Responsible AI Collaborative

Founding Report

March 28th, 2022

Contents

Founding Letter 4

Founding Mission 5

Founding Board and Observers 6

Organizational Structure 9

Year in Review 10

- Headlines 11
- Citations 11
- Taxonomies 13
- User Interface Changes 13
- Systems Changes 14
- Program Expenses 14
- Community Projects 16
- Credits 18
- Year Ahead 19
  - Hiring! 19
  - Impact Forecast 19
  - Feature Forecast 19
  - Finances 20

Committed Projects 21

Prospective Projects 26

Founding Letter

At the inaugural meeting of the Partnership on AI's Safety Critical AI working group in 2018, the assembled engineers and philosophers could not agree on the **top priorities for making AI safer**. Hearing the rising choir of calls for more workshops, convenings, working groups, etc., I defaulted to the ML researcher's prerogative and declared an intention to **collect data**. Now more than three years later, a growing community of AI incident researchers has amassed a collection of 1300+ incident reports in a database that continually makes failings of intelligent systems discoverable by developers, researchers, and policy makers. In a world increasingly populated with intelligent systems, we now have the ability to identify the most urgent areas requiring technical, process, and policy advancements.

We have answered Santayana's aphorism, "*those who cannot remember the past are condemned to repeat it...*"
...With data.

Keeping up with increasing AI capacities and deployments motivates the graduation of the Incident Database from the founding sponsorship of the Partnership on AI to launch the **Responsible AI Collaborative**. Our founding purpose is to advance the change thesis of the AI Incident Database and our organization is modeled on the community-centered governance of the Common Vulnerabilities and Exposures database.

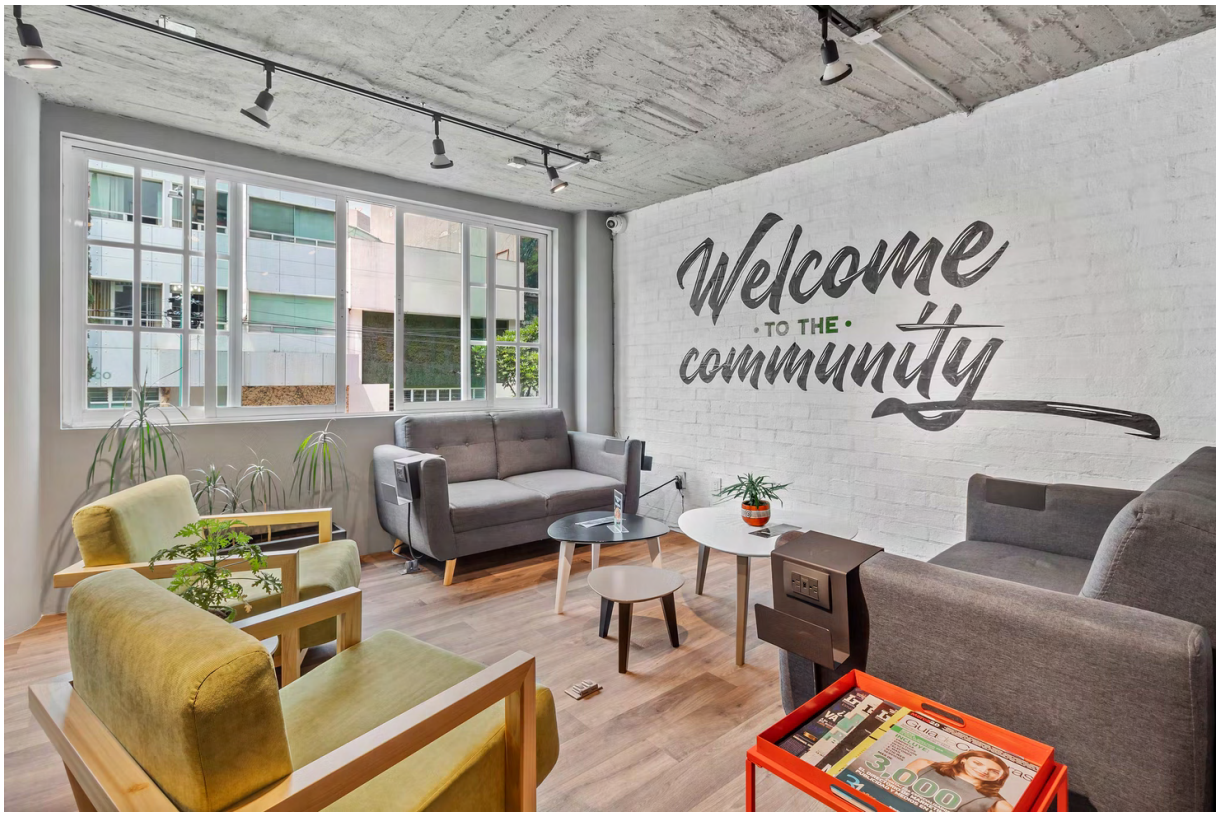
Recording and preventing the failures of engineered systems is of immense importance for the progress of society. Artificial intelligence can provide for the universal flourishing of people and society, or the debasing of our individual and collective values. It is only through collaboration that we will reliably produce a shared utopia. **Let's collaborate to make a better future!**

