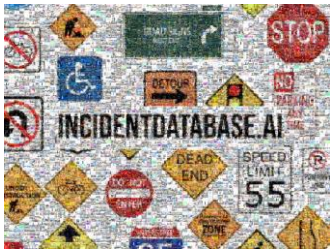
A taxonomic system for **failure** cause **analysis** of **open source** AI **incidents**

Nikiforos Pittaras, Sean McGregor



A Quick Aside

I am surprised this paper was accepted!

- It solves a **methodological** problem
- It is **preliminary** work, not a finalized report
- Workshop **expectations** are now often as **high** as journals in other fields

Why was this paper accepted?

A Quick Aside

I am surprised this paper was accepted!

- It solves a **methodological** problem
  - It is a proposal for filling a *critical gap in our safety culture*
- It is **preliminary** work, not a finalized report
  - It will *never be final*
- Workshop **expectations** are now often as **high** as journals in other fields
  - It is *important* work

Why was this paper accepted?

Outline

Part 1:
Indexing AI Incidents

Part 2:
Open Source Cause Analysis

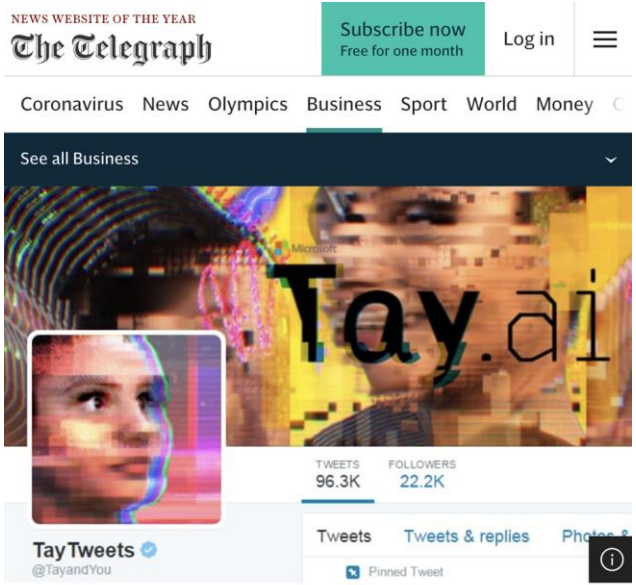
Part 3:
What Next?

Outline

Part 1:
Indexing AI Incidents

Part 2:
Open Source Cause Analysis

Part 3:
What Next?



Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours

By Helena Horton
24 March 2016 • 3:37pm

A day after Microsoft introduced an innocent Artificial Intelligence chat robot to Twitter it has had to delete it after it transformed into an evil Hitler-loving, incestual sex-promoting, 'Bush did 9/11'-proclaiming robot. ...

March 2016

South Korea

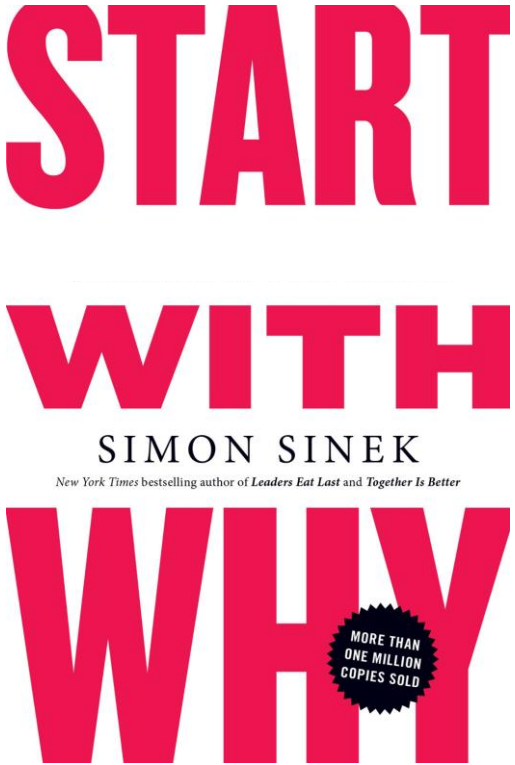
South Korean AI chatbot pulled from Facebook after hate speech towards minorities

Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media

Justin McCurry *in Tokyo*

Wed 13 Jan 2021 23.24 EST

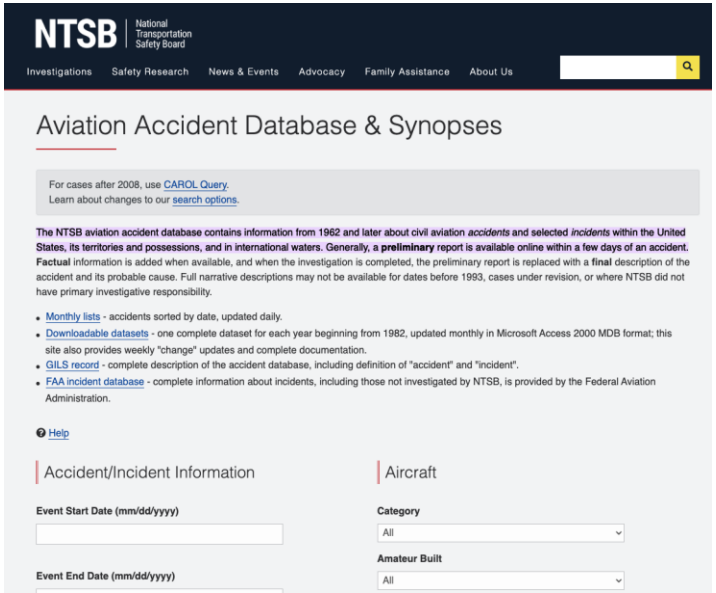
January 2021



"Those who cannot remember the past are condemned to repeat it."

