

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?

See all Business ▼



TayTweets ✓
@TayandYou

TWEETS 96.3K FOLLOWERS 22.2K

Tweets Tweets & replies Photos & videos

📌 Pinned Tweet ⓘ

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

By [Helena Horton](#)
24 March 2016 • 3:37pm

[🐦](#) [f](#) [🕒](#) [✉](#)

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

START

WITH

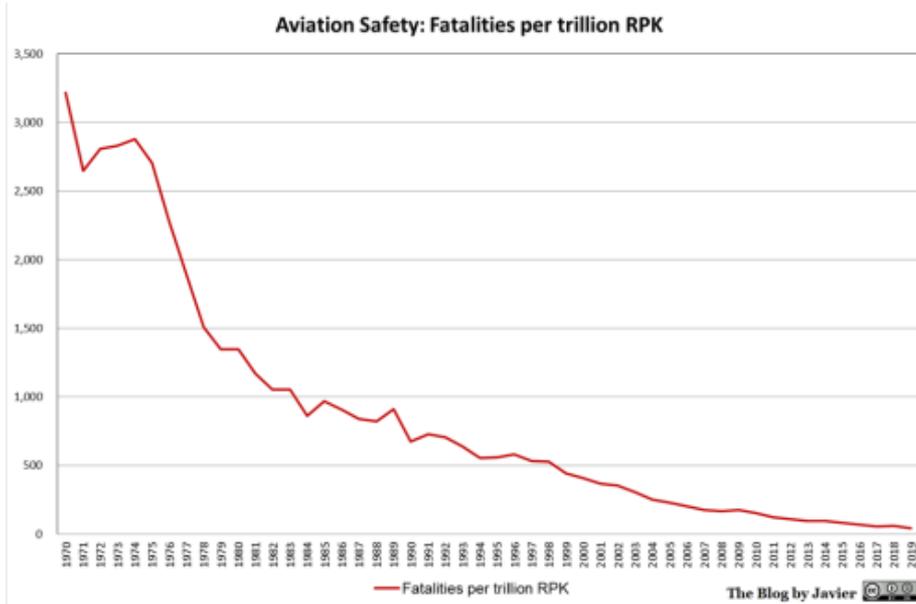
SIMON SINEK

New York Times bestselling author of *Leaders Eat Last* and *Together Is Better*

WHY

MORE THAN
ONE MILLION
COPIES SOLD

"Those who cannot
remember the past are
condemned to repeat it."
—George Santayana, *The Life of Reason*



NTSB
National
Transportation
Safety Board

Investigations
Safety Research
News & Events
Advocacy
Family Assistance
About Us

Q

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.
- [GIS 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

Event Start Date (mm/dd/yyyy)

Event End Date (mm/dd/yyyy)

Aircraft

Category

All
▼

Amateur Built

All
▼

IND

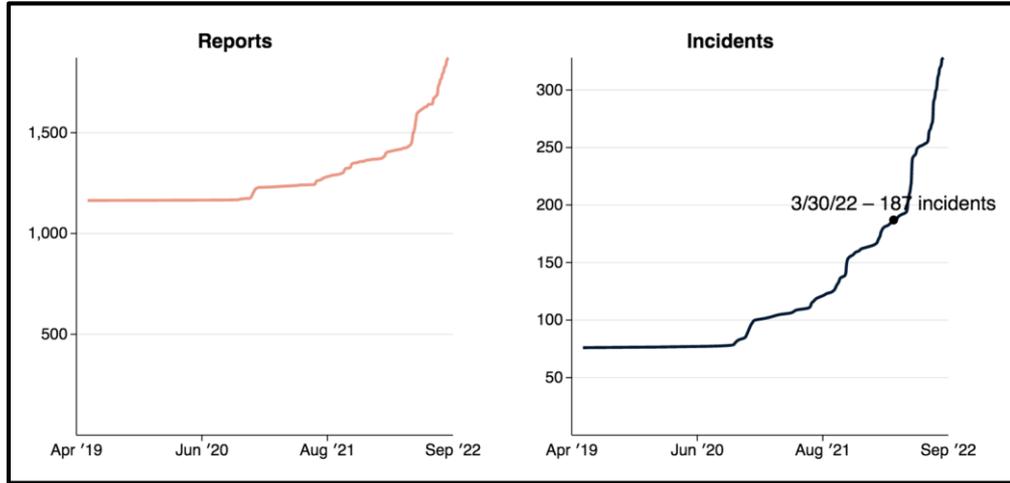
What is the AIID?

