Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro,[1] Carol J. Smith,[2] Tamara Zilovic[1]

HCI Institute - Carnegie Mellon University,[1] Software Engineering Institute, Carnegie Mellon University[2]
nikmart@cmu.edu, cjsmith@sei.cmu.edu, tzilovic@andrew.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

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.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

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.

DM22-0473

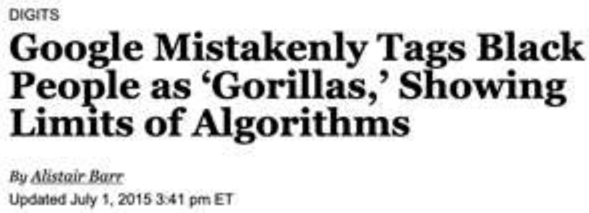
Image Source: NASA

Image Source: National Transportation Safety Board

Image Source: South Jordan Police Department via

Image Source:
Kyle Grillot, AFP/Getty Images via USA Today

Like other complex systems,
AI systems will fail.

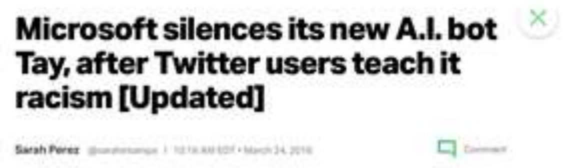


DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr

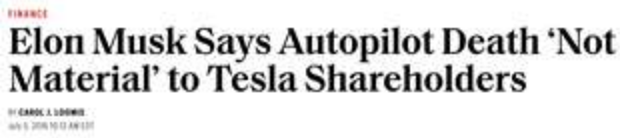
Updated July 1, 2015 3:41 pm ET



Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez [illegible] / 10:16 AM EDT • March 24, 2016

Comment

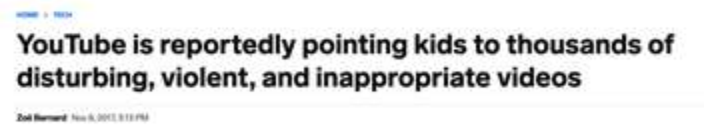


FINANCE

Elon Musk Says Autopilot Death 'Not Material' to Tesla Shareholders

BY CAROL J. LOOMIS

July 5, 2016 10:13 AM EDT



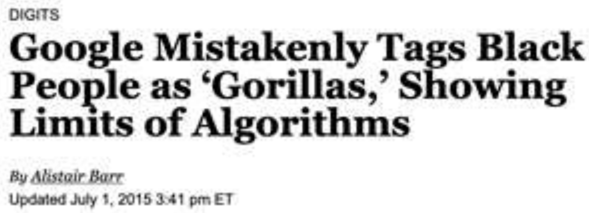
HOME > TECH

YouTube is reportedly pointing kids to thousands of disturbing, violent, and inappropriate videos

Zoë Bernard Nov 6, 2017, 9:13 PM

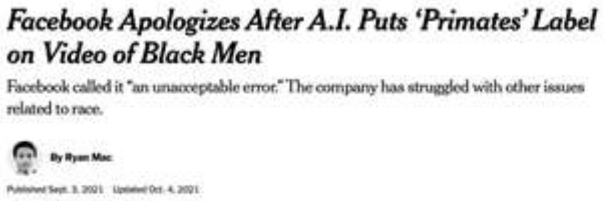
Failures of AI-enabled products and services have far-reaching effects.

But have we learned from our mistakes?


DIGITS

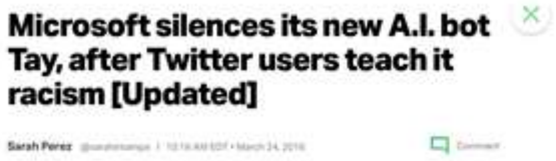
Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By *Alistair Barr*
Updated July 1, 2015 3:41 pm ET


Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

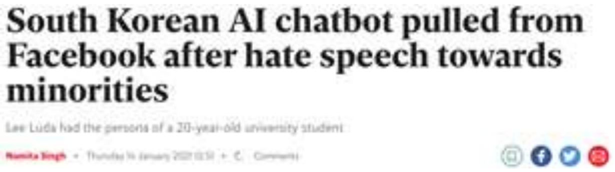
Facebook called it "an unacceptable error." The company has struggled with other issues related to race.

