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

This is a...problem
The Result of More AI

...with the Same Safety Culture
Incidents

#51, #68, #77, #261

#176, #289

#98, #207

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

**Entities Pages:**
See what is happening with individual companies

**Translation:**
Accepting Reports in 132 languages
Newer Features
Newer Features

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.
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?
What was the system supposed to do?

System Goals (G | I)

What is similar in the knowledge base?

Historical Incidents (H) (x ∈ AlID | L)

What do expert practitioners know about such incidents?

Technical Community Knowledge (T | I ∪ H)

Methods and Technologies (M | G, M_i, T)

Informed Speculation and Knowledge

Systematic Failures (F | M, F_i, T)

How was the system built?

Informs via M

Why did the system fail?

Informs via H

Selects by L = G

Selects by L = M

AI Incident (I)

Speculated or Known

Observed and Known
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.
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.
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 Technologies:**

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

**AI System Tasks:**

- Machine Translation
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

AI System Failures:
- Distributional Bias

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

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.
The error came 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 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:**
- 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.
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.

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 |
|------------------------------|------|
| **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 |
Preliminary Results

Known AI Goal

Discover:
- Autonomous Driving: 4 Incidents
- Hate Speech Detection: 3 Incidents
- Question Answering: 3 Incidents
- Automatic Skill Assessment: 3 Incidents
- Market Forecasting: 2 Incidents

Definition: An AI Goal which is almost certainly pursued by the AI system referenced in the incident.
Preliminary Results

This is only the beginning

Definition: An AI Goal which is almost certainly pursued by the AI system referenced in the incident.
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!
Thanks!

Collab Board Members

Incident and Taxonomy Editors

Some Past, Present, and Future Collaborating and Funding Orgs

Some big announcements coming soon...

Sean@incidentdatabase.ai
Thanks!

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