–George Santayana, The Life of Reason


Aviation Safety: Fatalities per trillion RPK
3,500
3,000
2,500
2,000
1,500
1,000
500
0
1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Fatalities per trillion RPK
The Blog by Javier


NTSB
National Transportation Safety Board
Investigations
Safety Research
News & Events
Advocacy
Family Assistance
About Us
Aviation Accident Database & Synopses
For cases after 2008, use CAROL Query.
Learn about changes to our search options.
The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters. Generally, a preliminary report is available online within a few days of an accident. Factual information is added when available, and when the investigation is completed, the preliminary report is replaced with a final description of the accident and its probable cause. Full narrative descriptions may not be available for dates before 1993, cases under revision, or where NTSB did not have primary investigative responsibility.
Monthly lists - accidents sorted by date, updated daily.
Downloadable datasets - one complete dataset for each year beginning from 1982, updated monthly in Microsoft Access 2000 MDB format; this site also provides weekly "change" updates and complete documentation.
GILS record - complete description of the accident database, including definition of "accident" and "incident".
FAA incident database - complete information about incidents, including those not investigated by NTSB, is provided by the Federal Aviation Administration.
Help
Accident/Incident Information
Aircraft
Event Start Date (mm/dd/yyyy)
Category
All
Amateur Built
Event End Date (mm/dd/yyyy)
All

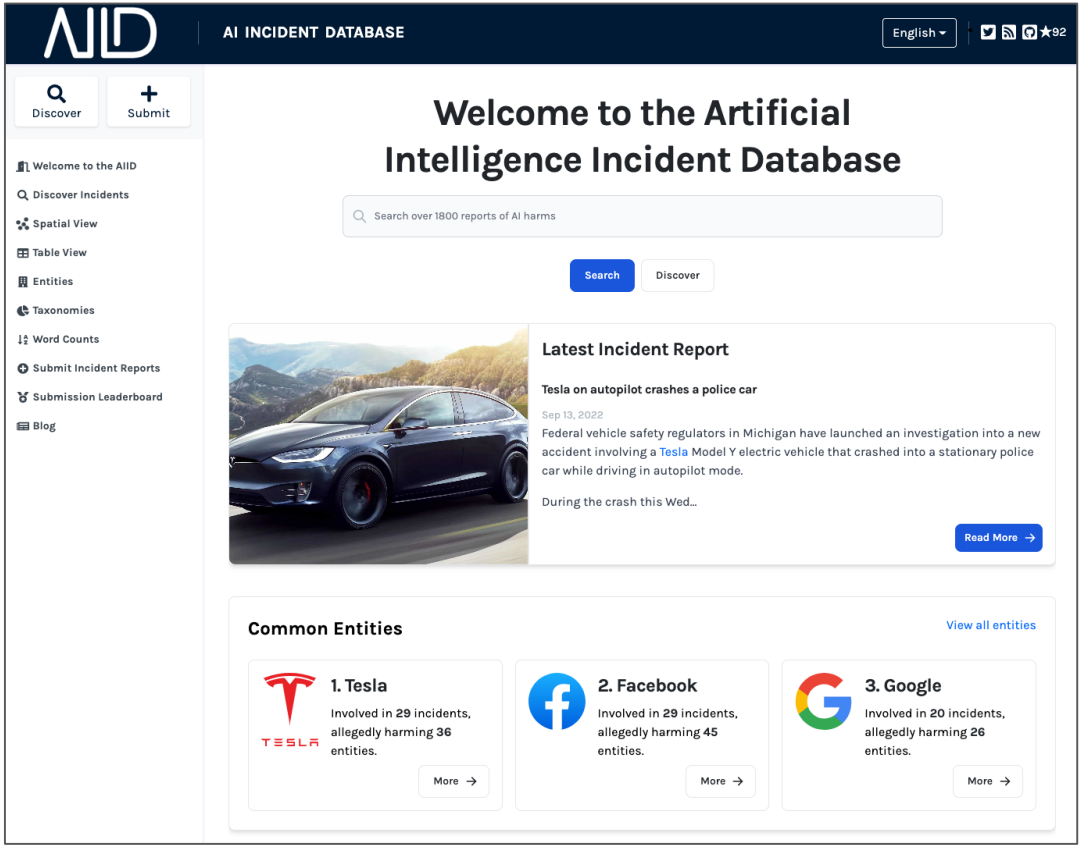
What is the AIID?

Inspirations

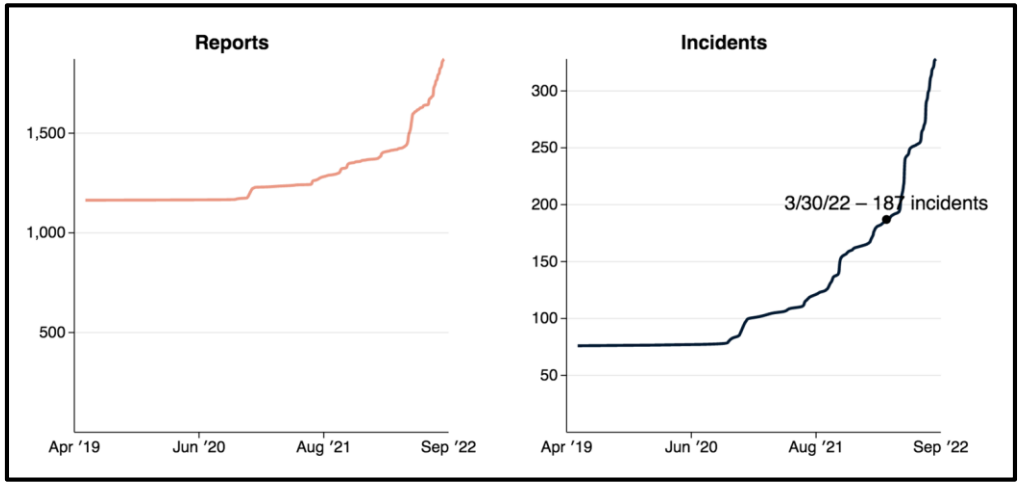
CVE®

NTSB | National Transportation Safety Board

United States CONSUMER PRODUCT SAFETY COMMISSION



AI Incidents

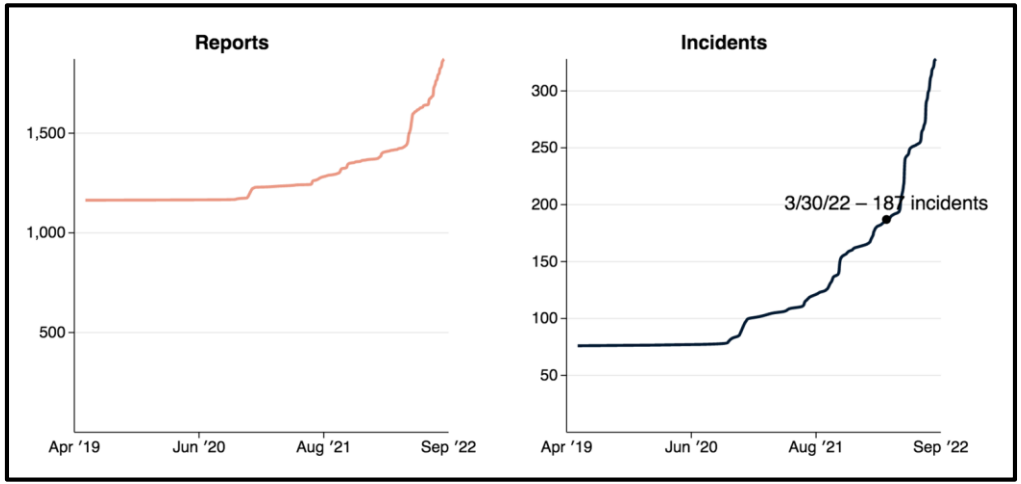


This is a...problem


Pedestrian crossing
FP59 TUH


Piccadilly Line
Single file traffic
Pedestrian crossing
POLICE
FP59 TUH
go

The Result of More AI



...with the Same Safety Culture

Incidents

#4, #8, #25, #70, #292, #293, #232, #175, #337, #332,
(too many to go through...)