By Ryan Mac
Published Sept. 3, 2021 Updated Oct. 4, 2021


Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

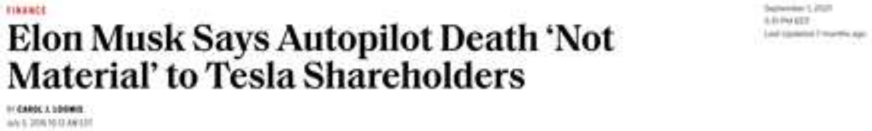
Sarah Perez @sarahintampa / 10:16 AM EDT • March 24, 2016

Comment


South Korean AI chatbot pulled from Facebook after hate speech towards minorities

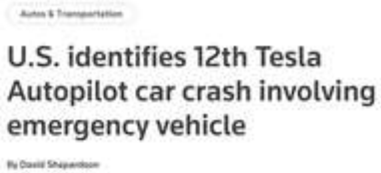
Lee Luda had the persona of a 20-year-old university student

Namita Singh • Thursday 14 January 2021 12:51 • Comments


FINANCE

Elon Musk Says Autopilot Death 'Not Material' to Tesla Shareholders

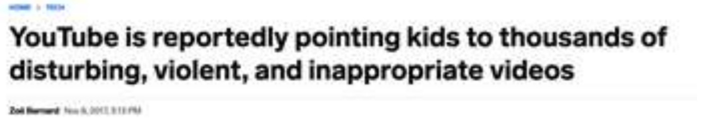
By CAROL J. LOOMIS
July 5, 2016 10:13 AM EDT


September 1, 2021
6:31 PM EDT
Last Updated 7 months ago

Autos & Transportation

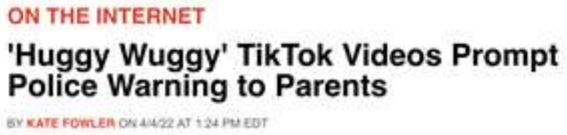
U.S. identifies 12th Tesla Autopilot car crash involving emergency vehicle

By David Shepardson


HOME > TECH

YouTube is reportedly pointing kids to thousands of disturbing, violent, and inappropriate videos

Zoë Bernard Nov 6, 2017, 9:13 PM

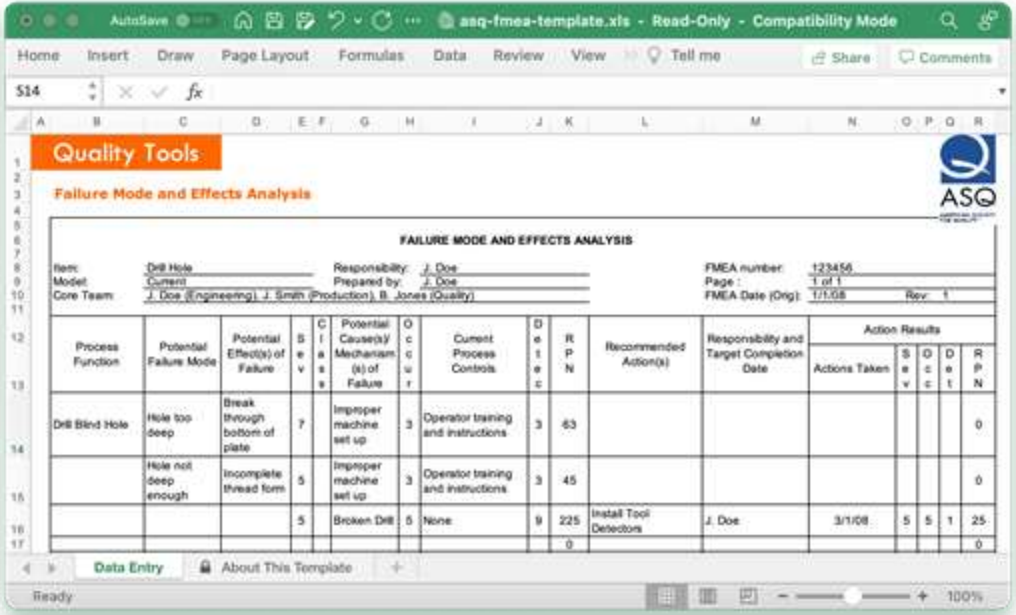

ON THE INTERNET

'Huggy Wuggy' TikTok Videos Prompt Police Warning to Parents

BY KATE FOWLER ON 4/4/22 AT 1:24 PM EDT

Hazard Analysis

Any **activity** that preemptively aims to identify and address potential safety and/or ethical concerns related to a system or product.



Failure Mode and Effects Analysis

FAILURE MODE AND EFFECTS ANALYSIS

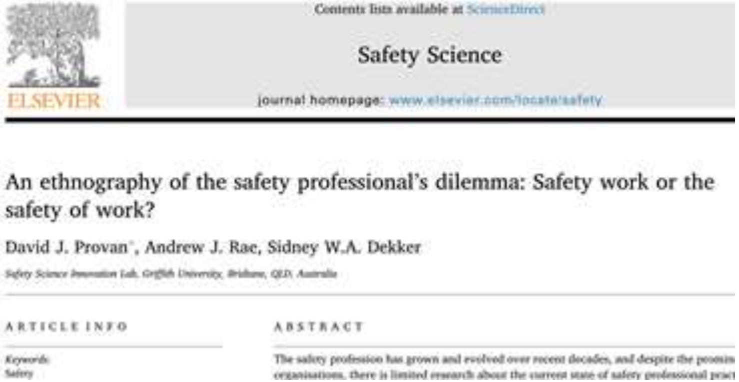
Item: Drill Hole
Model: Current
Core Team: J. Doe (Engineering), J. Smith (Production), B. Jones (Quality)
Responsibility: J. Doe
Prepared by: J. Doe
FMEA number: 123456
Page : 1 of 1
FMEA Date (Orig): 1/1/08 Rev: 1

