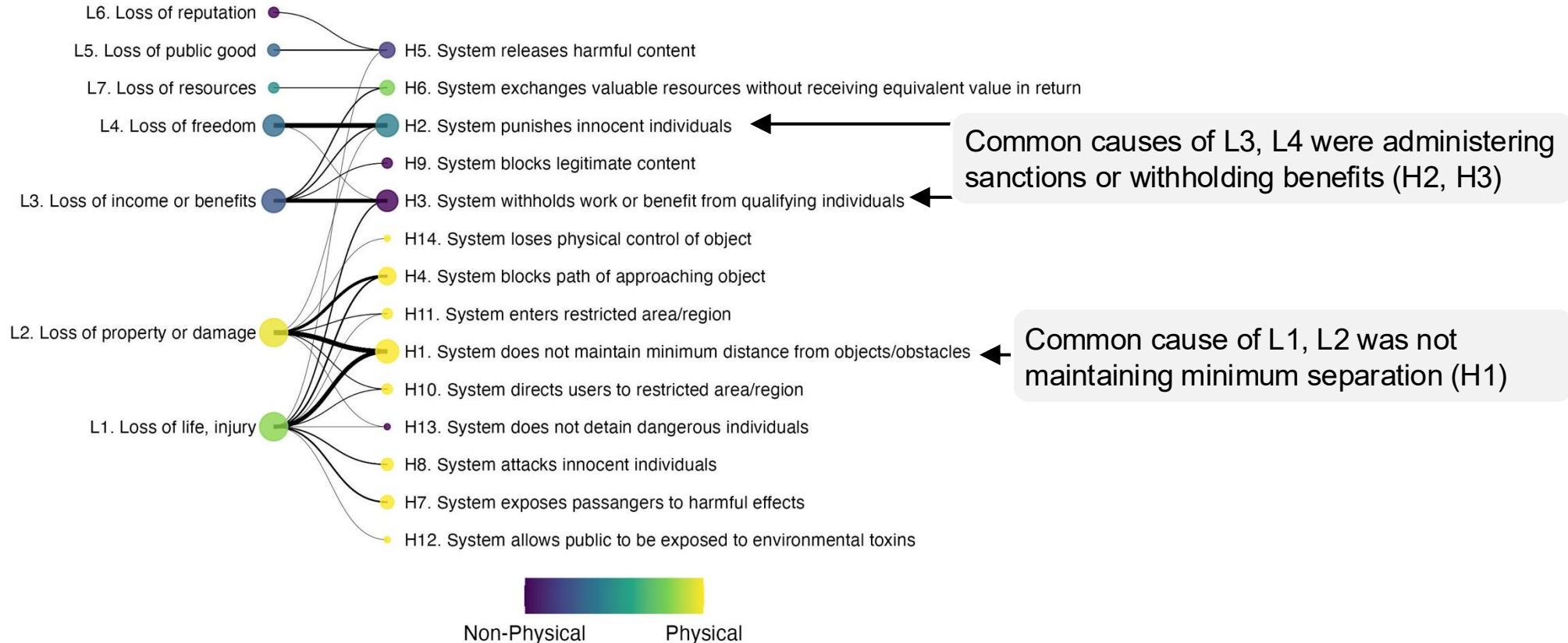
# Our analysis of 100 incidents revealed 7 losses



# Our analysis of 100 incidents revealed 14 hazards



# Our analysis of 100 incidents revealed 14 hazards



# Findings from quantitative analysis

- **Recurring set of losses and hazards across 100 incidents**
  - Most common losses involved life/injury and property/damage
  - Other common losses involved freedom and income/benefits
- **Losses affect physical *and* non-physical AI systems**
- **Losses and hazards are not AI-specific**
  - However, UCAs, scenarios and solutions may be AI-specific

How can we learn and apply from these incidents?

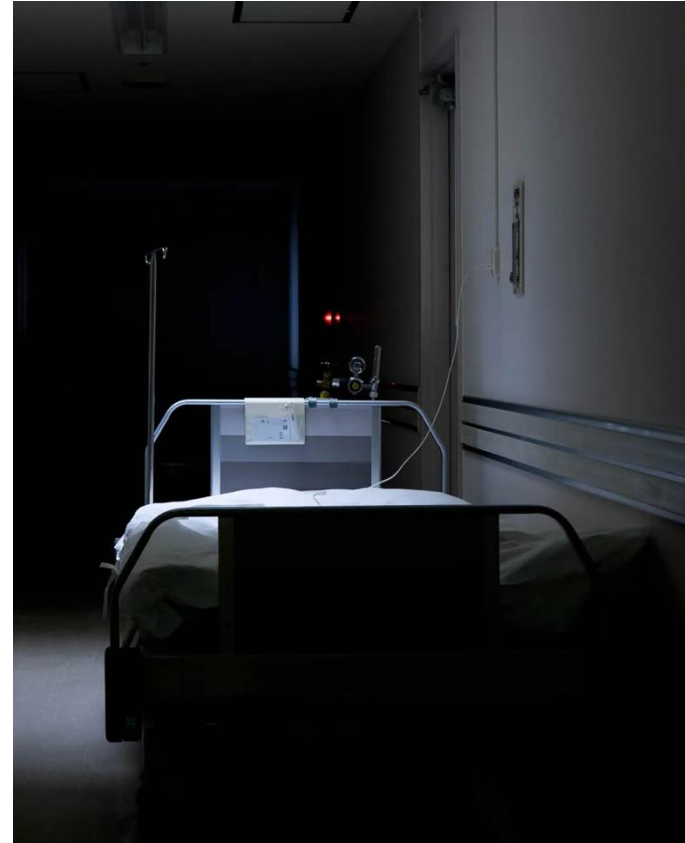
Applications of CAST to Mitigate Risks in AI Systems

