

Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro,¹ Carol J. Smith,² Tamara Zilovic¹

HCI Institute - Carnegie Mellon University,¹ Software Engineering Institute, Carnegie Mellon University²
nikmart@cmu.edu, cjsmith@sei.cmu.edu, tzilovic@andrew.cmu.edu

Copyright 2022 Carnegie Mellon University.

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.

[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.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0473



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 @sarahperez • 10:18 AM EDT • March 24, 2016

 Comment

FINANCE

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

By CAROL L. LOEWIS

July 5, 2016 10:13 AM EDT

HOME > TECH

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

Zolt Hornack • Nov 8, 2015, 9:19 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 @sarahperez • 10:18 AM EDT • March 24, 2016

Comment

FINANCE

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

By GARCIA LUDWIG
July 5, 2016 10:12 AM EDTSeptember 1, 2021
6:51 PM EDT
Last updated 7 months ago

Autos & Transportation

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

By David Shappellbaum

ON THE INTERNET

'Huggy Wuggy' TikTok Videos Prompt Police Warning to Parents

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

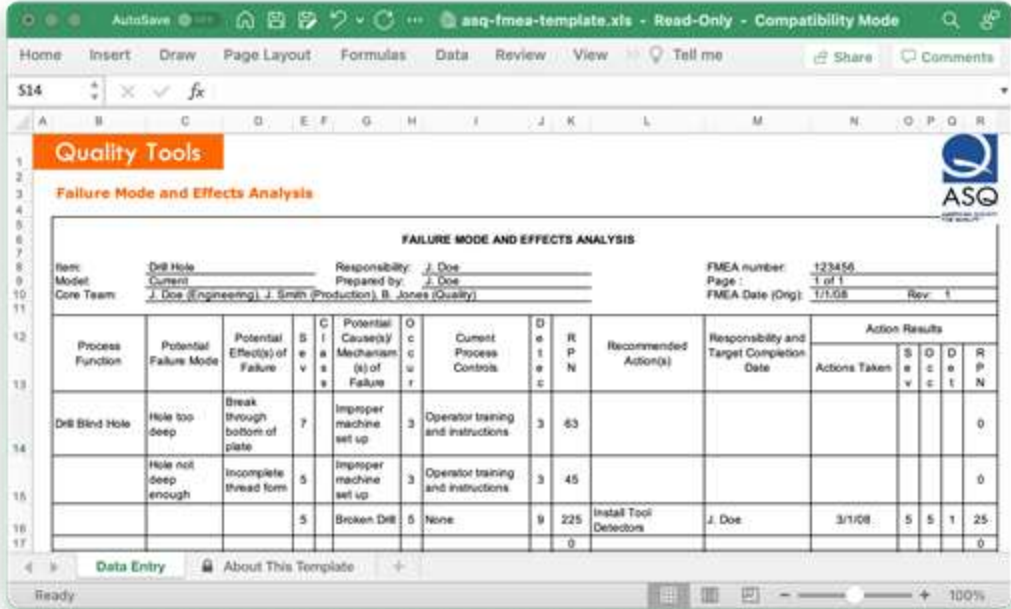
HOME > TECH

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

Zak Burnard Nov 8, 2015, 9:19 PM

Hazard Analysis

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



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?



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

David J. Provan^a, Andrew J. Rae, Sidney W.A. Dekker

^a 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 promise organisations, there is limited research about the current state of safety professional practice

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



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

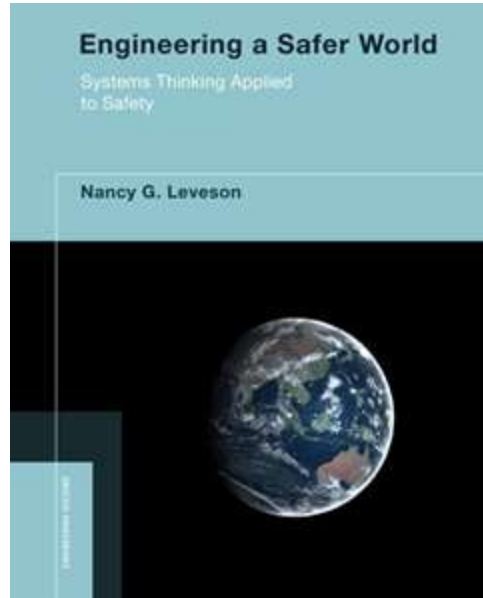
ARTICLE INFO

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 risker within the boundaries of existing theory, and