Process Function	Potential Failure Mode	Potential Effect(s) of Failure	Sev	Class	Potential Cause(s)/ Mechanism (s) of Failure	Occur	Current Process Controls	Detec	RPN	Recommended Action(s)	Responsibility and Target Completion Date	Action Results				
												Actions Taken	Sev	Occ	Det	RPN
Drill Blind Hole	Hole too deep	Break through bottom of plate	7		Improper machine set up	3	Operator training and instructions	3	63							0
	Hole not deep enough	Incomplete thread form	5		Improper machine set up	3	Operator training and instructions	3	45							0
			5		Broken Drill	5	None	9	225	Install Tool Detectors	J. Doe	3/1/08	5	5	1	25
									0							0

Source: American Society for Quality - FMEA Template: https://asq.org/-/media/public/learn-about-quality/data-collection-analysis-tools/asq-fmea-template.xls?la=en

Industry Hazard Analysis methods are generally driven by regulations, not safety science.

But what happens when there are currently no regulations for what you are developing?



Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

An ethnography of the safety professional's dilemma: Safety work or the safety of work?

David J. Provan*, Andrew J. Rae, Sidney W.A. Dekker

Safety Science Innovation Lab, Griffith University, Brisbane, QLD, Australia

ARTICLE INFO

Keywords:
Safety

ABSTRACT

The safety profession has grown and evolved over recent decades, and despite the promis
organisations, there is limited research about the current state of safety professional pract

Provan, D. J., Rae, A. J., & Dekker, S. W. A. (2019). An ethnography of the safety professional's dilemma: Safety work or the safety of work? Safety Science

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

Discussion

A manifesto for Reality-based Safety Science

Andrew Rae[a,*], David Provan[a], Hossam Aboelssaad[b], Rob Alexander[c]

[a] Griffith University, Brisbane, Australia
[b] University of Queensland, Brisbane, Australia
[c] University of York, York, United Kingdom

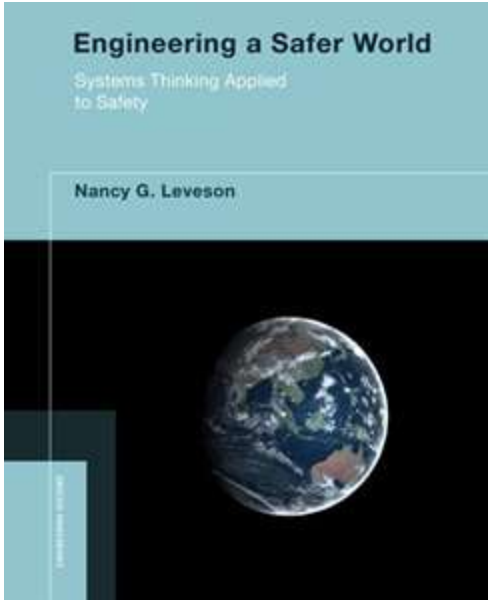
ARTICLE INFO

Keywords:
Reality-based Safety Science

ABSTRACT

In the field of safety science, we have stopped competing empirically. The theorists fight
notes and editorials, the empiricists tinker within the boundaries of existing theory, and

Rae, A., Provan, D., Aboelssaad, H., & Alexander, R. (2020). *A manifesto for Reality-based Safety Science.*



Engineering a Safe World: Systems Thinking Applied to Safety (2012) Nancy G. Leveson

Traditional models of hazard analysis assume linear causality.
As AI presents non-deterministic behavior, new methods must reflect more complex models of causality.

Preliminary hazard analysis generally encompasses generative and collaborative brainstorming sessions.

Gamification may increase engagement.

Judgment Call

How to play

Ballard, S., Chappell, K. M., & Kennedy, K. (2019). Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology

Figure 1: What Could Go Wrong? A digital card game to foster conversation around potential challenges and issues of autonomous vehicle technology

2.2 "What Could Go Wrong?" Card Game

Martelaro, N., & Ju, W. (2020). What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles.

Motivation …

Can we develop new **structured thinking methods** and **systems engineering tools** to support effective and engaging ways for *preemptively* considering failure modes in AI systems?