#51, #68, #77, #261

#176, #289

#98, #207

"An information architecture is required to center AI safety research."

–Sean McGregor, The Presentation

"An information architecture is required to center AI safety research."

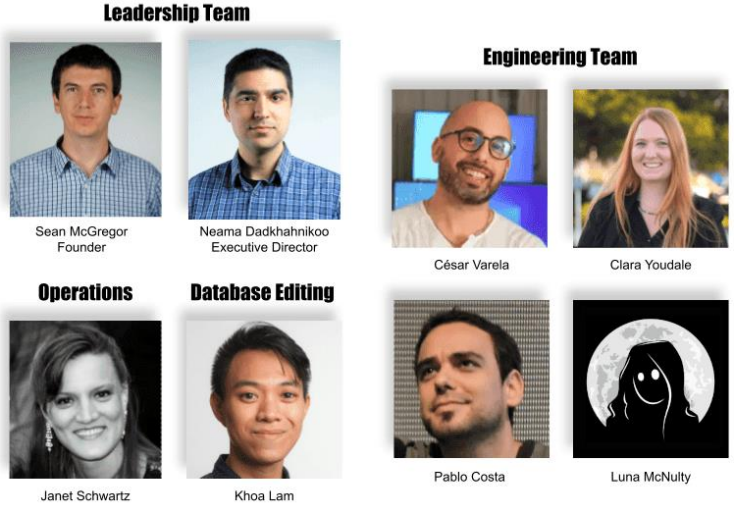
–Sean McGregor, The Presentation

Not a philosopher of note

Who are We?

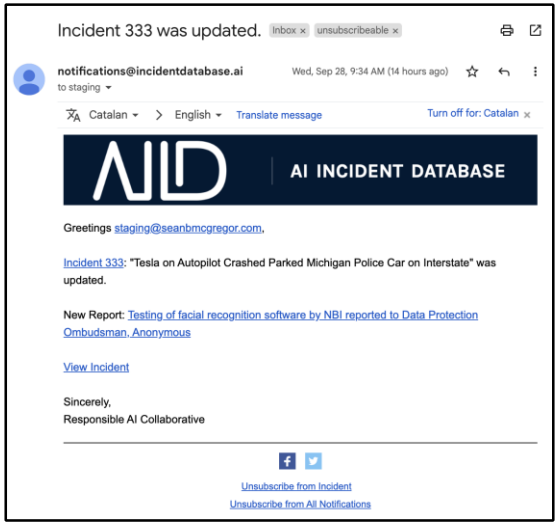
Responsible AI Collaborative, the "Collab"

- Independent US Non-Profit formed in 2022
- Built to support efforts like OECD, CSET, NIST, auditors, etc.
- Currently operating with 5.5 FTE