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



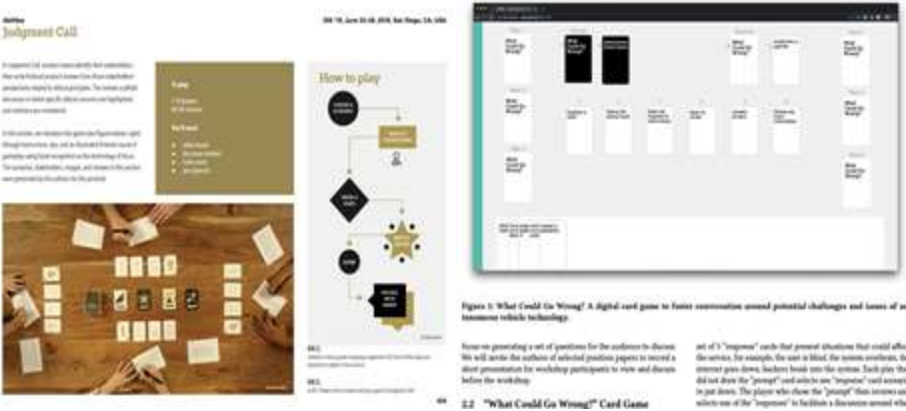
Engineering a Safe World: Systems Thinking Applied to Safety (2012) Nancy G. Leveson

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.



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

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

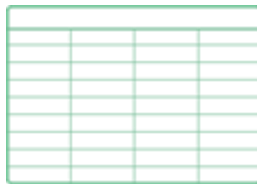
Initial Findings



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

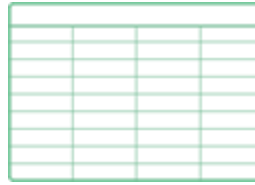
Initial Findings



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



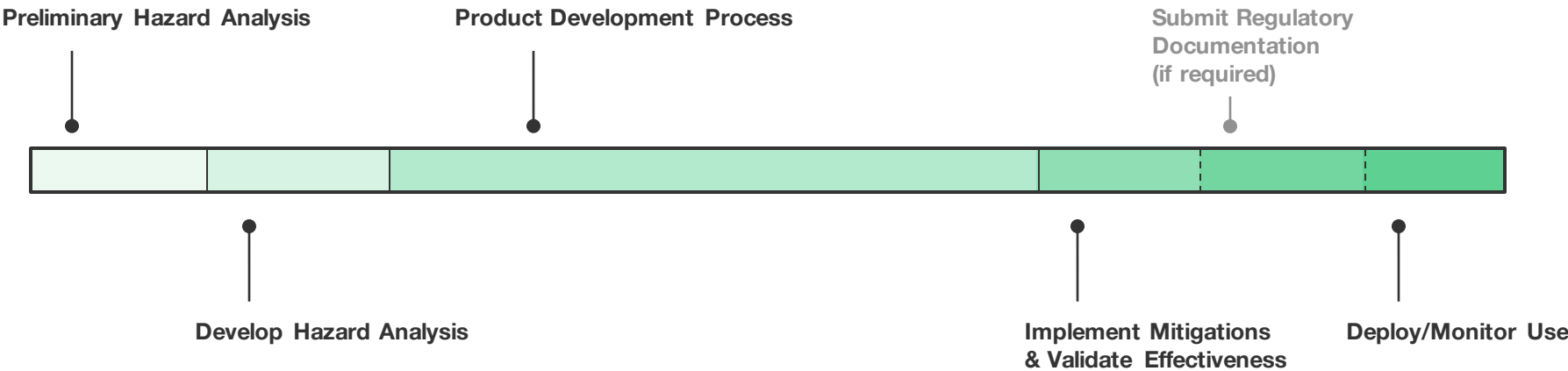
Role of **company culture** in the support of these efforts

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?**



Preliminary Hazard Analysis



What is done?

- Group brainstorming/ 'whiteboarding' sessions involving multiple disciplines
- 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)

Develop Hazard Analysis



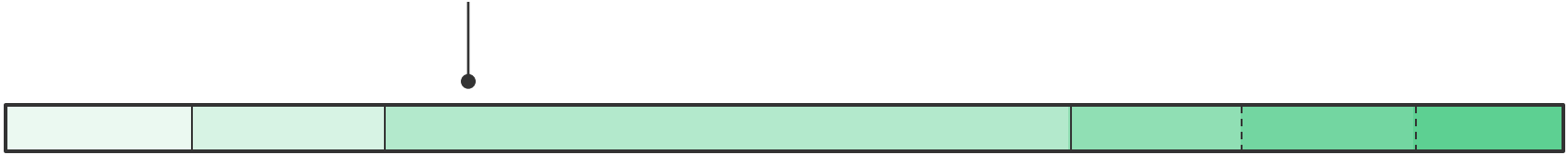
What is done?