11

Semi-Structured Interviews

~ 30 minutes/each

Discussions focused on …

- Current hazard analysis process
- What is/isn't working well
- Unique considerations of hazard analysis for AI-based systems
- Challenges with hazard/ risk considerations

9

Survey Responses (Recently Launched)

~ 20 minutes to complete

Questions focus on …

- Recent professional experiences surrounding hazard analysis
- Standards employed (ISO, IEC, IEEE)
- Formal hazard analysis processes used
- Tooling used (spreadsheets, project management software …)
- Satisfaction with current processes & tools

Initial Findings

Incompatibility of such processes with modern development practices

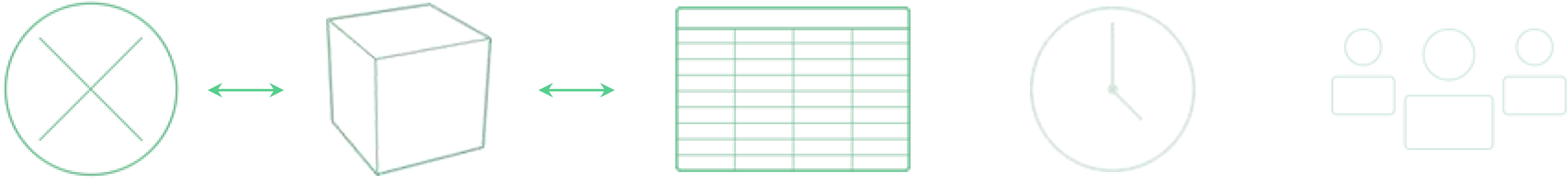
Unique challenges posed by working with non-deterministic ML systems

Limited Tooling available to support hazard analysis activities

Time pressures inherent to competitive markets

Role of **company culture** in the support of these efforts

Initial Findings



Incompatibility of such processes with modern development practices

Unique challenges posed by working with non-deterministic ML systems

Limited Tooling available to support hazard analysis activities

Time pressures inherent to competitive markets

Role of **company culture** in the support of these efforts

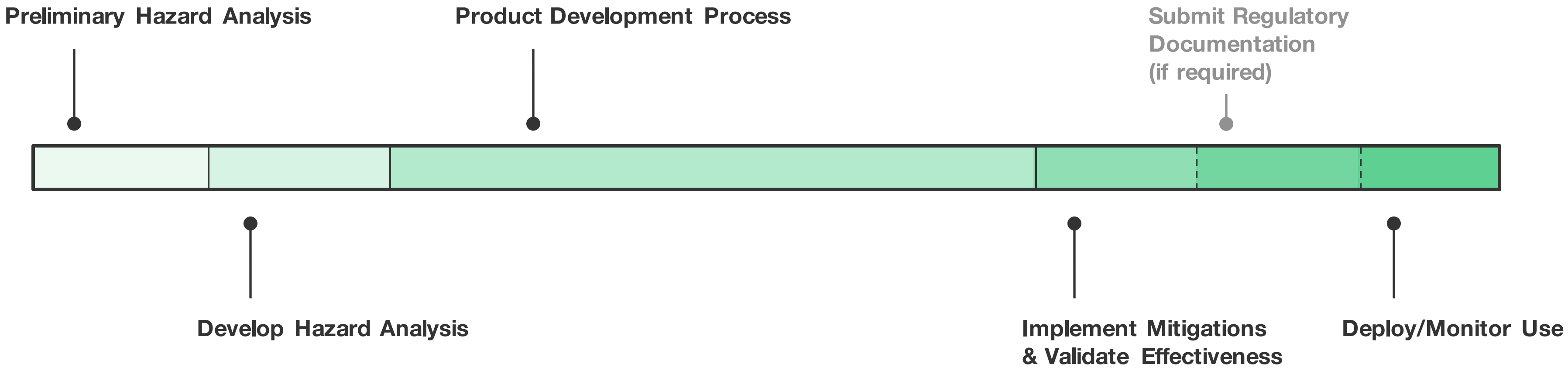
FOR THOSE IN INDUSTRY ....

Are you doing Hazard Analysis in your practice?
If so, how and when?

FOR THOSE IN ACADEMIA ....

Are you teaching Hazard Analysis in your classes?
If so, what and why?

HAZARD ANALYSIS TIMELINE



Preliminary Hazard Analysis

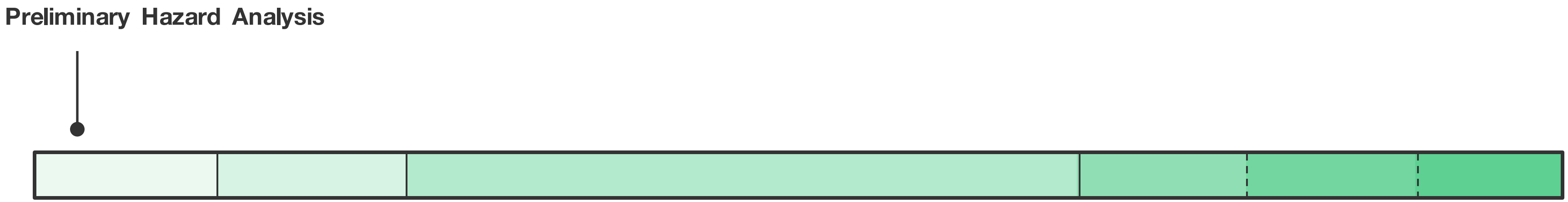
Product Development Process

Submit Regulatory Documentation (if required)