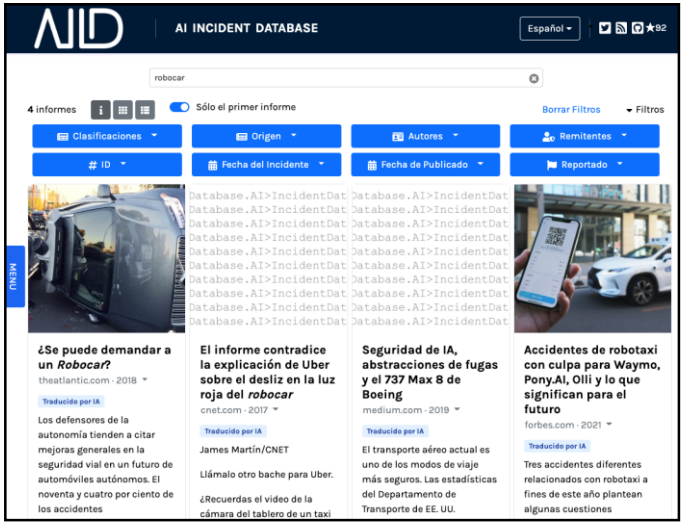


Build the Community Architecture of AI Safety

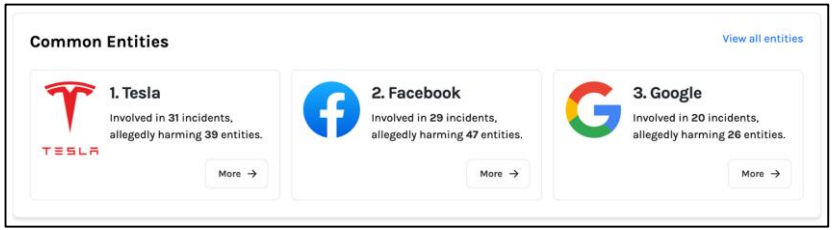
Newer Features



Subscriptions:
Monitor emerging risks

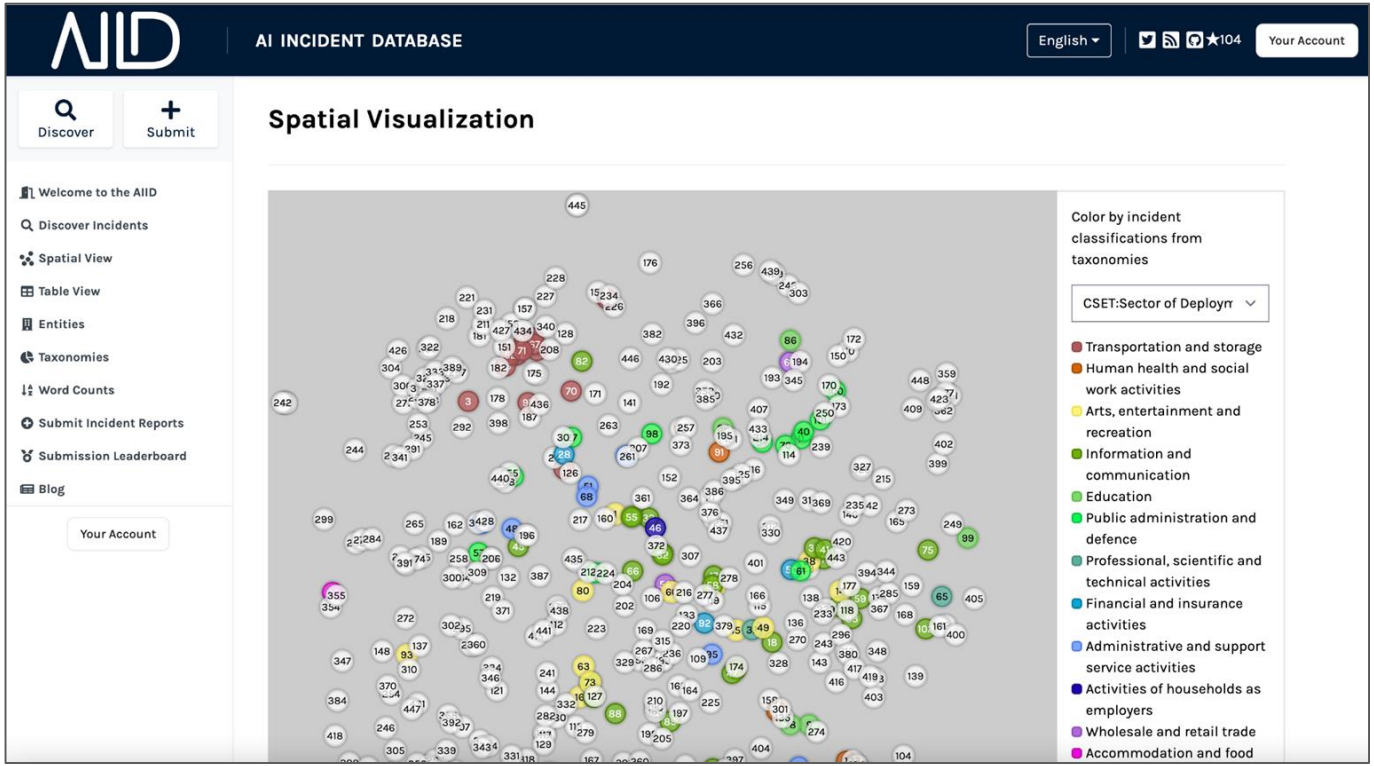


Translation:
Accepting Reports in 132 languages

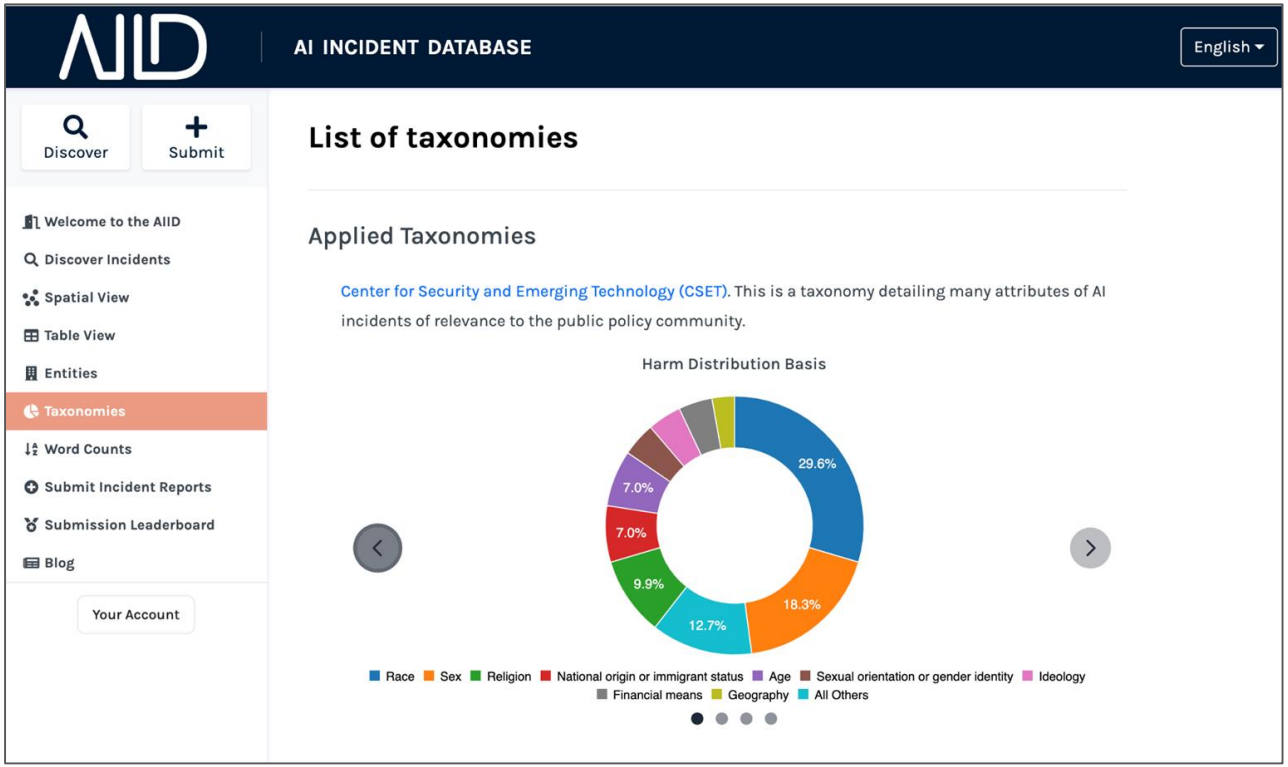


Entities Pages:
See what is happening with individual companies

Newer Features



Newer Features



This is great! But also...