Sincerely,
Sean McGregor, Founder, Responsible AI Collaborative, sean@incidentdatabase.ai

Founding Mission

Identify, define, and catalog artificial intelligence incidents

— Credit: www.cve.org



Founding Board and Observers

Due to their history of contributing to the AI Incident Database project, the following people and organizations form the founding board of the RAIC.

Helen Toner

Helen Toner is Director of Strategy at Georgetown's Center for Security and Emerging Technology (CSET). She previously worked as a Senior Research Analyst at Open Philanthropy, where she advised policymakers and grantmakers on AI policy and strategy. Between working at Open Philanthropy and joining CSET, Helen lived in Beijing, studying the Chinese AI ecosystem as a Research Affiliate of Oxford University's Center for the Governance of AI. Helen has written for Foreign Affairs and other outlets on the national security implications of AI and machine learning for China and the United States, as well as testifying before the U.S.-China Economic and Security Review Commission. She is a member of the board of directors for OpenAI. Helen holds an MA in Security Studies from Georgetown, as well as a BSc in Chemical Engineering and a Diploma in Languages from the University of Melbourne.

Contributions: AI incident research and oversight of the CSET taxonomy.

Patrick Hall

Patrick Hall is principal scientist at bnh.ai, a D.C.-based law firm specializing in AI and data analytics. Patrick also serves as visiting faculty at the George Washington University School of Business. Before co-founding bnh.ai, Patrick led responsible AI efforts at the machine learning software firm H2O.ai, where his work resulted in one of the world's first commercial solutions for explainable and fair machine learning. Among other academic and technology media writing, Patrick is the primary author of popular e-books on explainable and responsible machine learning. Patrick studied computational chemistry at the University of Illinois before graduating from the Institute for Advanced Analytics at North Carolina State University.

Contributions: Patrick is the leading contributor of incident reports to the AI Incident Database Project.

Sean McGregor

Sean McGregor founded the AI Incident Database project and recently left a position as machine learning architect at the neural accelerator startup Syntiant so he could focus on the assurance of intelligent systems full time. Dr. McGregor's work spans neural accelerators for energy efficient inference, deep learning for speech and heliophysics, and reinforcement learning for wildfire suppression policy. Outside his paid work, Sean organized a series of workshops at major academic AI conferences on the topic of "AI for Good" and is currently developing an incentives-based approach to making AI safer through audits and insurance.

Contributions: Sean volunteers as a project maintainer and editor of the AI Incident Database (AIID) project.

Kit Harris

Kit leads on grant investigations and researches promising fields at Longview Philanthropy. He also writes about Longview recommendations and research for Longview client philanthropists. Prior to focusing on high-impact philanthropic work, Kit worked as a credit derivatives trader with J.P. Morgan. During that time, he donated the majority of his income to high-impact charities. Kit holds a first class degree in mathematics from the University of Oxford.

Contributions: Kit serves as board observer, provides strategic advice, and is RAIC's point of contact at Longview.

Neama Dadkhahnikoo

Neama Dadkhahnikoo is an expert in artificial intelligence and entrepreneurship, with over 15 years of experience in technology development at startups, non profits, and large companies.

He currently serves as a Product Manager for Vertex AI, Google Cloud's unified platform for end-to-end machine learning. Previously, Mr. Dadkhahnikoo was the Director of AI and Data Operations at the XPRIZE Foundation, CTO of CaregiversDirect (AI startup for home care), co-founder and COO of Textpert (AI startup for mental health), and a startup consultant. He started his career as a Software Developer for The Boeing Company.