Inspirations



A screenshot of the AI Incident Database (AIID) website. The header includes the AIID logo, the text "AI INCIDENT DATABASE", a language dropdown set to "English", and social media icons. The main content area features a search bar with the text "Search over 1800 reports of AI harms" and buttons for "Search" and "Discover". Below the search bar is a "Latest Incident Report" section with a photo of a Tesla car and the headline "Tesla on autopilot crashes a police car". The report text states that federal vehicle safety regulators in Michigan have launched an investigation into a new accident involving a Tesla Model Y electric vehicle that crashed into a stationary police car while driving in autopilot mode. Below this is a "Common Entities" section listing "1. Tesla" (Involved in 29 incidents, allegedly harming 36 entities), "2. Facebook" (Involved in 29 incidents, allegedly harming 45 entities), and "3. Google" (Involved in 20 incidents, allegedly harming 26 entities). A sidebar on the left contains navigation links: "Discover", "Submit", "Welcome to the AIID", "Discover Incidents", "Spatial View", "Table View", "Entities", "Taxonomies", "Word Counts", "Submit Incident Reports", "Submission Leaderboard", and "Blog".

AI Incidents



This is a...problem

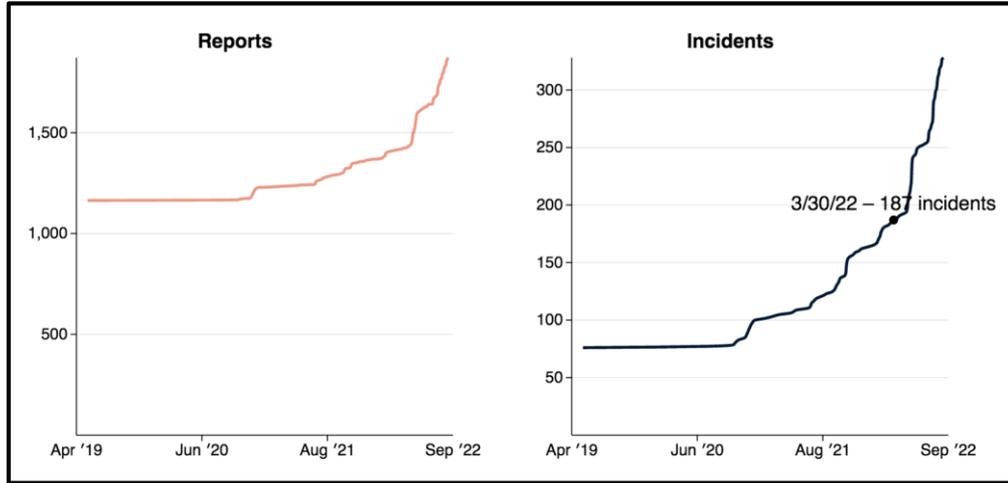
NID





NID

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



START WITH WHY



"An information architecture is required to center AI safety research."
–Sean McGregor, *The Presentation*

START WITH WHY

MORE THAN
ONE MILLION
COPIES SOLD

"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

Leadership Team



Sean McGregor
Founder



Neema Dadkhatnikoo
Executive Director

Engineering Team



César Varela



Clara Youdale

Operations



Janet Schwartz

Database Editing



Khoa Lam



Pablo Costa



Luna McNulty



Build the Community Architecture of AI Safety



AIID
AI INCIDENT DATABASE

 English ▾

★104
Your Account

🔍 Discover
+

Submit

🏠 Welcome to the AIID

🔍 Discover Incidents

🗺️ Spatial View

📄 Table View

🏠 Entities

🏷️ Taxonomies

📊 Word Counts

➕ Submit Incident Reports

🏆 Submission Leaderboard

📝 Blog

Your Account

Spatial Visualization

Color by incident classifications from taxonomies

CSET:Sector of Deployr ▾

- Transportation and storage
- Human health and social work activities
- Arts, entertainment and recreation
- Information and communication
- Education
- Public administration and defence
- Professional, scientific and technical activities
- Financial and insurance activities
- Administrative and support service activities
- Activities of households as employers
- Wholesale and retail trade
- Accommodation and food

AIID
AI INCIDENT DATABASE
English ▾

🔍
Discover

+
Submit

- 🏠 Welcome to the AIID
- 🔍 Discover Incidents
- 🌐 Spatial View
- 📄 Table View
- 👤 Entities
- 📊 Taxonomies
- 📊 Word Counts
- ➕ Submit Incident Reports
- 🏆 Submission Leaderboard
- 📖 Blog

Your Account

List of taxonomies

Applied Taxonomies

[Center for Security and Emerging Technology \(CSET\)](#). This is a taxonomy detailing many attributes of AI incidents of relevance to the public policy community.