Develop Hazard Analysis

Implement Mitigations & Validate Effectiveness

Deploy/Monitor Use



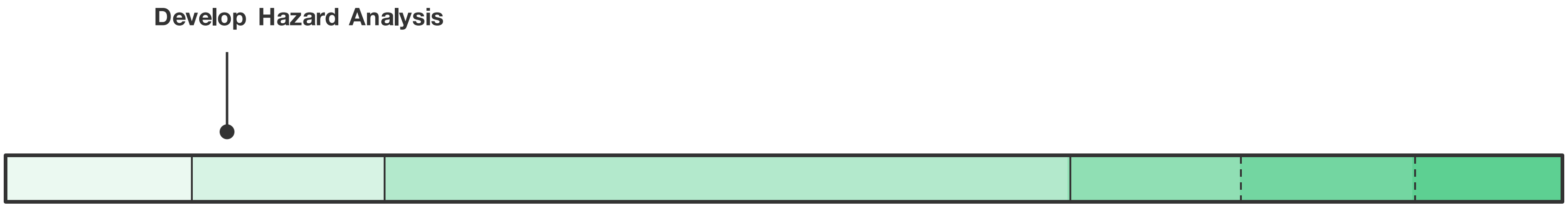
What is done?

- Group brainstorming/ ‘whiteboarding’ sessions involving multiple disciples
- Exploring combinations of components that may lead to hazardous situations (hazard, trigger event) as well as potential consequences

What opportunities are there?

- Better understanding of what has gone wrong with similar predicate products (Known Problems Analysis)
- Access to relevant and easily-searchable incident databases
- Better communication across disciplines (especially between technical & non-technical fields)

HAZARD ANALYSIS TIMELINE

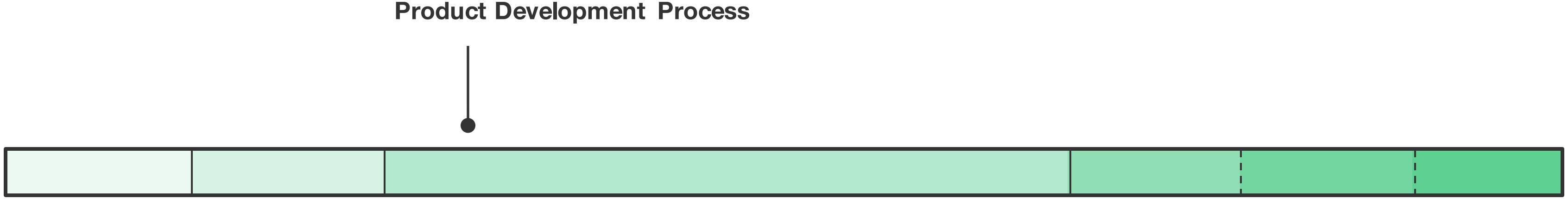


What is done?

- Smaller group of people (esp safety engineers) will work to refine & formalize content generated from preliminary HA
- Assignment of severity usually happens in this part of the process (negligible to catastrophic)
- Prioritization - often influenced by severity and probability of occurrence

What opportunities are there?

- Tool that could more accurately predict the probability of something occurring
- Tool that could suggest potential attack vectors
- Tailoring of risk assessment framework to better suit considerations of particular product
- System to better gauge the potential societal impacts of components/ algorithms, which is much more unique to AI development

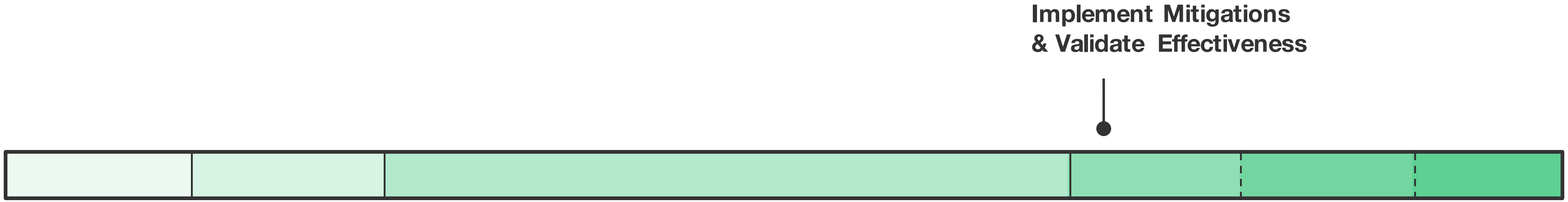


What is done?