Mr. Dadkhahnikoo holds an MBA from UCLA Anderson; an MS in Project Management from Stevens Institute of Technology; and a BA in Applied Mathematics and Computer Science, with a minor in Physics, from UC Berkeley.

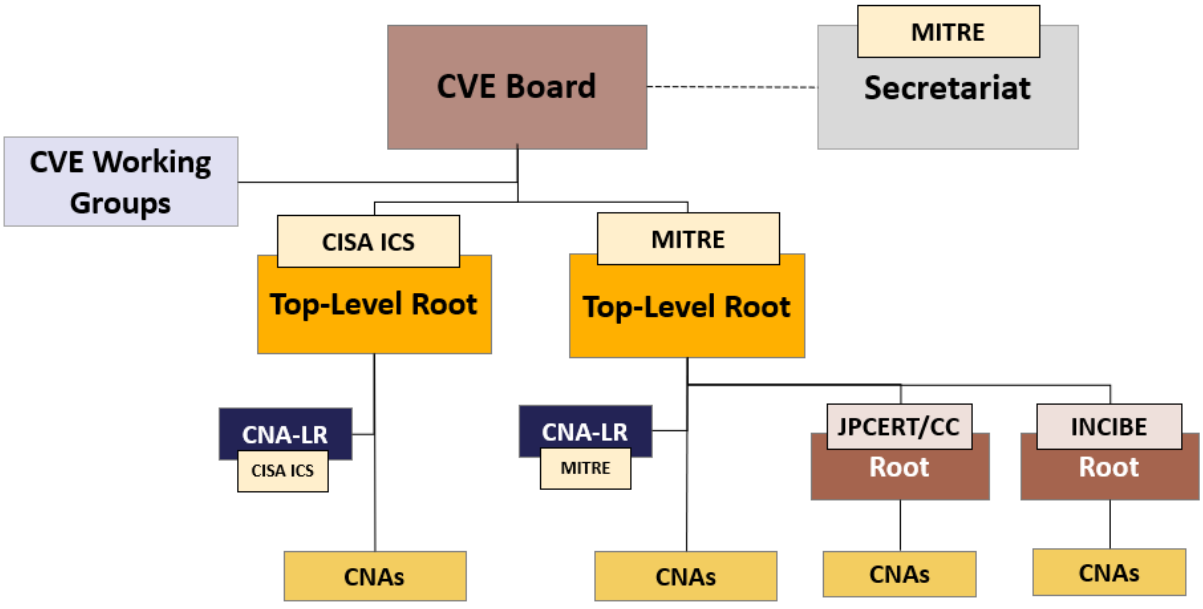
Contributions: Neama serves as volunteer of RAIC outside his role as product manager at Google.

All are welcome to apply or nominate people and organizations for the Responsible AI Collaborative (RAIC) board. The base requirement is substantial ongoing contributions to RAIC's programming.

Apply for the Board

Organizational Structure

The RAIC's organizing structure is heavily inspired by that of the Common Vulnerabilities and Exposures (CVE) database, which is highly participatory in nature and structured for the involvement of multiple stakeholders in the governance of the projects.