# CAST Analysis of an AI Incident

# Incident 124. High-risk care management program

A health system used an ML model to decide who to enroll in a “high-risk care management” program. The model gave each patient a “risk” score. Patients with the highest scores were automatically enrolled in the program.

Black patients enrolled in the program were significantly sicker than white patients, suggesting that the model systematically assigned lower risk scores to Black patients. This disparity was subsequently confirmed.



# 1. Assemble basic information

## Losses, hazards, and safety constraints

<b>Losses</b>	Loss of life/injury [L1]
<b>Hazards</b>	System did not enroll high-risk patients in care management program [H1]
<b>Safety Constraints</b>	System must enroll high-risk patients in care management program [SC1]

Post-hoc analysis of patients enrolled in high-risk program revealed racial disparity, which led to identification of L1 and H1

# 1. Assemble basic information

## Proximal events

ID	What happened?
1	Health systems implement “high-risk care management” programs to proactively treat patients with complex health needs.
2	However, due to the cost of the program, only a limited number of individuals can be enrolled
3	Developers trained an ML algorithm on historic insurance claims to predict future costs, using this as a proxy for ‘medical need.’
4	Hospitals used the algorithm to assign risk scores to patients.
5	Patients above the 97th percentile were automatically enrolled, while those above the 55th percentile were referred for evaluation.
6	A retrospective analysis found that black patients admitted to the program were significantly sicker than white patients with the same score, suggesting unequal access to care for non-admitted black patients.

# 1. Assemble basic information

## Proximal events

ID	What happened?
1	Health systems implement “high-risk care management” programs to proactively treat patients with complex health needs.
2	However, due to the cost of the program, only a limited number of individuals can be enrolled
3	Developers trained an ML algorithm on historic insurance claims to predict future costs, using this as a proxy for ‘medical need.’
4	Hospitals used the algorithm to assign risk scores to patients.
5	Patients above the 97th percentile were automatically enrolled, while those above the 55th percentile were referred for evaluation.
6	A retrospective analysis found that black patients admitted to the program were significantly sicker than white patients with the same score, suggesting unequal access to care for non-admitted black patients.

Certain groups underutilize essential medical services, decoupling claims and need