- Iterative product development process
- Continuous changes to individual systems that may be tightly coupled with other systems overseen by different disciples
- Demonstrated need for traceability of requirements and design history states

What opportunities are there?

- Overall monitoring system that is cognizant of the coupling between and complexities of individual components
- Tool to help with traceability and version control
- Ability to provide quick, regular access for engineers to review (HA involves a lot of different review)
- Tool that could allow effective comparison of tradeoffs



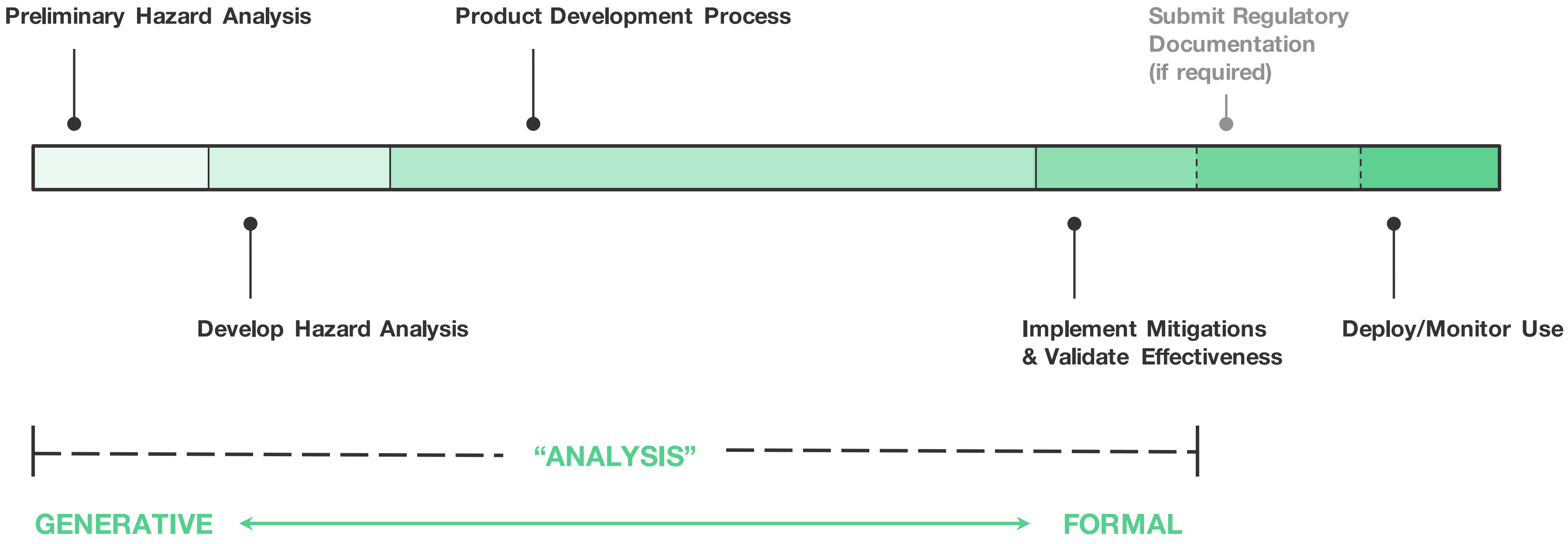
What is done?

- Implementing mitigations and validating the effectiveness of those mitigations typically occurs at later stages of product development
- If any residual risk remains, might have to do a residual risk assessment after
- Auditing and sub-auditing reports

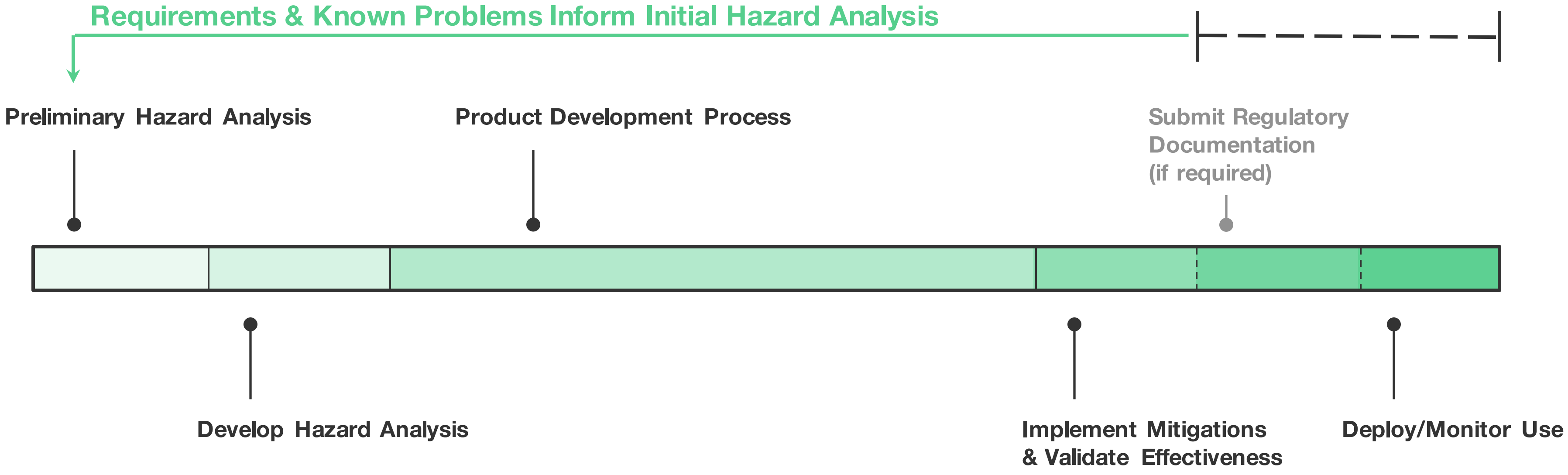
What opportunities are there?