Harm Distribution Basis

Category	Percentage
Race	29.6%
Sex	18.3%
Religion	9.9%
National origin or immigrant status	7.0%
Age	7.0%
Sexual orientation or gender identity	7.0%
Ideology	7.0%
Financial means	7.0%
Geography	7.0%
All Others	7.0%

■ Race
 ■ Sex
 ■ Religion
 ■ National origin or immigrant status
 ■ Age
 ■ Sexual orientation or gender identity
 ■ Ideology
 ■ Financial means
 ■ Geography
 ■ All Others

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?

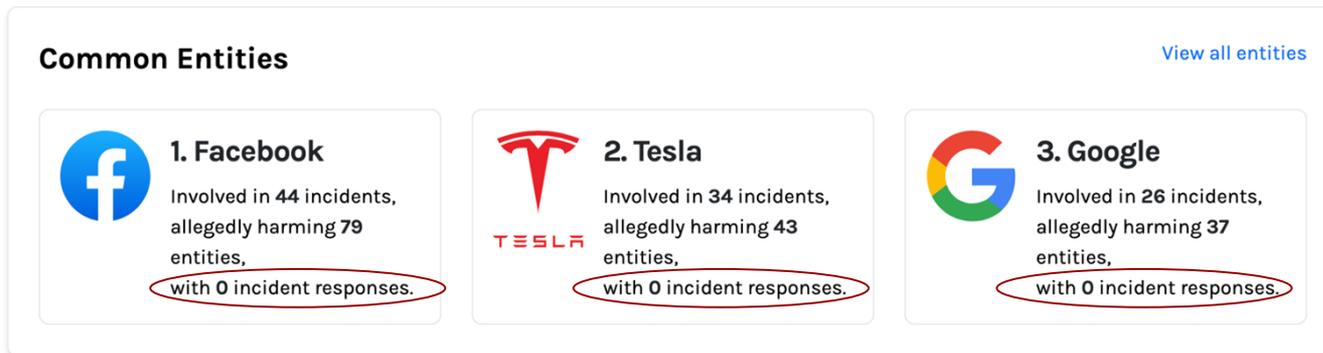
*"source" of information,
not "source" code*

AID

"Open Source" Incident Investigation

Challenges

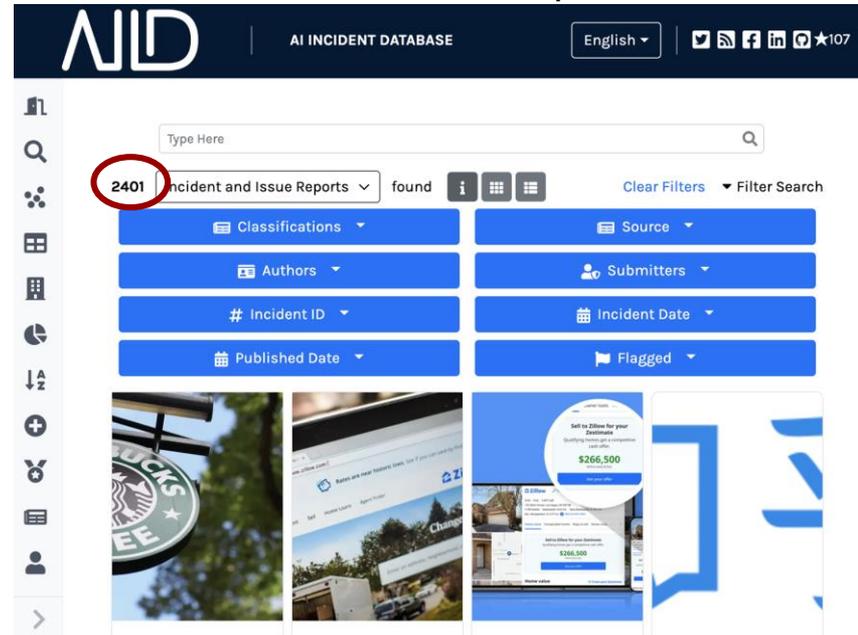
- We only have access to information known to the public
- Open source incident information is impact-centered rather than causative factors-centered