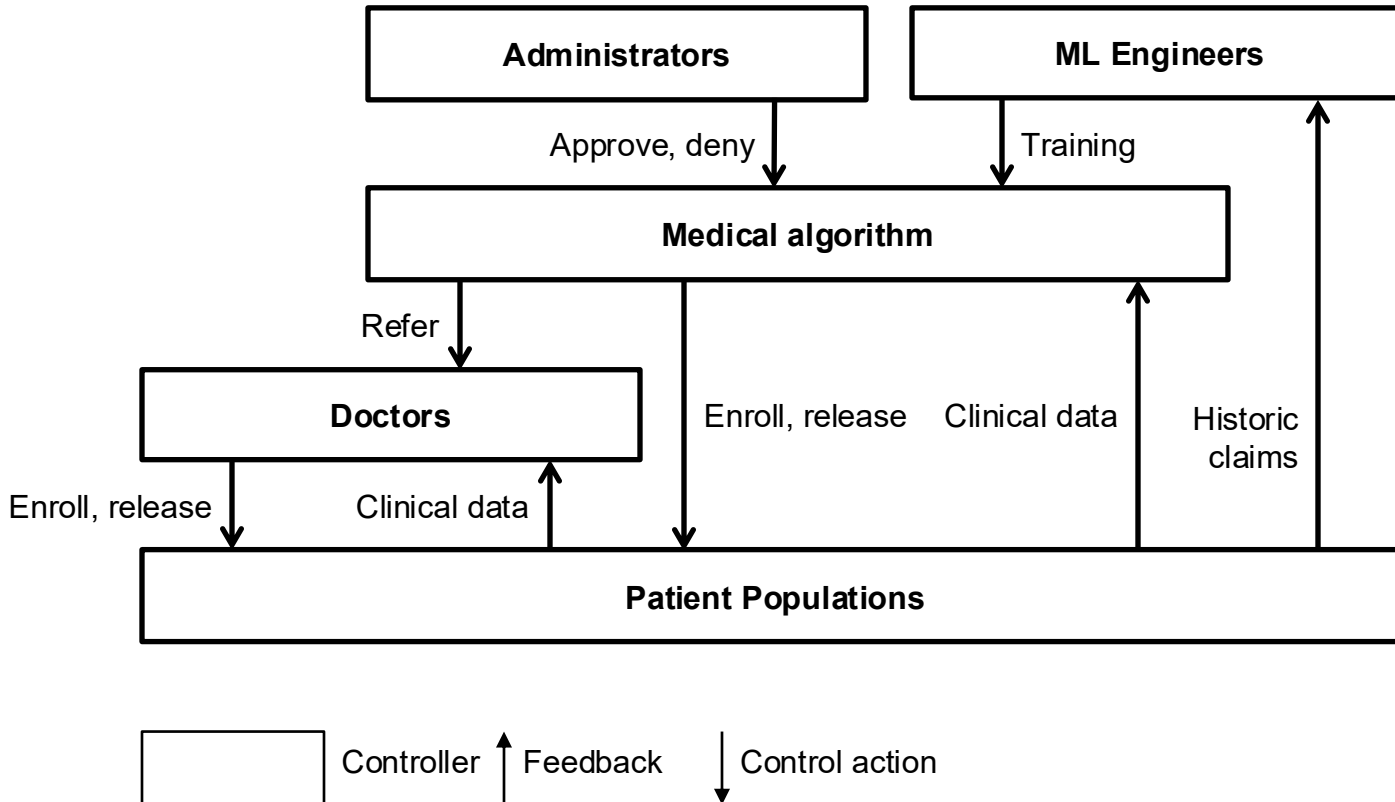
Algorithm underpredicts needs of individuals from those groups



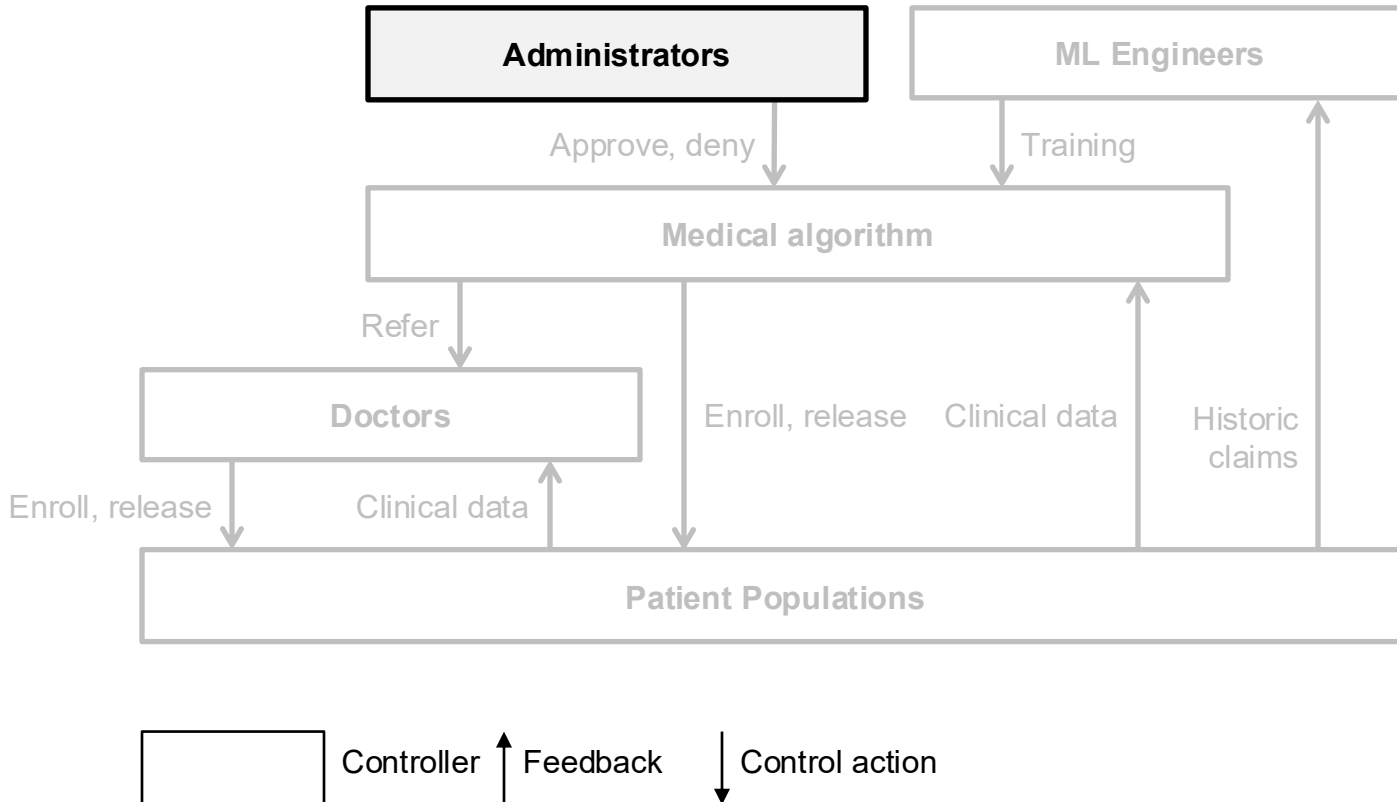
Benefits are withheld from those individuals