- Tools that can easily add new potential hazards found to original list: During testing, may encounter series of events that might lead to a new hazard, then must update original list

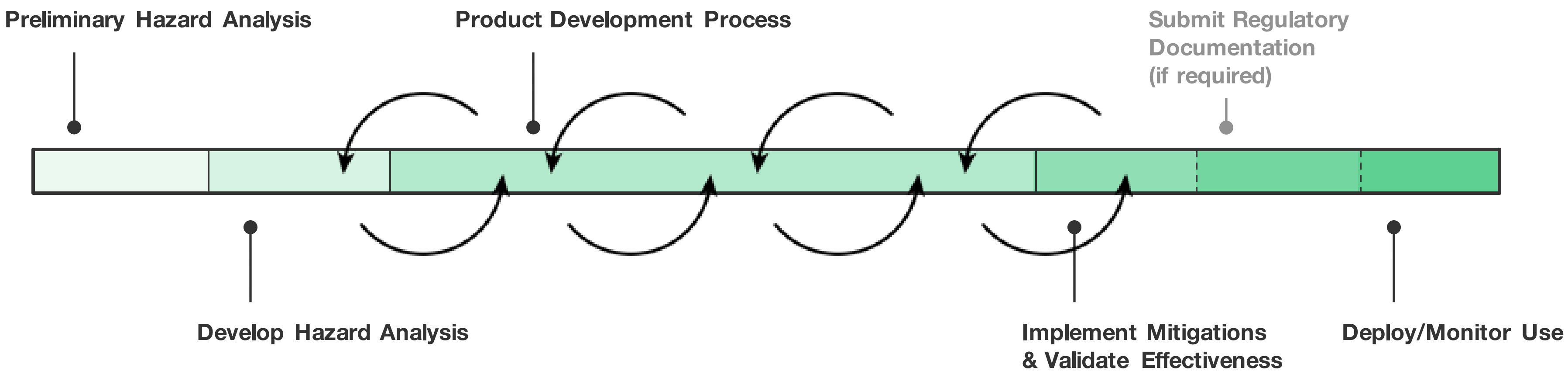
HAZARD ANALYSIS TIMELINE


Preliminary Hazard Analysis
Product Development Process
Submit Regulatory Documentation (if required)
Develop Hazard Analysis
Implement Mitigations & Validate Effectiveness
Deploy/Monitor Use
"ANALYSIS"
GENERATIVE
FORMAL

HAZARD ANALYSIS TIMELINE


Requirements & Known Problems Inform Initial Hazard Analysis
Preliminary Hazard Analysis
Product Development Process
Submit Regulatory Documentation (if required)
Develop Hazard Analysis
Implement Mitigations & Validate Effectiveness
Deploy/Monitor Use

HAZARD ANALYSIS TIMELINE



Preliminary Hazard Analysis

Product Development Process

Submit Regulatory Documentation (if required)

Develop Hazard Analysis

Implement Mitigations & Validate Effectiveness

Deploy/Monitor Use

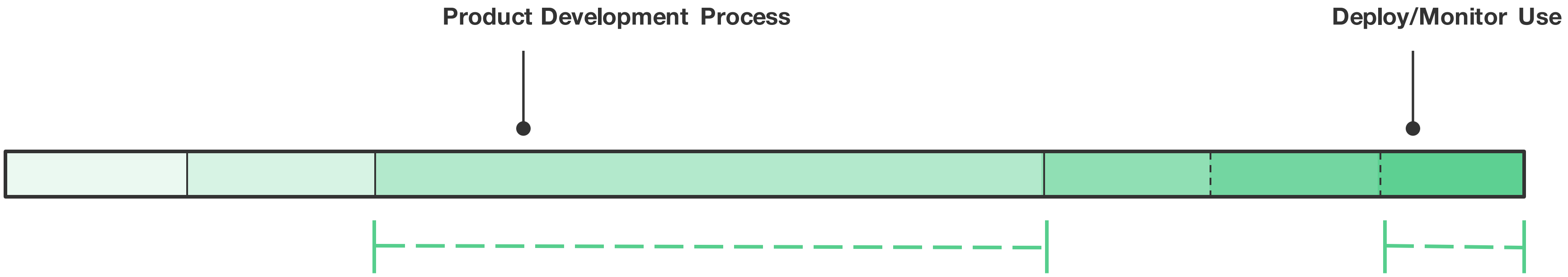
DISCUSSION - TIMELINE

FOR THOSE IN INDUSTRY ....