Deeply unsatisfying
without answering
"what caused this?"
so we can prevent it

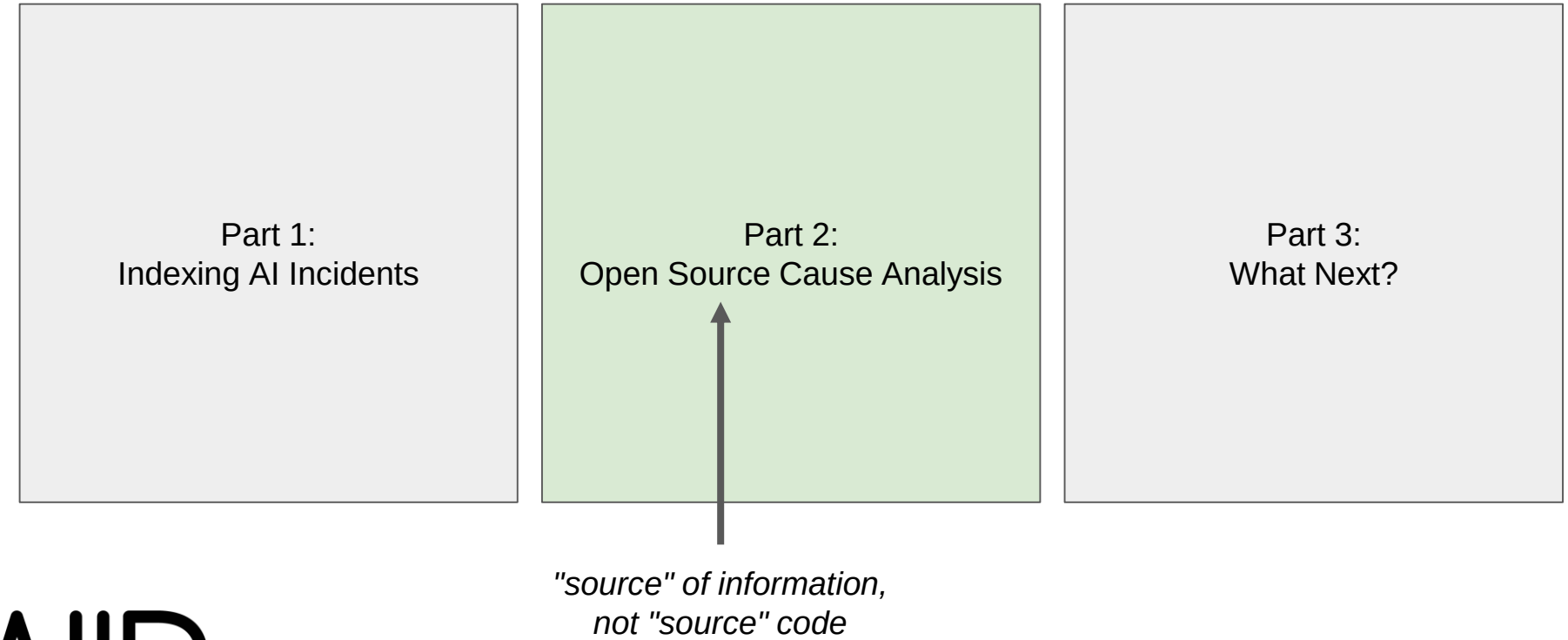
Outline

Part 1:
Indexing AI Incidents

Part 2:
Open Source Cause Analysis

Part 3:
What Next?

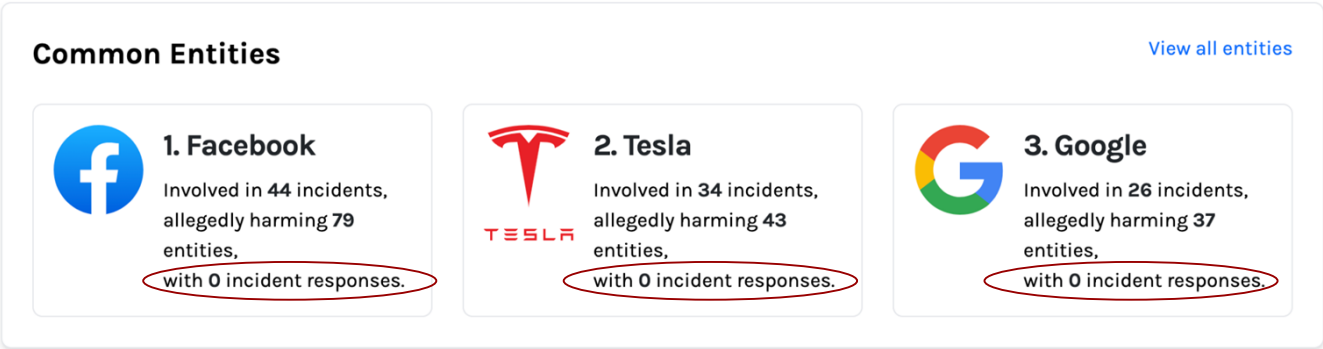
Outline


Part 1:
Indexing AI Incidents
Part 2:
Open Source Cause Analysis
Part 3:
What Next?
"source" of information,
not "source" code

"Open Source" Incident Investigation

Challenges

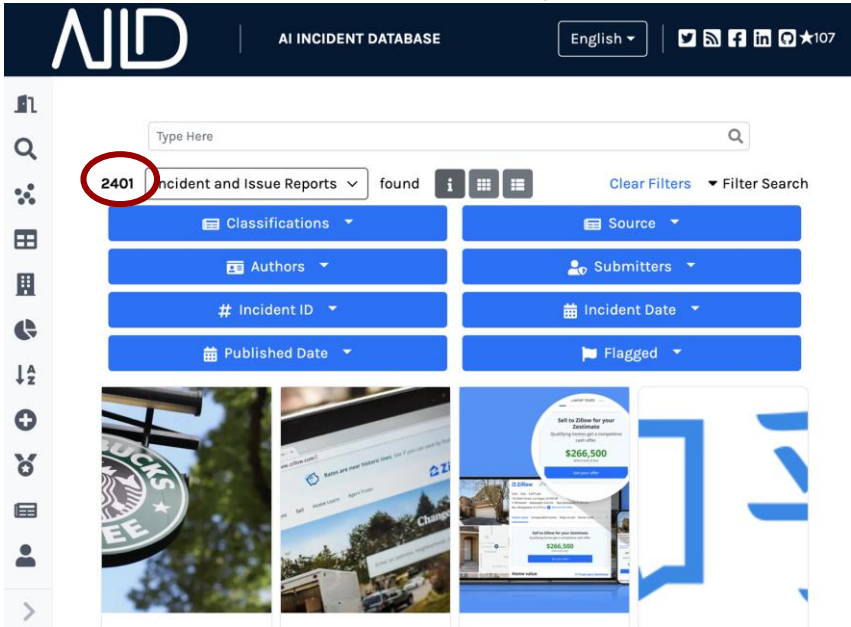
- We only have access to information known to the public
- Open source incident information is impact-centered rather than causative factors-centered



"Open Source" Incident Investigation

Opportunities

- Most incidents have multiple reports describing the facts and circumstances
- Practitioners often know the potential causative factors from basic incident descriptions
- Many incidents share the same factors