## 2. Model the control structure

### Boundaries based on healthcare system



### 3. Analyze each component in the loss



#### Safety Responsibilities

Ensure effective distribution of medical resources

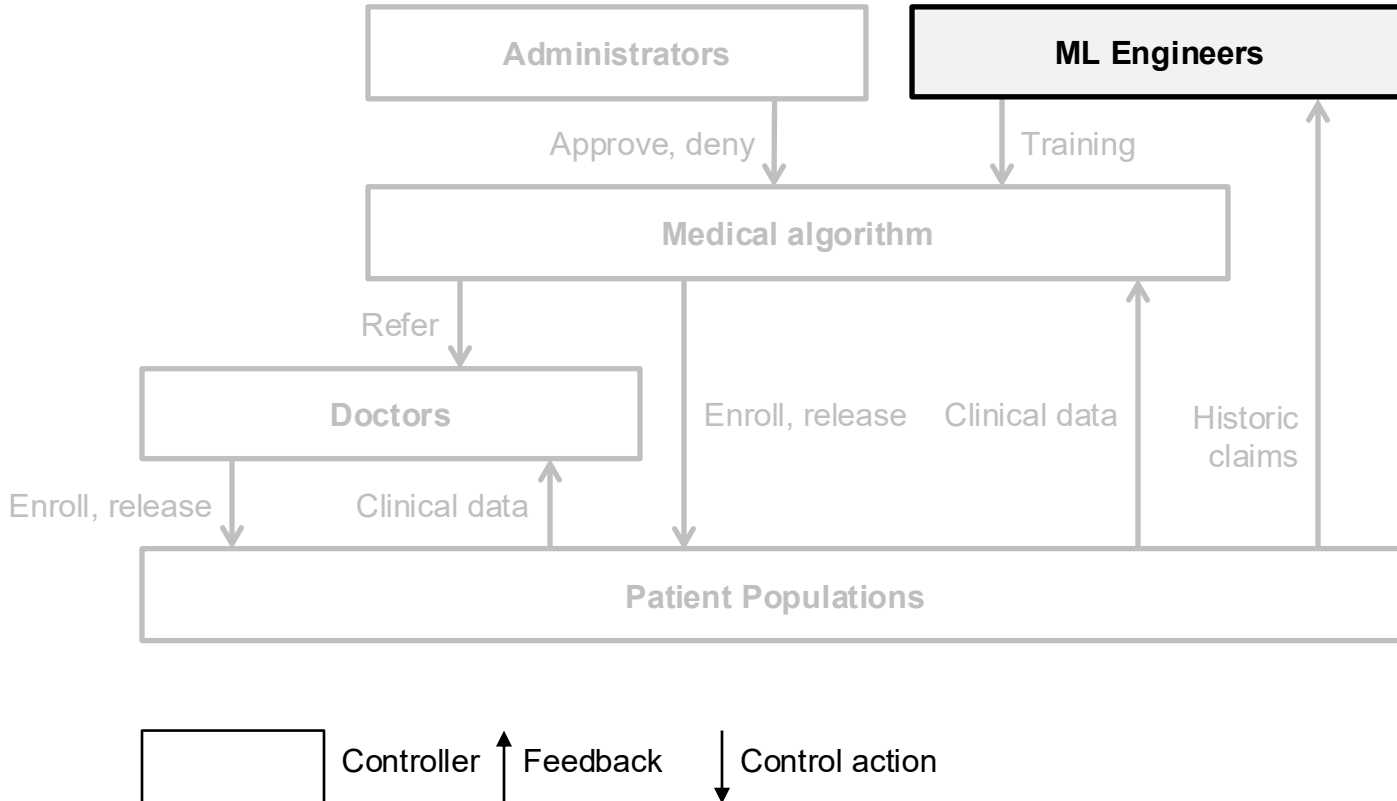
#### Role in loss

Approved use of model when model predicted an imperfect proxy variable for *patient need*