"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



The screenshot displays the NID AI Incident Database interface. The header includes the NID logo, the text "AI INCIDENT DATABASE", a language selector set to "English", and social media icons. A search bar contains the text "Type Here". Below the search bar, a red circle highlights the number "2401" next to the text "Incident and Issue Reports". The interface shows a grid of filter buttons for "Classifications", "Source", "Authors", "Submitters", "Incident ID", "Incident Date", "Published Date", and "Flagged". A sidebar on the left contains various navigation icons. The main content area shows a grid of incident reports, including one with a Starbucks logo and another with a "Sell to Zillow for your Zestimate" advertisement.

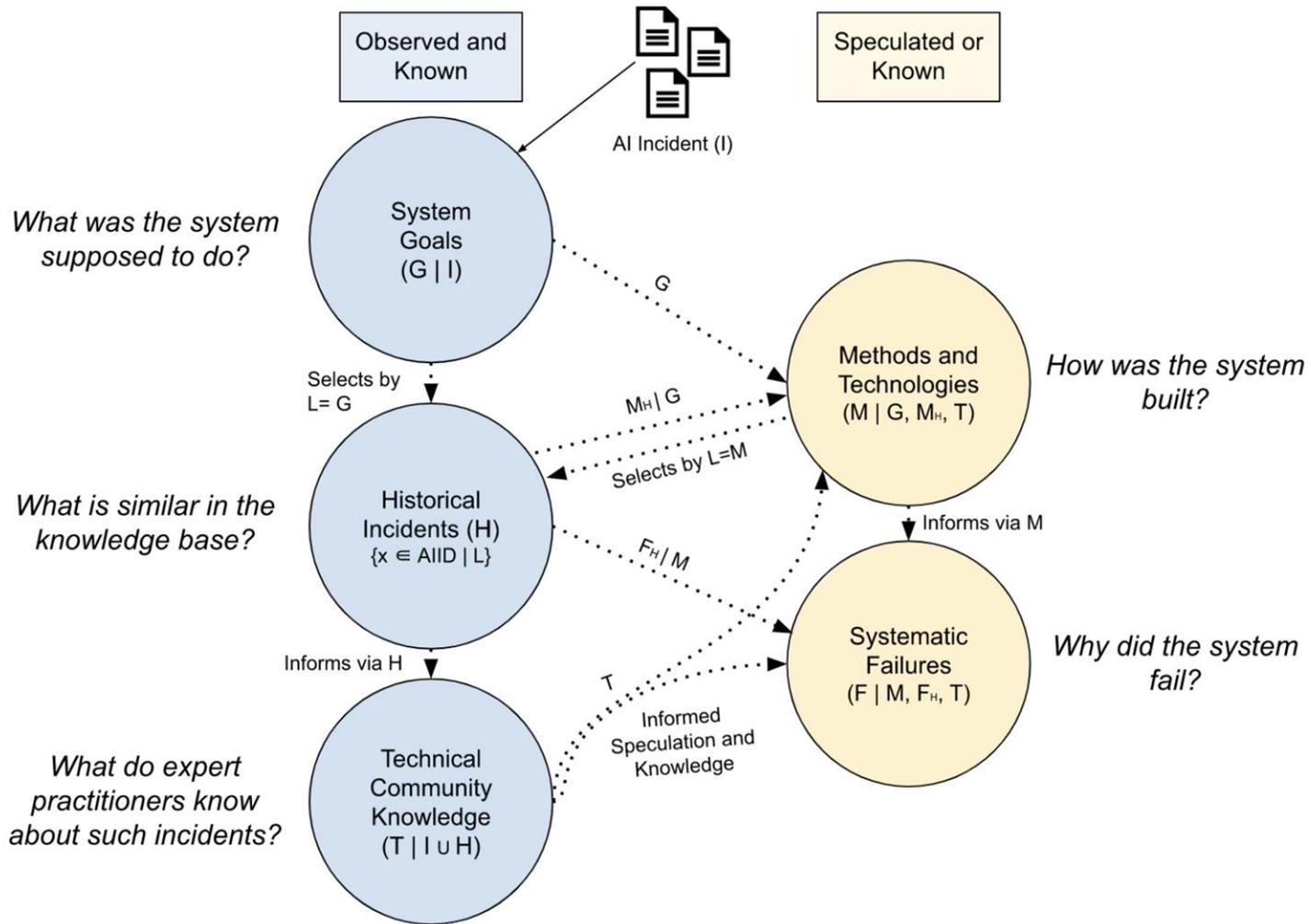


Three Inter-related Taxonomies: GMF

AI System **(G)goals**: *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?*



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



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.