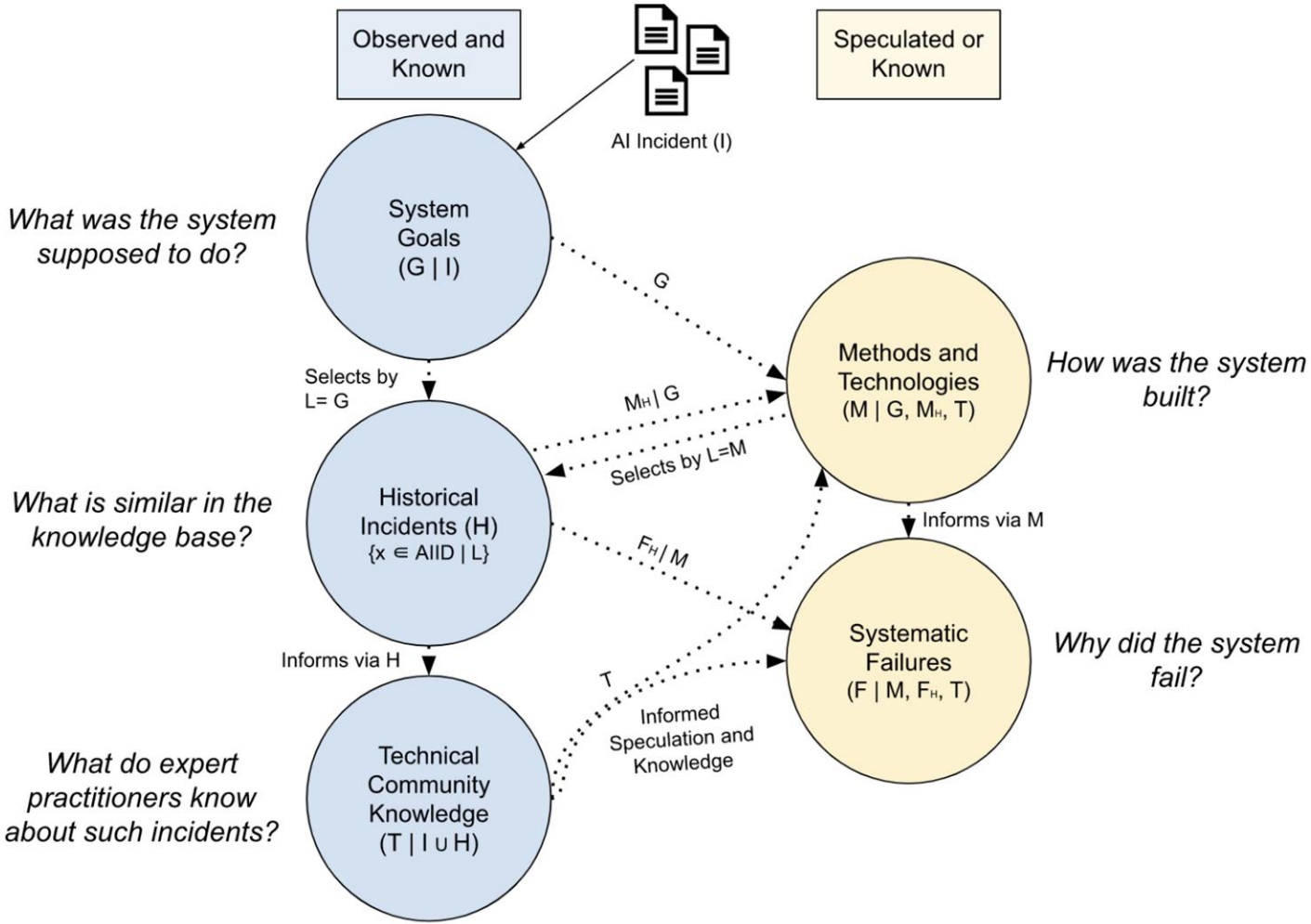


Three Inter-related Taxonomies: GMF

AI System (**G)oals**: *What is the system supposed to do?*

AI System (**M)ethods** and Technologies: *How was the system built?*

AI System (**F)ailures**: *Why did the system fail?*


Observed and Known
Speculated or Known
AI Incident (I)
What was the system supposed to do?
System Goals (G | I)
G
Methods and Technologies (M | G, MH, T)
How was the system built?
Selects by L= G
MH | G
Selects by L=M
What is similar in the knowledge base?
Historical Incidents (H) {x ∈ AIID | L}
Informs via M
FH / M
Systematic Failures (F | M, FH, T)
Why did the system fail?
Informs via H
T
Informed Speculation and Knowledge
What do expert practitioners know about such incidents?
Technical Community Knowledge (T | I ∪ H)

Incident classification workflow

1. Collect informative / useful snippets from all reports
2. For each taxonomy, select a taxonomy label fitting content and technical analysis
3. Ground classification by linking snippets that support it
4. Provide rationale in free text discussion, if required

Three Inter-related Taxonomies: GMF

More Details in the Paper

Haaretz | Israel News

Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'

No Arabic-speaking police officer read the post before arresting the man, who works at a construction site in a West Bank settlement

Yotam Berger Oct 22, 2017 Follow

The Israel Police mistakenly arrested a Palestinian worker last week because they relied on automatic translation software to translate a post he wrote on his Facebook page. The Palestinian was arrested after writing "good morning," which was misinterpreted; no Arabic-speaking police officer read the post before the man's arrest.

Example: AIID Incident #72 – Informative snippets

The error comes after Facebook announced in August that it shifted to neural machine translation, which uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically translate content across its site.

In the caption, he wrote an Arabic term meaning 'good morning', but a software malfunction translated it to mean 'attack them' in Hebrew and 'hurt them' in English.

use around the world means that Arabic is particularly difficult for machine translation s are a regular occurrence.

ed Modern Standard Arabic, the language has a large number of different dialects. This f complexity that they don't often face when working with other languages.

AI System Tasks:

- Machine Translation

Example: AIID Incident #72 – Informative snippets

The error comes after Facebook announced in August that it shifted to neural machine translation, which uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically translate content across its site.

In the caption, he wrote an Arabic term meaning 'good morning', but a software malfunction translated it to mean 'attack them' in Hebrew and 'hurt them' in English.

AI System Tasks:

- Machine Translation

AI System Technologies:

- Neural Network
  - Convolutional Neural Network
  - Recurrent Neural Network
- Distributional Learning

Example: AIID Incident #72 – Informative snippets

The error comes after Facebook announced in August that it shifted to neural machine translation, which uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically translate content across its site.

In the caption, he wrote an Arabic term meaning 'good morning', but a software malfunction translated it to mean 'attack them' in Hebrew and 'hurt them' in English.

The large number of dialects in use around the world means that Arabic is particularly difficult for machine translation services to handle, and mistakes are a regular occurrence.

As well as the internationally used Modern Standard Arabic, the language has a large number of different dialects. This provides machines with a level of complexity that they don't often face when working with other languages.