- 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

Product Development Process



What is done?

- Iterative product development process
- Continuous changes to individual systems that may be tightly coupled with other systems overseen by different disciplines
- 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**

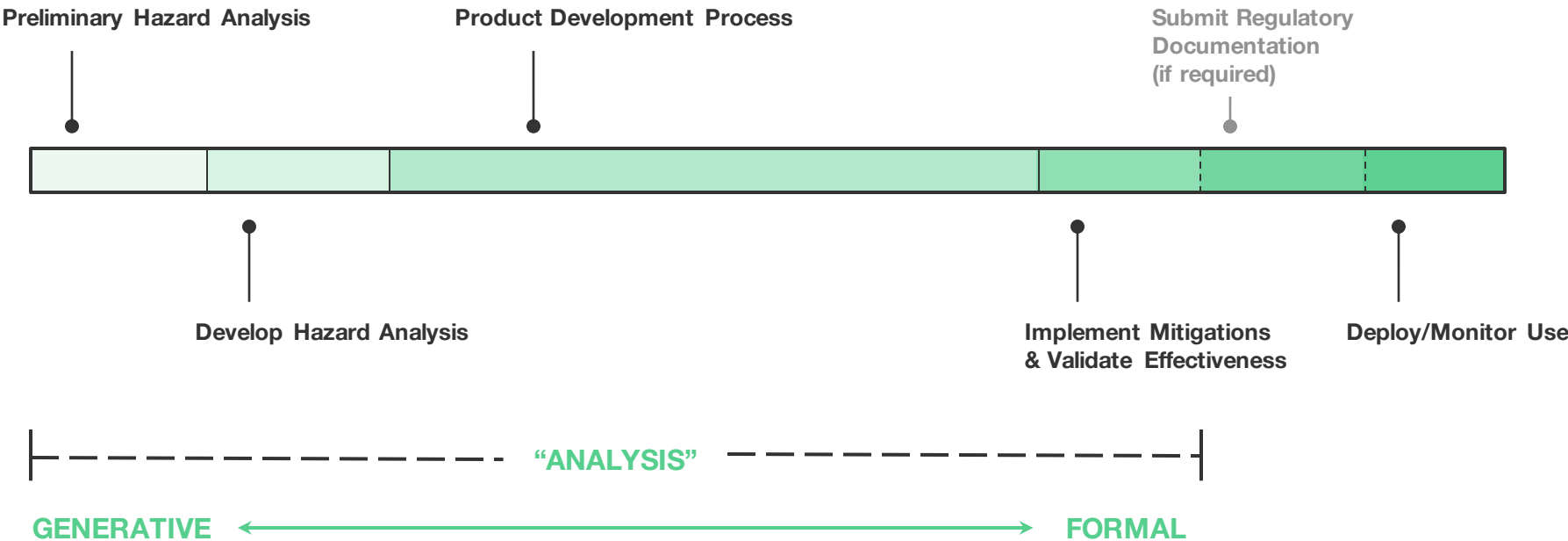