Inspiring organizational structure for the AI Incident Database

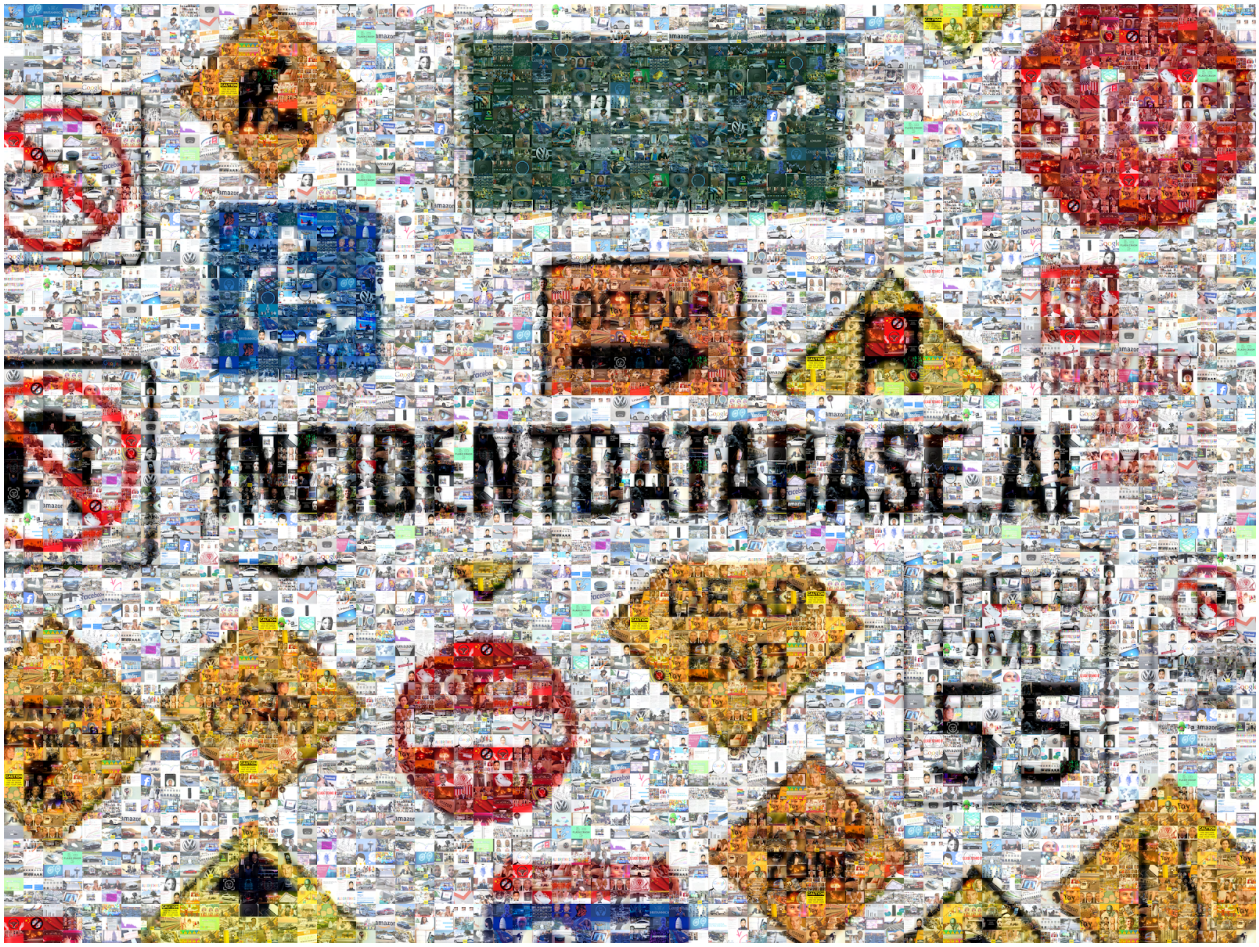
At launch, the AIID is RAIC's sole activity. If in the future RAIC engages in multiple activities requiring their own governance, then the CVE structure will be adopted for the AIID, RAIC's board will be converted to solely a fiscal sponsor, and new projects will be individually organized.

RAIC Board Charter

Responsible AI Collaborative

Year in Review

March 28, 2022



Each of the images in the mozaic are images associated with incident reports in the AIID

Headlines

- The number of **incidents** in the database **doubled**
- **First taxonomy** of AI incidents developed
- System funding recently expanded by **5x total budget** to date
- The system architecture has been **completely reworked**

VentureBeat WIRED VICE

Citations

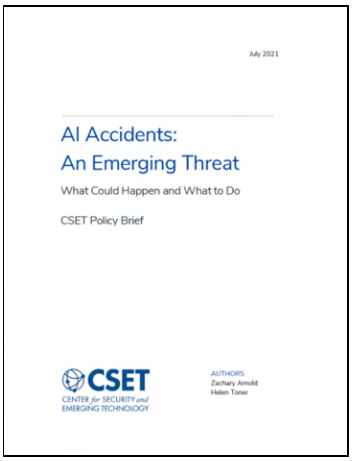
Google Scholar indexed research papers and books.