#### Why?

- Black-box model
- Missing feedback about effectiveness during deployment

### 3. Analyze each component in the loss



#### Safety Responsibilities

Create model for effective distribution of medical resources

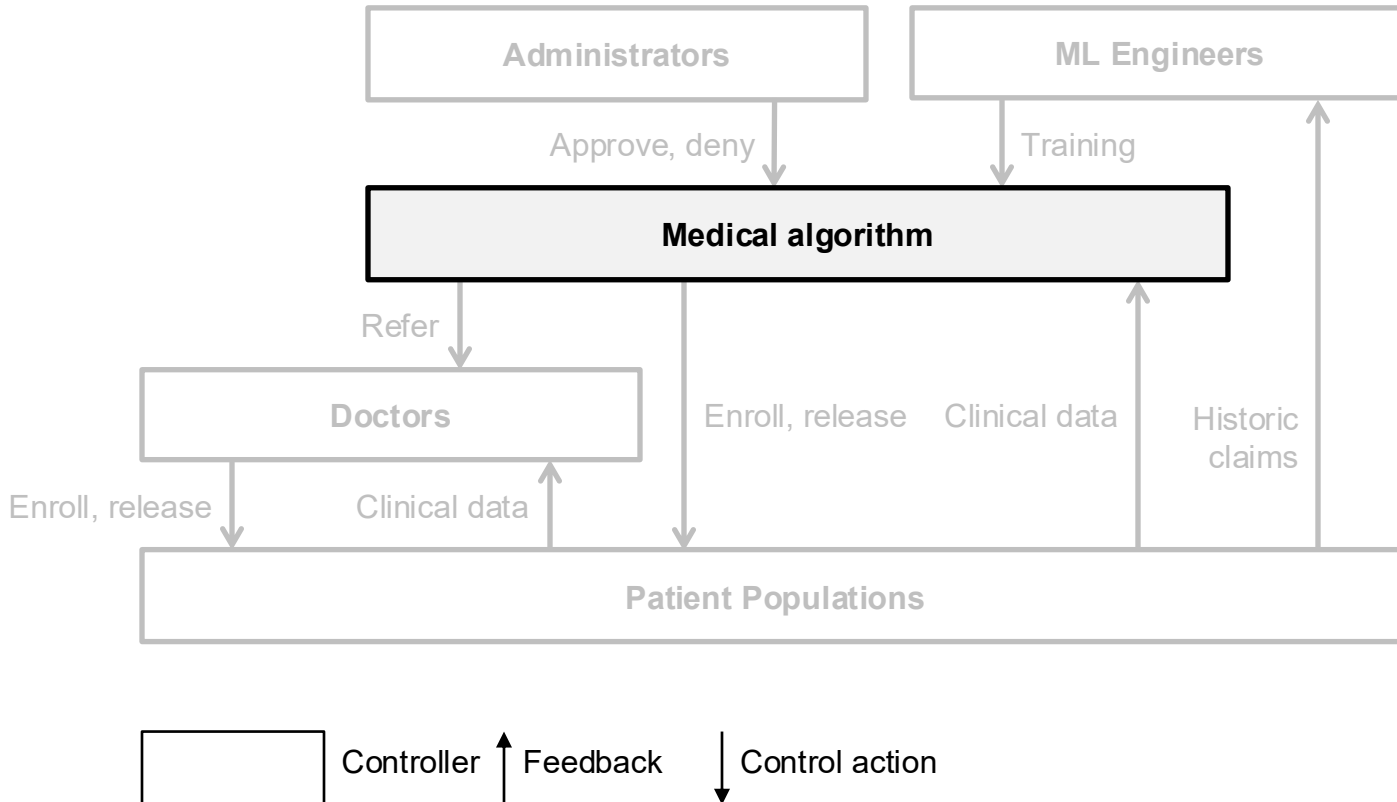
#### Role in loss

Selected proxy variable when proxy variable gave an imperfect measure of *patient need*

#### Why?

- Limited awareness of medical and social context
- Better measure of medical need not available
- Missing feedback about effectiveness during deployment