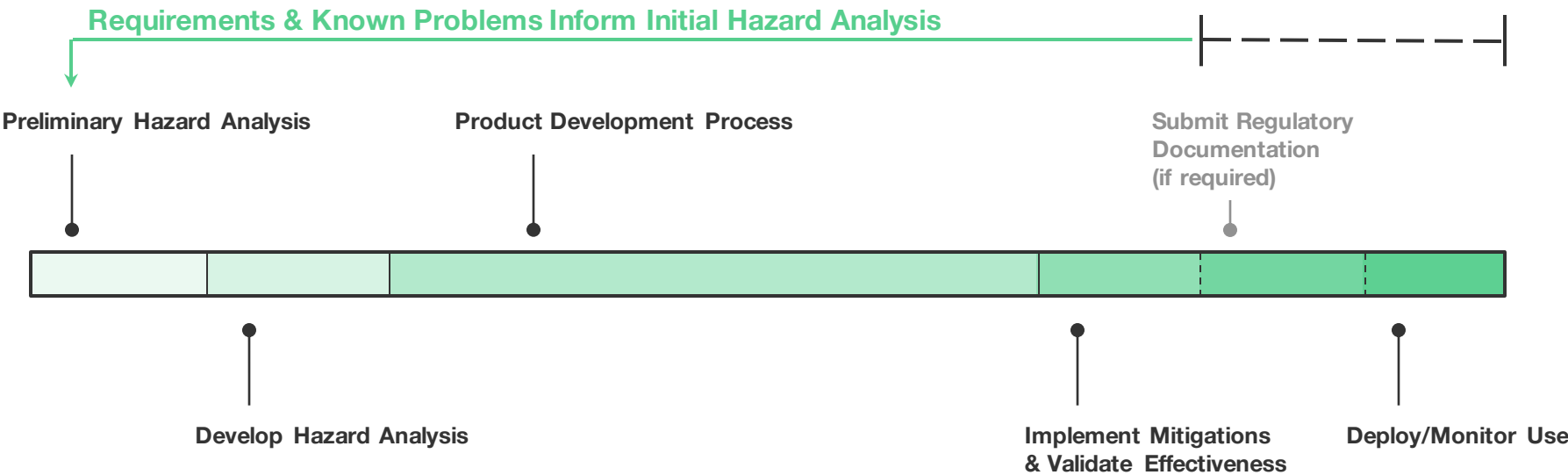
What is done?

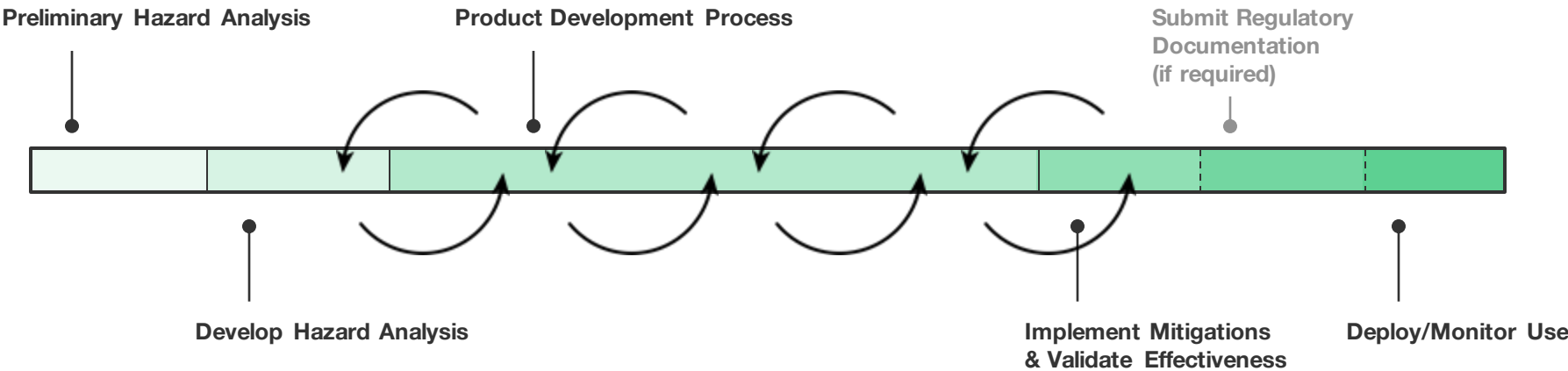
- Implementing mitigations and validating the effectiveness of those mitigations typically occurs at later stages of product development
- If any residual risk remains, might have to do a residual risk assessment after
- Auditing and sub-auditing reports