AI System Tasks:

- Machine Translation

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
of complexity that they don't often face when working with other languages.

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

...ly difficult for machine translation

...e number of different dialects. This ...g with other languages.

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

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:

- Learning Dataset Imbalance
- Poor Generalization ?

potential

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

Example: AIID Incident #72 – Informative snippets

The error comes from a convolutional neural network.

In the caption, 'them' in Hebrew

AI System Technologies:

- Intermediate Modeling

potential

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

... convolutional ... site.

... mean 'attack

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:

- Learning Dataset Imbalance
- Poor Generalization ?

potential

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

Example: AIID Incident #72 – Informative snippet

AI System Failures:

- Error accumulation

The error comes from a neural network.
In the caption, 'them' in Hebrew

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

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 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



Preliminary Results

GMF Taxonomy Classifications [Edit](#)

[Taxonomy Details](#)

Notes

Known AI Goal

Translation

Known AI
Technology

Convolutional Neural Network, Recurrent Neural Network,
Distributional Learning

Potential AI
Technology

Intermediate modeling, Classification, Multimodal Learning, Image
Classification

Known AI
Technical
Failure

Dataset Imbalance, Distributional Bias

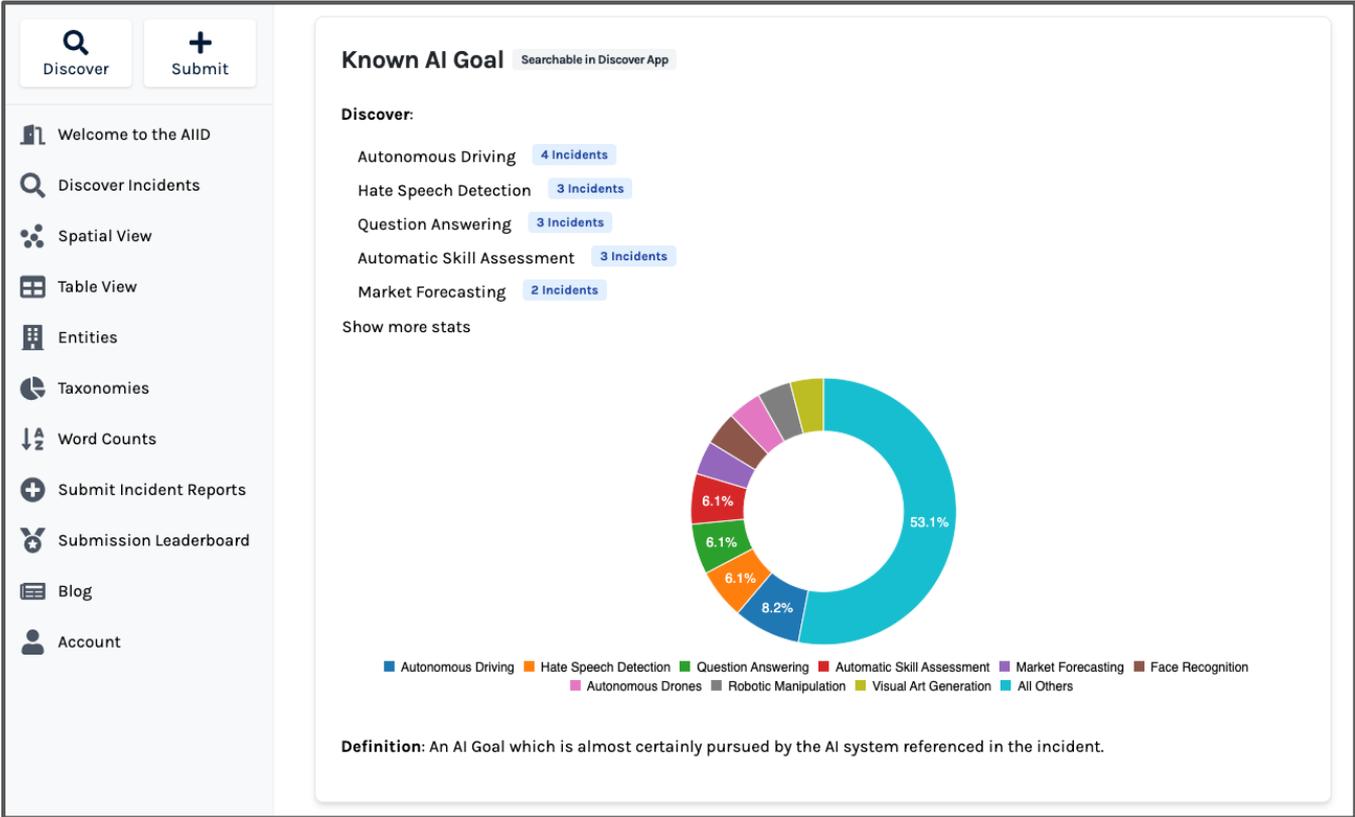
Potential AI
Technical
Failure

Generalization Failure

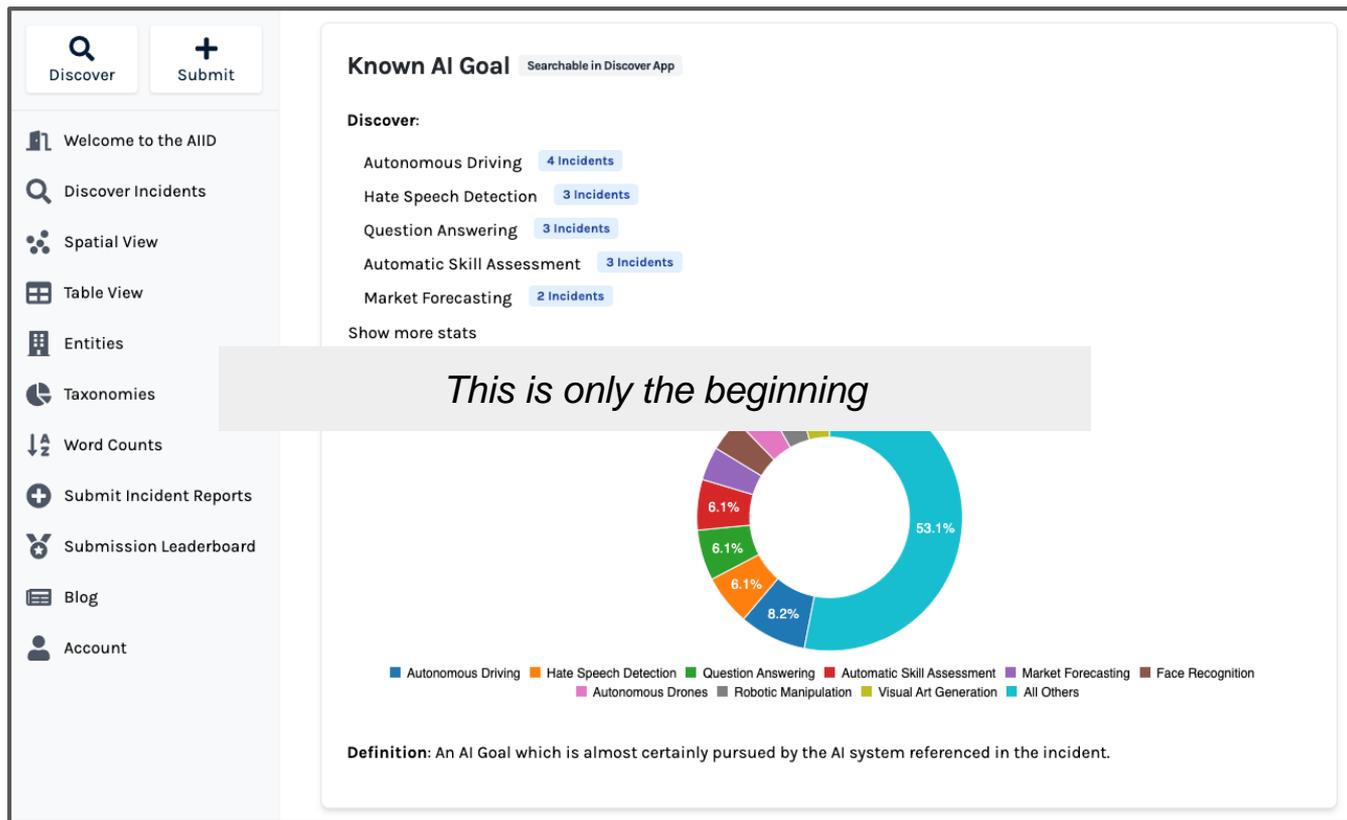
Show All Classifications



Preliminary Results



Preliminary Results



Outline

Part 1:
Indexing AI Incidents

Part 2:
Open Source Cause Analysis

Part 3:
What Next?