### 3. Analyze each component in the loss



#### Safety Responsibilities

Assign patients with complex needs to high-risk care management program

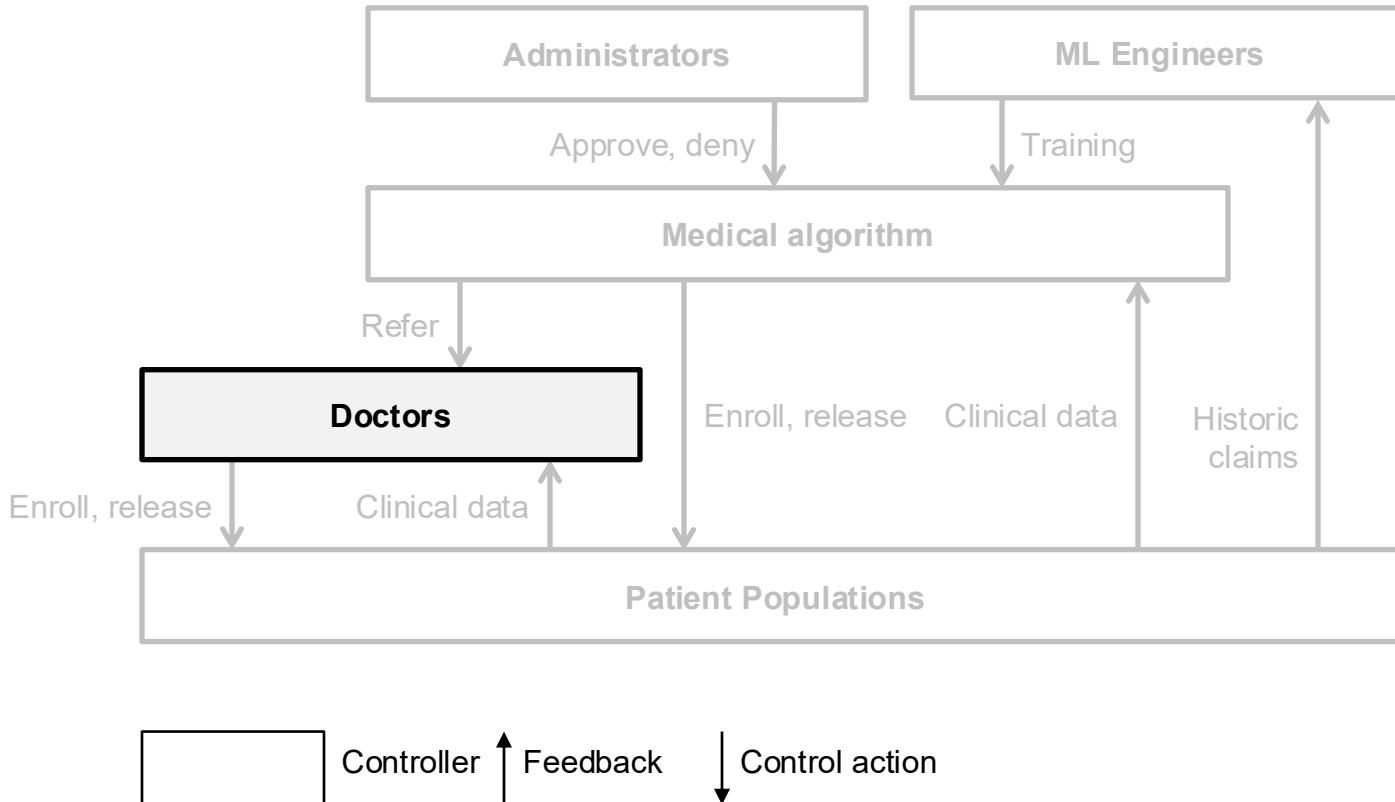
#### Role in loss

Did not assign patients to high-risk management program when patients had complex needs

#### Why?

- Essentially, the model was predicting future claims rather than future needs

### 3. Analyze each component in the loss



#### Safety Responsibilities

Assign patients with complex needs to high-risk care management program

#### Role in loss

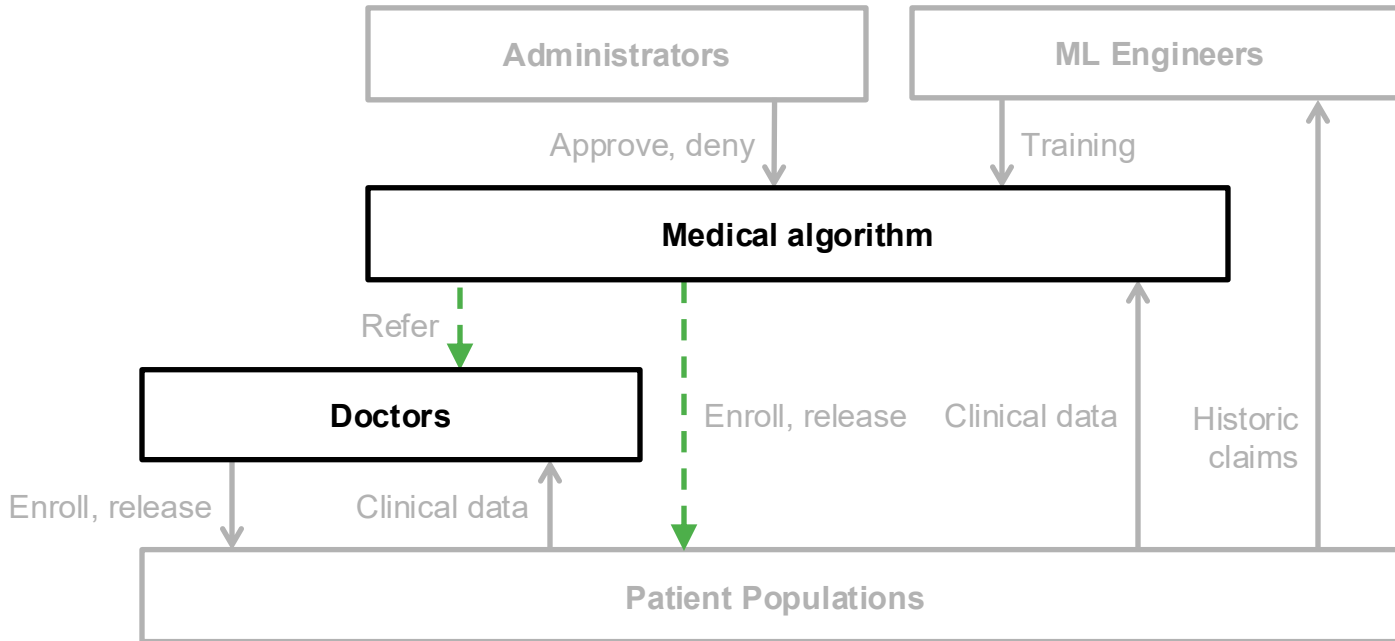
Did not assign patients to high-risk management program when patients had complex needs