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





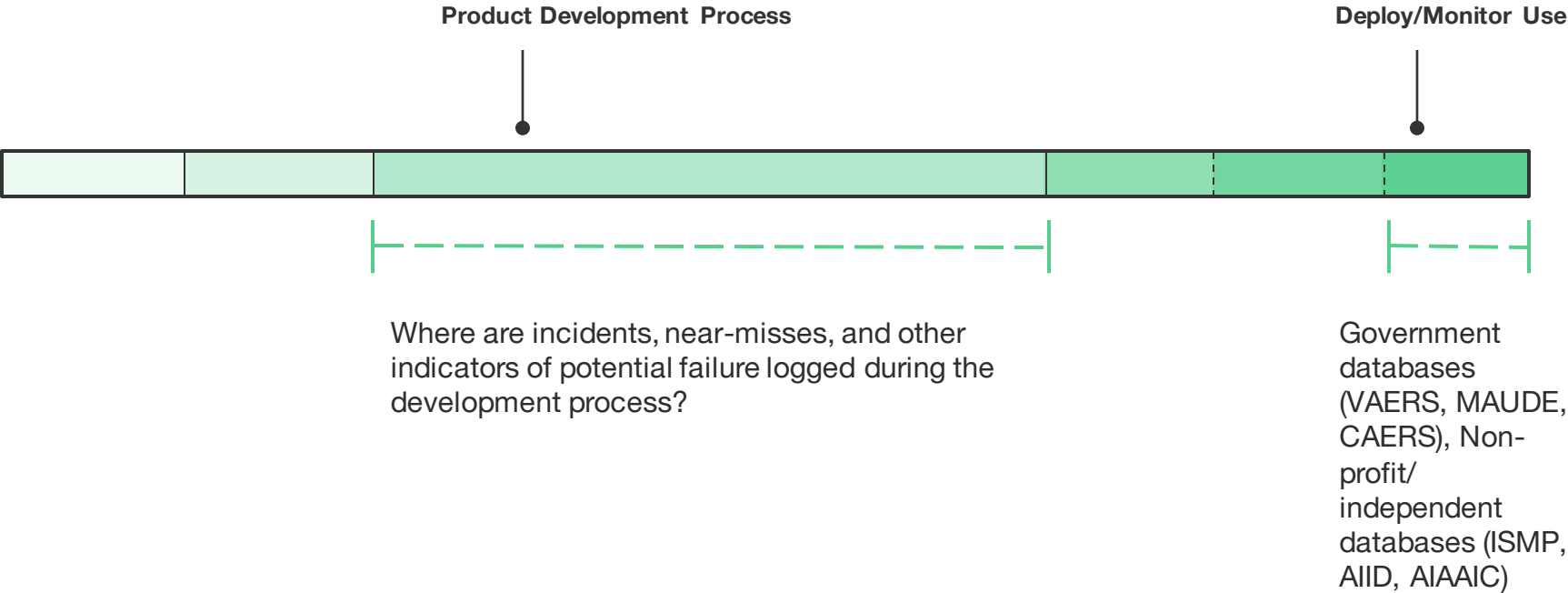


FOR THOSE IN INDUSTRY

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?



FOR THOSE IN INDUSTRY

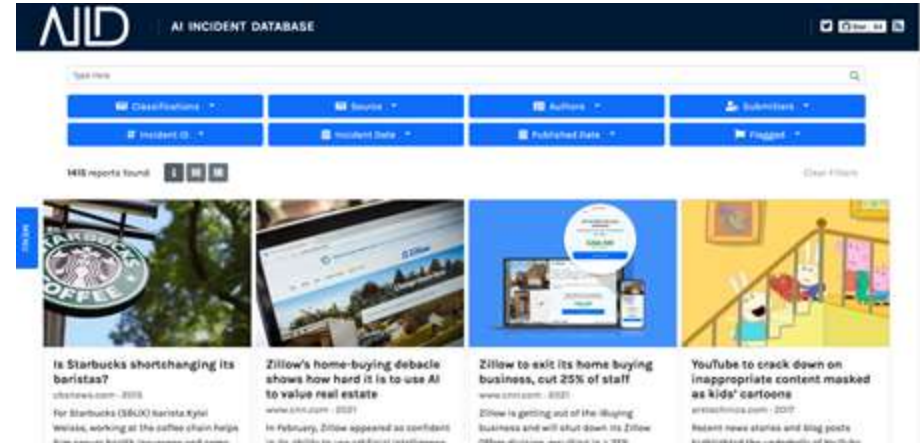
How does your organization **document incidents/harms** internally and/or externally?

FOR THOSE IN ACADEMIA

Do you **teach any documentation techniques** in your classes?

“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>



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>

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
2022 Software Engineering Institute, Software Engineering Institute, Carnegie Mellon University
https://www.sei.cmu.edu/hazards-ai-eng/

Abstract
Identifying, analyzing, and mitigating hazards in machine learning systems is a complex task. This paper explores the challenges of hazard analysis in machine learning systems and presents a framework for identifying and mitigating hazards in machine learning systems. The framework consists of three main components: hazard identification, hazard analysis, and hazard mitigation. The framework is designed to be used by machine learning engineers and system designers to identify and mitigate hazards in machine learning systems. The framework is designed to be used by machine learning engineers and system designers to identify and mitigate hazards in machine learning systems.



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