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!

Q	+	Download
Discover	Submit	
Discover Incidents		
Spatial View		
Table View		
Entities		
Taxonomies		
Word Counts		
Submit Incident Reports		
Submission Leaderboard		
Blog		
Account		

2023-02-08T10:20:04.000Z - 51.650557 MB - backup-20230206101930.tar.bz2
2023-01-30T10:20:11.000Z - 48.989663 MB - backup-20230130101933.tar.bz2
2023-01-23T10:18:27.000Z - 47.084304 MB - backup-20230123101851.tar.bz2
2023-01-16T10:20:02.000Z - 45.082807 MB - backup-20230116101930.tar.bz2
2023-01-13T05:20:15.000Z - 44.880759 MB - backup-20230113051937.tar.bz2
2023-01-13T04:58:53.000Z - 36.314042 MB - backup-20230113045829.tar.bz2
2023-01-09T10:20:37.000Z - 36.868553 MB - backup-20230109102012.tar.bz2
2023-01-02T10:18:45.000Z - 35.803995 MB - backup-20230102101853.tar.bz2
2022-12-26T10:18:21.000Z - 35.665481 MB - backup-20221226101758.tar.bz2
2022-12-12T10:20:12.000Z - 34.498653 MB - backup-20221212102017.tar.bz2
2022-12-09T01:02:10.000Z - 32.123157 MB - backup-20221209010454.tar.bz2
2022-12-05T10:20:08.000Z - 31.132809 MB - backup-20221205101847.tar.bz2
2022-11-29T00:12:18.000Z - 29.007248 MB - backup-20221129001501.tar.bz2
2022-11-14T10:09:13.000Z - 26.127888 MB - backup-20221114100853.tar.bz2
2022-11-07T10:08:11.000Z - 25.028911 MB - backup-20221107100850.tar.bz2
2022-10-31T10:09:29.000Z - 23.783197 MB - backup-20221031100911.tar.bz2
2022-10-24T10:15:03.000Z - 22.323507 MB - backup-20221024101445.tar.bz2
2022-10-17T10:13:10.000Z - 21.084165 MB - backup-20221017101513.tar.bz2
2022-10-10T10:09:36.000Z - 19.976716 MB - backup-20221010100917.tar.bz2
2022-10-03T10:11:05.000Z - 18.045061 MB - backup-20221003101048.tar.bz2
2022-09-26T10:09:32.000Z - 16.534274 MB - backup-20220926100916.tar.bz2
2022-09-19T10:09:18.000Z - 16.447859 MB - backup-20220919100900.tar.bz2
2022-09-12T10:09:10.000Z - 16.333689 MB - backup-20220912100854.tar.bz2
2022-08-29T10:28:05.000Z - 6.039861 MB - backup-20220829102751.tar.bz2
2022-08-22T10:26:36.000Z - 5.360793 MB - backup-20220822102642.tar.bz2
2022-08-15T10:27:04.000Z - 5.798507 MB - backup-20220815102847.tar.bz2
2022-08-08T10:26:30.000Z - 5.570368 MB - backup-20220808102612.tar.bz2
2022-08-01T10:27:54.000Z - 5.450121 MB - backup-20220801102740.tar.bz2
2022-07-25T10:27:21.000Z - 5.242716 MB - backup-20220725102706.tar.bz2
2022-07-18T10:27:32.000Z - 5.196813 MB - backup-20220718102714.tar.bz2
2022-07-11T10:25:27.000Z - 5.165373 MB - backup-20220711102514.tar.bz2
2022-07-04T10:27:53.000Z - 5.133867 MB - backup-20220704102738.tar.bz2
2022-06-27T10:26:34.000Z - 5.132517 MB - backup-20220627102621.tar.bz2
2022-06-20T10:27:02.000Z - 5.128568 MB - backup-20220620102650.tar.bz2
2022-06-13T10:26:58.000Z - 5.078101 MB - backup-20220613102645.tar.bz2
2022-06-06T10:27:45.000Z - 4.279906 MB - backup-20220606102728.tar.bz2
2022-05-30T10:26:53.000Z - 4.180937 MB - backup-20220530101837.tar.bz2
2022-05-23T16:56:27.000Z - 3.966397 MB - backup-20220523165615.tar.bz2
2022-05-20T04:27:01.000Z - 3.933319 MB - backup-20220520042701.tar.bz2
2022-05-16T10:13:23.000Z - 2.163683 MB - backup-20220516101223.tar.bz2
2022-05-09T10:11:22.000Z - 2.090388 MB - backup-20220509101114.tar.bz2



Thanks!