#### Why?

- Model made automated assignments
- Trusted the model
- Exhibited similar bias to model when assigning patients

# 4. Identify control structure flaws

## Early system design



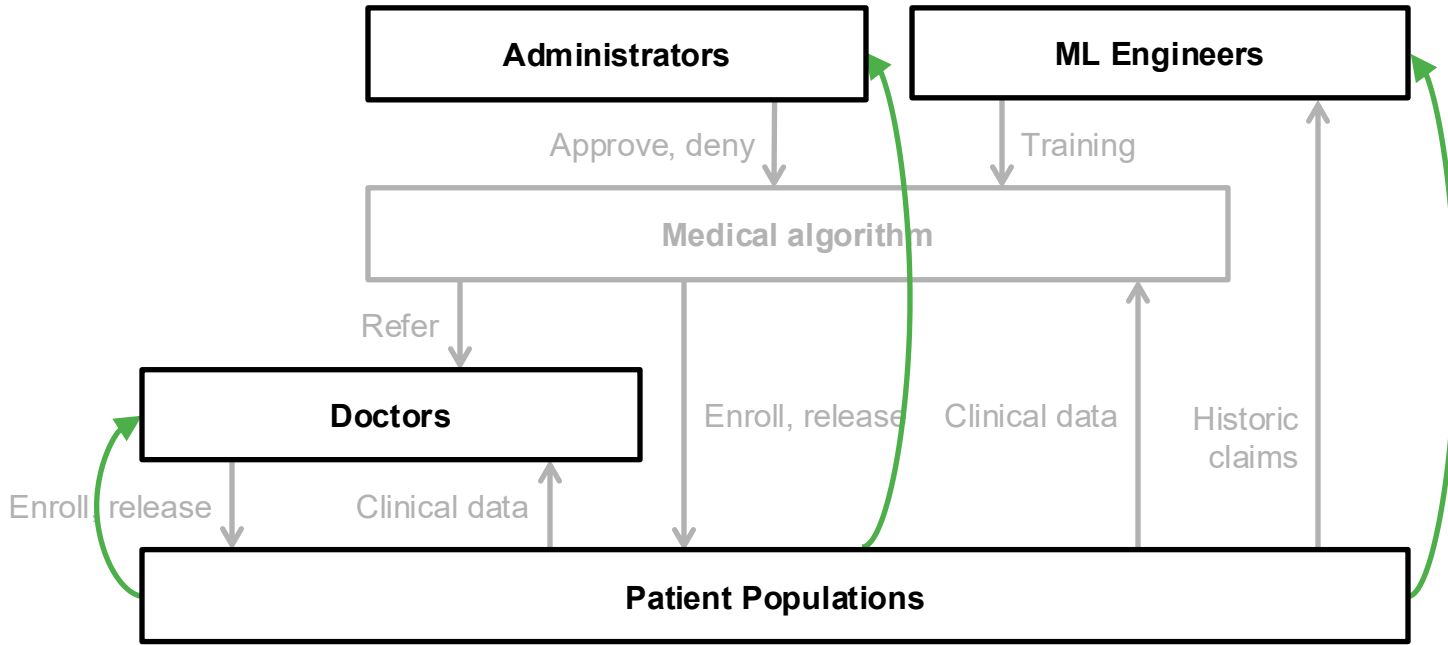
- Medical algorithm has a high degree of decision authority
- Alternate architectures include:
  - Give recommendation to doctor
  - Give risk score to doctor
  - Highlight risk factors for doctor
  - Audit doctors' decisions