POLICY FORUM

ARTIFICIAL INTELLIGENCE

Filling gaps in trustworthy development of AI

Incident sharing, auditing, and other concrete mechanisms could help verify the trustworthiness of actors


July 2021

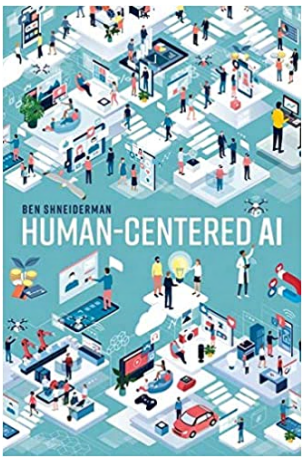
AI Accidents: An Emerging Threat

What Could Happen and What to Do

CSET Policy Brief

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

AUTHORS Zachary Arnold Helen Toner



- Arnold, Z., Toner, H., CSET Policy. "AI Accidents: An Emerging Threat." (2021).
- Schwartz, Reva, et al. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." (2022).
- Aliman, Nadisha-Marie, Leon Kester, and Roman Yampolskiy. "Transdisciplinary AI Observatory—Retrospective Analyses and Future-Oriented Contradistinctions." Philosophies 6.1 (2021): 6.
- Shneiderman, Ben. Human-Centered AI. Oxford University Press, 2022.

- Nor, Ahmad Kamal Mohd, et al. "Abnormality Detection and Failure Prediction Using Explainable Bayesian Deep Learning: Methodology and Case Study of Real-World Gas Turbine Anomalies." (2022).
- Falco, Gregory, and Leilani H. Gilpin. "A stress testing framework for autonomous system verification and validation (v&v)." 2021 IEEE International Conference on Autonomous Systems (ICAS). IEEE, 2021.
- Petersen, Eike, et al. "Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions." arXiv preprint arXiv:2107.09546 (2021).
- John-Mathews, Jean-Marie. AI ethics in practice, challenges and limitations. Diss. Université Paris-Saclay, 2021.
- Macrae, Carl. "Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety and Sociotechnical Sources of Risk." Safety and Sociotechnical Sources of Risk (June 4, 2021) (2021).
- Weissinger, Laurin, AI, Complexity, and Regulation (October 14, 2021). OUP Handbook on AI Governance, Forthcoming
- Hong, Matthew K., et al. "Planning for Natural Language Failures with the AI Playbook." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.
- Ruohonen, Jukka. "A Review of Product Safety Regulations in the European Union." arXiv preprint arXiv:2102.03679 (2021).
- Kalin, Josh, David Noever, and Matthew Ciolino. "A Modified Drake Equation for Assessing Adversarial Risk to Machine Learning Models." arXiv preprint arXiv:2103.02718 (2021).
- Aliman, Nadisha Marie, and Leon Kester. "Epistemic defenses against scientific and empirical adversarial AI attacks." CEUR Workshop Proceedings. Vol. 2916. CEUR WS, 2021.
- John-Mathews, Jean-Marie. L'Éthique de l'Intelligence Artificielle en Pratique. Enjeux et Limites. Diss. université Paris-Saclay, 2021.
- Xie, Xuan, Kristian Kersting, and Daniel Neider. "Neuro-Symbolic Verification of Deep Neural Networks." arXiv preprint arXiv:2203.00938 (2022).
- Please reach out if your publication is missing from the list.

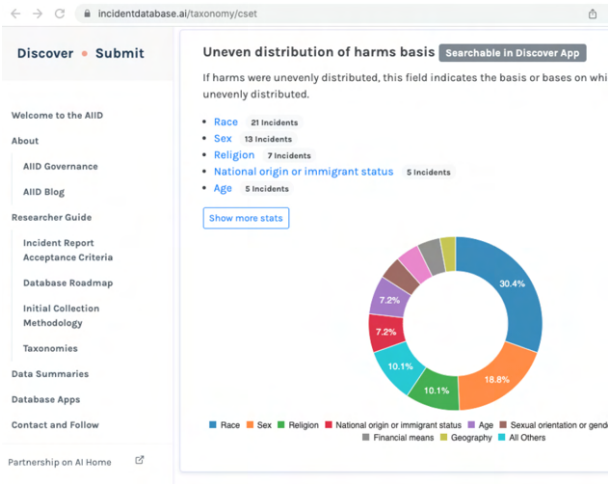
Taxonomies

2021 saw the first taxonomy added to the AI Incident Database: The CSET taxonomy has been applied to 100 incident records for a total of more than 2,000 classifications.