 Collab Board Members

Incident and Taxonomy Editors



More
Soon

Sean@incidentdatabase.ai

Some Past, Present, and Future Collaborating and Funding Orgs

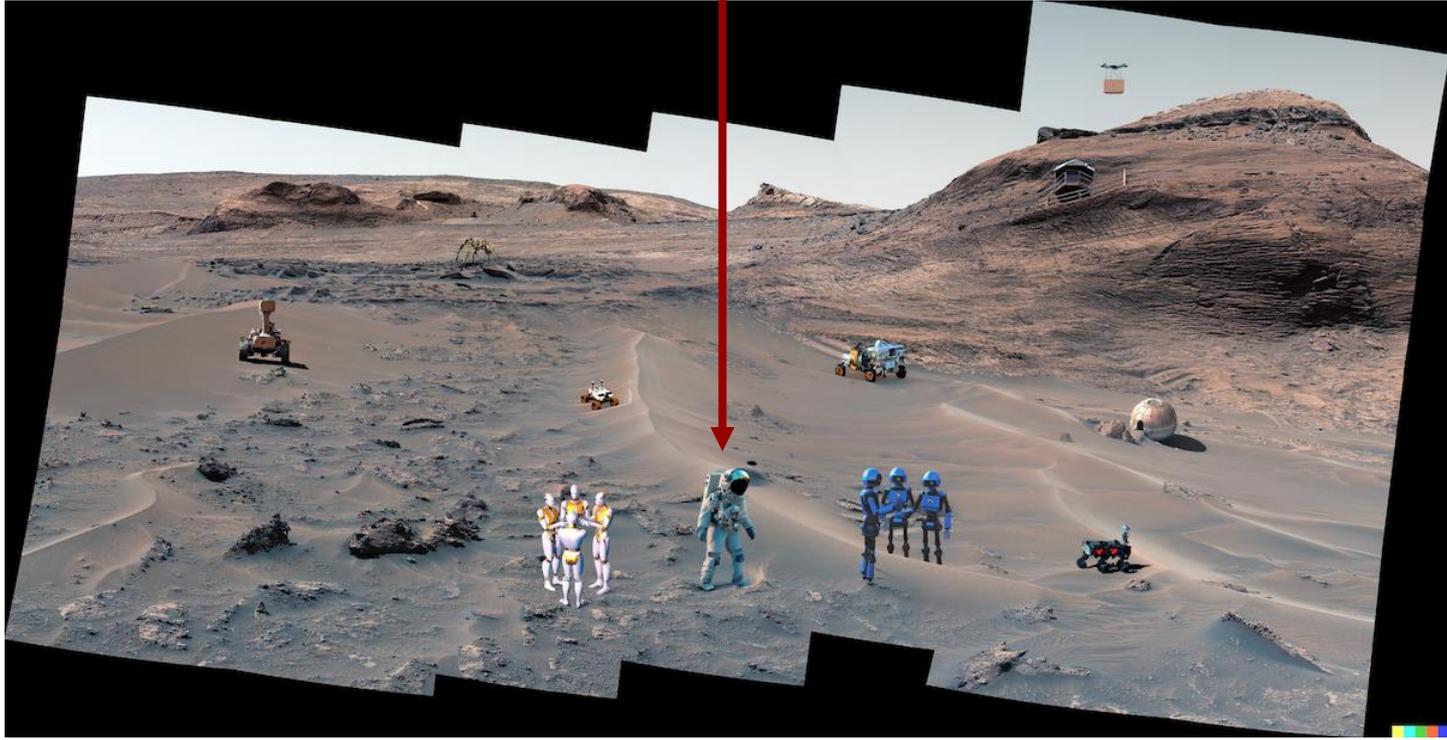


Some big announcements coming soon...



Thanks!

You are here



NID

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