Would these work? Only if doctors don't have the same bias

# 4. Identify control structure flaws

## Missing feedback

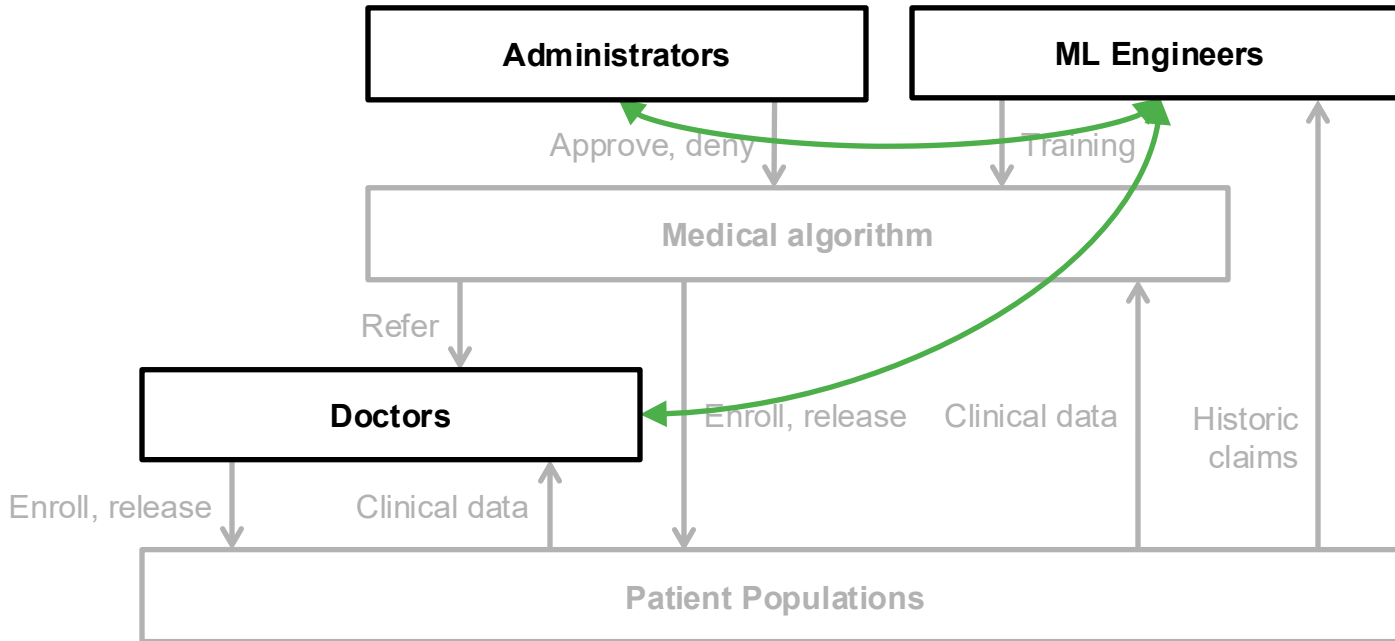
- Doctors, administrators, and ML engineers lack feedback about decisions and performance of model in situ



Candidate measure – Active number of chronic conditions in patients enrolled in program

# 4. Identify control structure flaws

## Communication and coordination



- Administrators and doctors lack awareness of how model assigns risk scores
- Administrators and doctors have social and domain knowledge that could enhance model design

## 5. Create Improvement Program

### **REC-1. Include SMEs and diverse panels in algorithm design and review**

- Rec Context: ML engineers may be unaware of biases and limitations present in model and or training data. SMEs and diverse panels have the proper knowledge to identify these problems.

### **REC-2. Select different label for prediction model**

- Rec Context: Insurance claims data are the result of complex aggregation. Predict an alternate measure of health, such as number of active chronic health conditions.

## 5. Create Improvement Program

### **REC-3. Collect and publish de-identified results of algorithmic decisions**

- Rec Context: Doctors, administrators, and patient populations lack feedback about group-level trends in model decisions. Collecting and publishing these data enables continuous monitoring and oversight.

### **REC-4. Explore alternate human-machine architectures**

- Rec Context: The model could be used in different ways. For example, to make a *recommendation* to clinical providers, or to *audit* their decisions. A qualitatively different human-machine architecture might be better suited for allocating medical resources in an effective and equitable manner.

# Findings from CAST analysis

- **AI incidents stem from more than just AI itself**
  - System theory helps identify and mitigate contributing factors across all controllers and their interactions
- **Certain UCAs and scenarios recur across AI systems**
  - Solutions include transparency, SME involvement, and continuous monitoring
- **Human-in-the-loop does not eliminate risk**
  - Humans may over trust algorithms
  - Humans may share similar biases

Applications of CAST to Mitigate Risks in AI Systems

# Discussion

# Contributions

- **Established that system and operational risks have manifest as losses in fielded AI systems**
- **Demonstrated that physical and non-physical AI systems can both produce losses**
- **Showed how CAST to be used to create safer systems following losses**