[About the CSET Taxonomy]

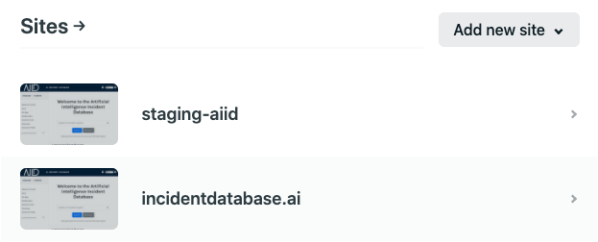
User Interface Changes

- Updated landing page
- Leaderboard page
- Widgets for recent incident reports
- Updated discover application
- Updated faceting
- New taxonomy view
- New map view
- New timeline view
- New logo
- New color scheme
- Redesign of discover page
- See more



Systems Changes

- Weekly database snapshot availability
- Complete staging environment
- Additional REST API endpoints
- Google geocoding API integration
- Programmatic Algolia index updates
- New user authentication strategy
- Weekly snapshotting to S3
- Support for translation of incident reports to Spanish
- Continuous integration testing
- More



Program Expenses

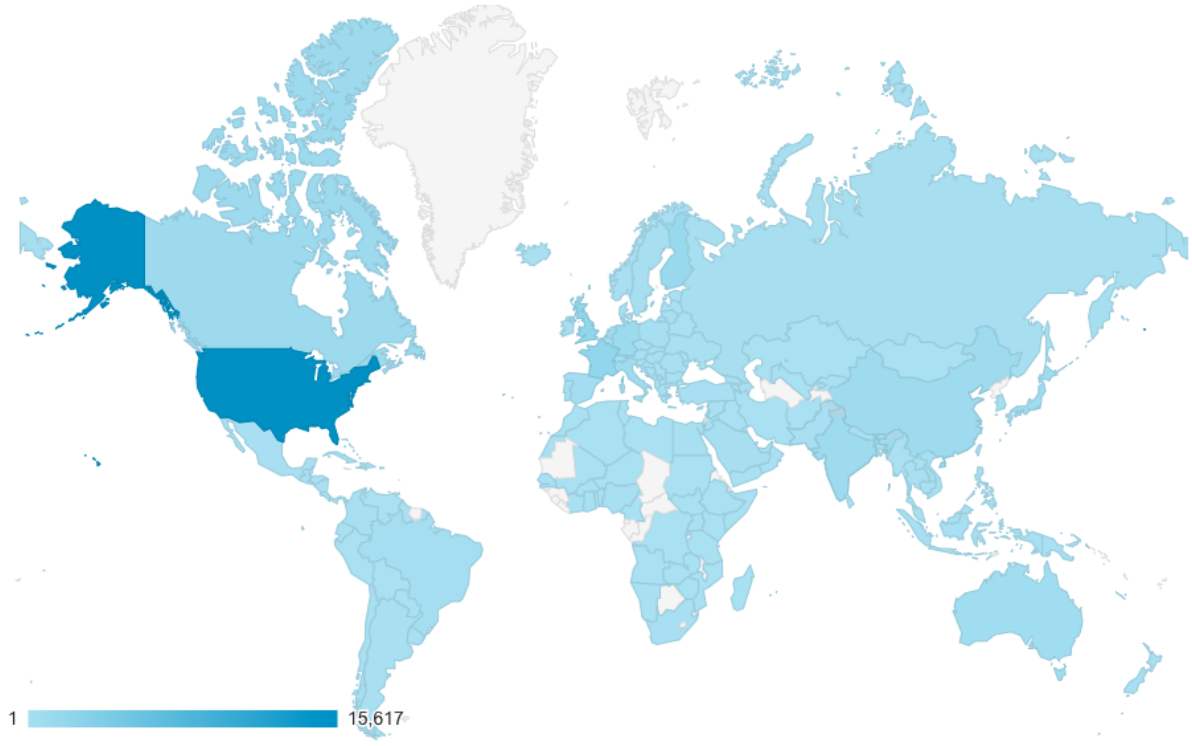
Paid during calendar year 2021. **Thank you to the Partnership on AI for its support of these costs**.