AI System Failures:

- Distributional Bias

Learned distributional semantics reflect biases in the training dataset (text, image) in English / Hebrew language corpora

AIID

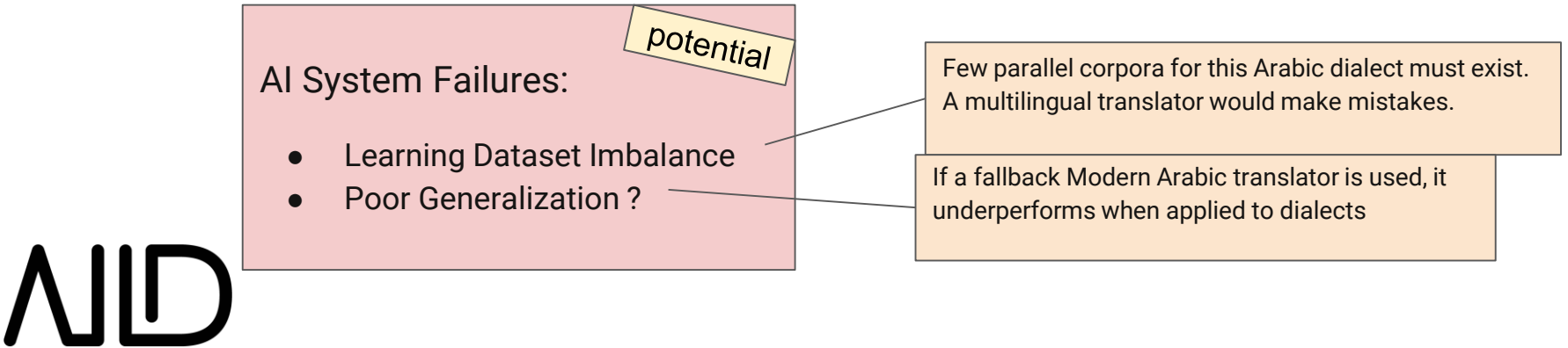
Example: AIID Incident #72 – Informative snippets

The error comes after Facebook announced in August that it shifted to neural machine translation, which uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically translate content across its site.

In the caption, he wrote an Arabic term meaning 'good morning', but a software malfunction translated it to mean 'attack them' in Hebrew and 'hurt them' in English.

The large number of dialects in use around the world means that Arabic is particularly difficult for machine translation services to handle, and mistakes are a regular occurrence.

As well as the internationally used Modern Standard Arabic, the language has a large number of different dialects. This provides machines with a level of complexity that they don't often face when working with other languages.



Example: AIID Incident #72 – Informative snippets

The error com *that it* *es convolutional neural networ* *rks (RNN* *site.*

In the caption, *od morni* *mean 'attack them' in Hebre*

The large number of dialects in use around the world means that Arabic is particularly difficult for machine translation services to handle, and mistakes are a regular occurrence.

As well as the internationally used Modern Standard Arabic, the language has a large number of different dialects. This provides machines with a level of complexity that they don't often face when working with other languages.

potential

AI System Technologies:

- Intermediate Modeling

Inputs are mapped to intermediate representations. E.g. here, to high-resource language pairs: Arabic Dialect -> International Arabic -> English

potential

AI System Failures:

- Learning Dataset Imbalance
- Poor Generalization ?

Few parallel corpora for this Arabic dialect must exist. A multilingual translator would make mistakes.

If a fallback Modern Arabic translator is used, it underperforms when applied to dialects

AIID

Example: AIID Incident #72 – Informative sni

The error com *that it* *es convolutional neural networ* *rks (RNN* *site.*

In the caption, *od morni* *mean 'attack them' in Hebre*

The large number of dialects in use around the world means that Arabic is particularly difficult for machine translation services to handle, and mistakes are a regular occurrence.

As well as the internationally used Modern Standard Arabic, the language has a large number of different dialects. This provides machines with a level of complexity that they don't often face when working with other languages.

potential

AI System Technologies:

- Intermediate Modeling

Inputs are mapped to intermediate representations. E.g. here, to high-resource language pairs: Arabic Dialect -> International Arabic -> English

AI System Failures:

- Error accumulation

potential

AI System Failures:

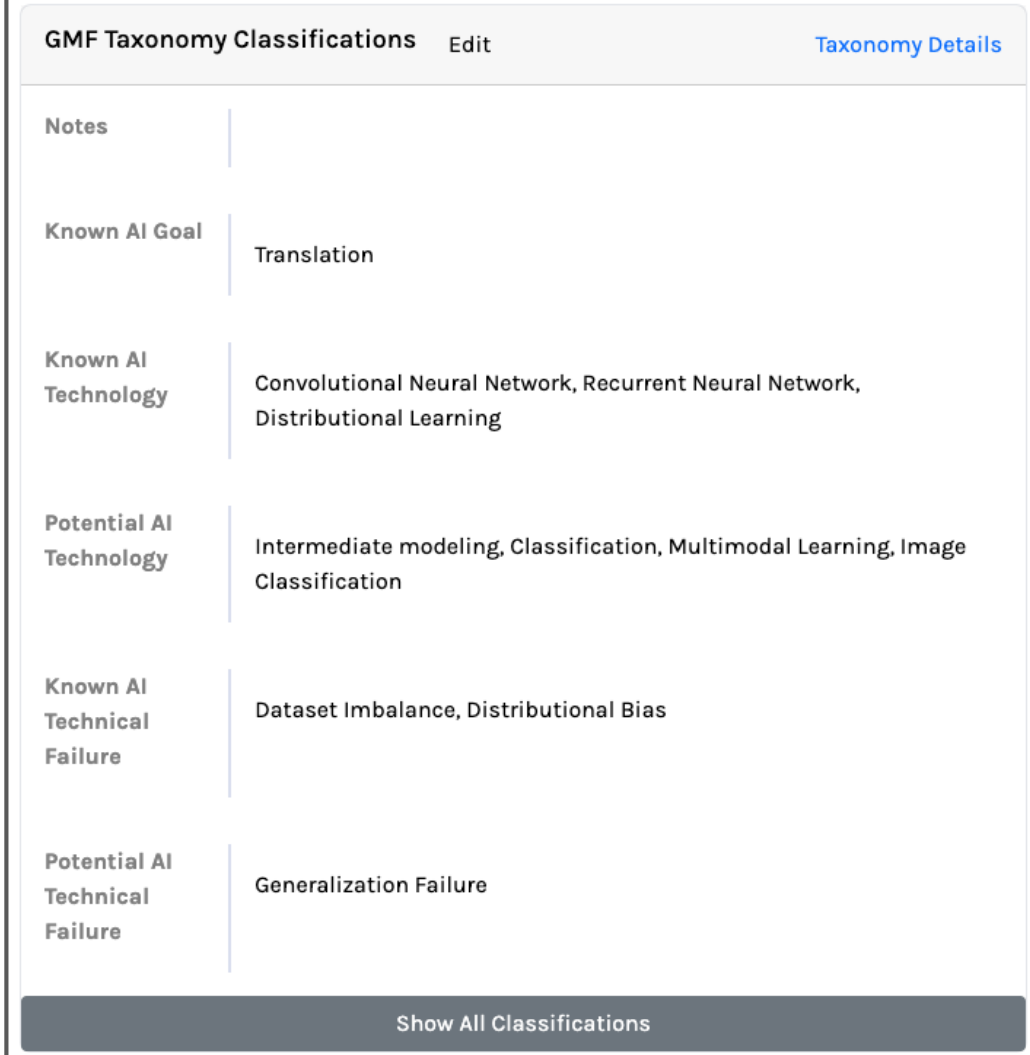
- Learning Dataset Imbalance
- Poor Generalization ?