Is this timeline representative of your experiences in industry?
If not, how have your experiences diverged?

FOR THOSE IN ACADEMIA ....

At what stage of the development process do you feel has the most opportunity for research and improvement?



Where are incidents, near-misses, and other indicators of potential failure logged during the development process?

Government databases (VAERS, MAUDE, CAERS), Non-profit/ independent databases (ISMP, AIID, AIAAIC)

DISCUSSION - TIMELINE

FOR THOSE IN INDUSTRY ....

How does your organization document incidents/harms internally and/or externally?

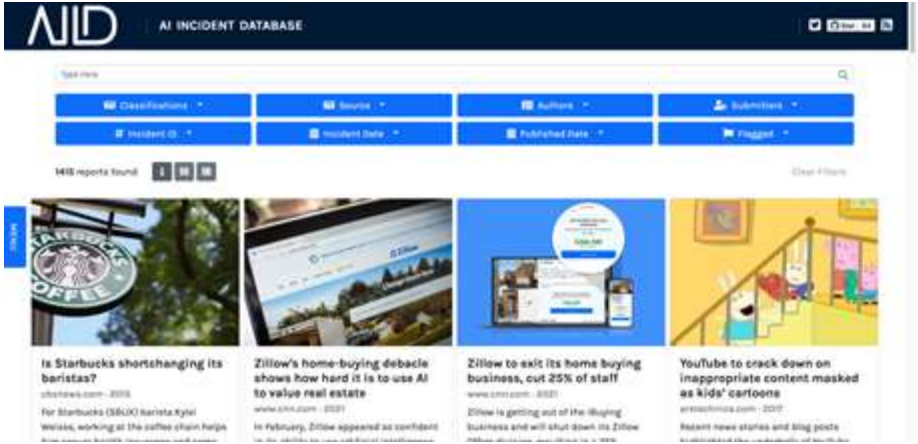
FOR THOSE IN ACADEMIA ....

Do you teach any documentation techniques in your classes?

AI INCIDENT DATABASE

"Much like the transportation sector before it (e.g., FAA and FARS) and more recently computer systems, intelligent systems require a repository of problems experienced in the real world so that future researchers and developers may mitigate or avoid repeated bad outcomes."

Source: AI Incident Database: https://incidentdatabase.ai/about



1416 reports found

Is Starbucks shortchanging its baristas?
cbsnews.com · 2015
For Starbucks (SBUX) barista Kylei Weisse, working at the coffee chain helps

Zillow's home-buying debacle shows how hard it is to use AI to value real estate
www.cnn.com · 2021
In February, Zillow appeared so confident

Zillow to exit its home buying business, cut 25% of staff
www.cnn.com · 2021
Zillow is getting out of the iBuying business and will shut down its Zillow

YouTube to crack down on inappropriate content masked as kids' cartoons
arstechnica.com · 2017
Recent news stories and blog posts

Source: AI Incident Database: https://incidentdatabase.ai/apps/discover

CYBERSECURITY

Source: NIST - National Vulnerabilities Database
https://nvd.nist.gov/

TRANSPORTATION

Source: Federal Aviation Administration -
Accident & Incident Data:
https://www.ntsb.gov/Pages/AviationQuery.aspx

Source: NHTSA - Fatality Analysis Reporting System:
https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

MEDICAL/PRODUCTS

Source: US FDA - MAUDE:
https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm

Source: US CDC & FDA - VAERS:
https://vaers.hhs.gov/data.html

Source: US FDA - CAERS:
https://www.fda.gov/food/compliance-enforcement-food/cfsan-adverse-event-reporting-system-caers#files

DISCUSSION - DATABASE USES

FOR THOSE IN INDUSTRY ....

How might the AI Incident Database provide the most use to you?

Are there other databases that you have successfully used?

FOR THOSE IN ACADEMIA ....

How might you leverage the AI Incident Database in your research or teaching?

Nikolas Martelaro
nikmart@cmu.edu

Carol J. Smith
cjsmith@sei.cmu.edu

Tamara Zilovic
tzilovic@andrew.cmu.edu

LEARN MORE + SHARE YOUR EXPERIENCES

Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro, Carol J. Smith, Tamara Zilovic

Paper presented at the AAAI Symposium on AI Engineering

tinyurl.com/hazards-ai-eng

Survey re: hazard analysis practice in industry

tinyurl.com/hazards-ai-survey