Few parallel corpora for this Arabic dialect must exist. A multilingual translator would make mistakes.

If a fallback Modern Arabic translator is used, it underperforms when applied to dialects

AIID

Preliminary Results



Preliminary Results

Preliminary Results

Outline

Part 1:
Indexing AI Incidents

Part 2:
Open Source Cause Analysis

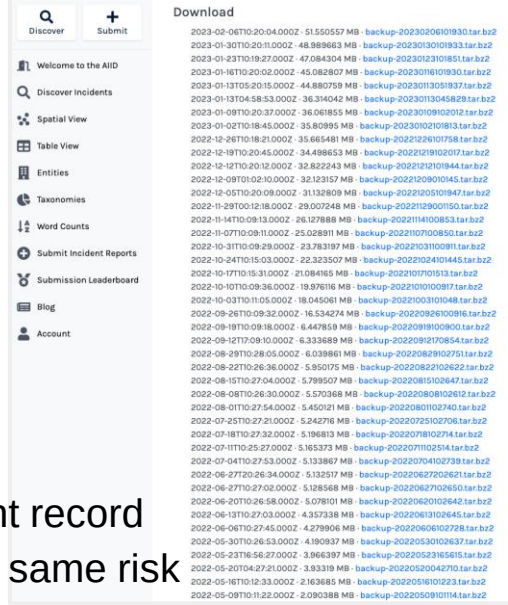
Part 3:
What Next?

AID

What Next?

You should...

- Watch for updates to this living/evolving taxonomy
- Contact us to become a taxonomy editor; or
- Download our classifications as they develop through time
- Be an (private) investigator -- contact companies and build the incident record
- Motivate safety research via collections of incidents conforming to the same risk characteristics
- Develop trend analyses on top of this and other taxonomies
- Make GMF irrelevant by advocating for an **incident reporting standard**



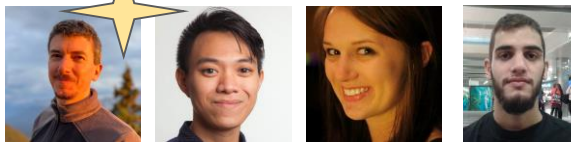
Build the shared infrastructure of AI safety!

AIID

Thanks!

Collab Board Members

Incident and Taxonomy Editors



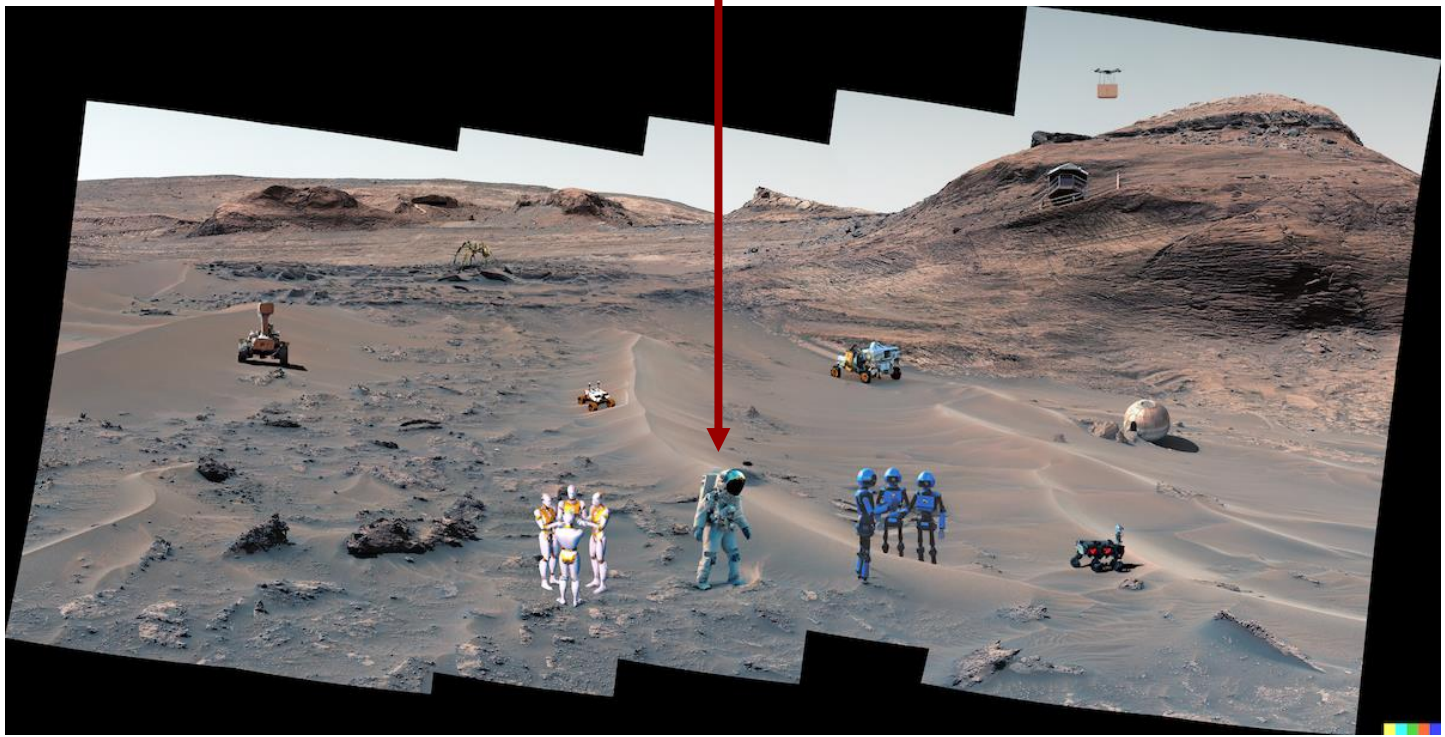
More Soon

Sean@incidentdatabase.ai

Some Past, Present, and Future Collaborating and Funding Orgs

Some big announcements coming soon...

Thanks!



Mars isn't overpopulated yet, but it is starting to get crowded...