PARTNERSHIP ON AI

Expense	Cost (USD)
Google Cloud Hosting	485.61
Netlify Build Expenses	548.80
MongoDB Realm Hosting	239.33

Algolia Index Hosting	70.00
Web Developer Pay	102,777.83
AAAI Conference Registration	274.00
Picture Mosaics	35.95
Amazon Web Services	0.54
AirBnB	617.84
Food and Drink	23.76
Total	**105,073.66**

Community Projects



Persons from 90+ percent of the world have visited the AI Incident Database. It is among our goals to facilitate the development of community projects representing this full diversity of viewpoints. Please consider integrating the languages, incidents, and analysis of your perspective

Project: CSET Taxonomy
Who: This is a project of the Center for Security and Emerging Technology at Georgetown University
Description: The Center for Security and Emerging Technology (CSET) taxonomy is a general taxonomy of AI incidents. There are a large number of classified attributes, including ones pertaining to safety, fairness, industry, geography, timing, and cost.All classifications within the CSET taxonomy are first applied by one CSET annotator and reviewed by another CSET annotator before the classifications are finalized. The combination of a rigorously defined coding set and the completeness with which it has been applied make the CSET taxonomy the AIID's gold standard for taxonomies. Nevertheless, the CSET taxonomy is an ongoing effort

and you are invited to report any errors you may discover in its application.

Project: Related Database Integration
Who: Several Databases
Description: We are currently talking with several databases about integrating the central incident identifiers of the AIID with more specialized systems tailored to security and policy needs.

Project: Vector Embeddings of Incident Text
Who: This is a project of an Oregon State University capstone team working over the course of their senior year.
Description: The main aim of the OSU Capstone team is to create an API that the AI Incident Database will be able to utilize in its future growth. The API will make requests to a Natural Language Processing model (Longformer) with a given input text and return a list of AI Incidents from the database that the input text is most similar to. The project is currently in a state where both the model and the API are working as expected. The model uses a tokenized collection of existing database entries and runs a cosine similarity between each incident's tokenized representation and the input text's tokenized representation. The current state of the API allows for both GET and POST requests, with the arguments for these calls being the text and the number of incidents it should return, with a negative number returning all the incidents in order from most similar to least. All code is still in personal repositories or personal AWS accounts and will be integrated in the coming months. During this time, we will also be extensively documenting our project to ensure future use and expansion will be easy.

Project: Incident Curation
Who: Various individuals and organizations
Description: Appropriately judging and accepting potential incidents into the AIID is a difficult human-centered task. We are now assembling a group of volunteer editors to accelerate the process.

Project: Database Visualization
Who: Various individuals that have reached out to the AIID project
Description: The AI Incident Database is built to be an extensible infrastructure for others to provide, among other inputs, visualizations and custom views into the database. At present people are working to integrate the Observable visualization site into the AIID so that a large

number of useful visualizations can be generated and listed on the AIID website.

Credits

Maintainer of the Year

César Varela

Submitters of the Year

Roman Lutz and Kate Perkins

Taxonomy of the Year

Georgetown Center for Security and Emerging Technology

PAI Staff

The following Partnership on AI staff deserve special credit for their efforts to advance the AI Incident Database (AIID) program.

Jingying Yang
Christine Custis

Open Source Contributors

People that have contributed more than one pull request to the code base or graphic to the site.

César Varela
Alex Muscă
Chloe Kam
JT McHorse
Seth Reid

Incident Editors

Incident Contributors

People that have contributed a large number of incidents to the database.

Patrick Hall (Burt and Hall LLP)
Kate Perkins (Intel)
Roman Lutz (Max Planck Institute for Intelligent Systems, formerly Microsoft)
Sam Yoon (as contractor to PAI, now at Kennedy School of Government)
Catherine Olsson (Anthropic, formerly of Google)
Roman Yampolskiy (University of Louisville)

The following people have collected a large number of incidents that are pending ingestion.

Zachary Arnold, Helen Toner, Ingrid Dickinson, Thomas Giallella, and Nicolina Demakos (Center for Security and

People that resolve incident submissions to the database.

Sean McGregor

Taxonomy Editors

Organizations or people that have contributed taxonomies to the database.

Center for Security and Emerging Technology (CSET)

Emerging Technology, Georgetown)
Charlie Pownall via AI, algorithmic and automation incident and controversy repository (AIAAIC)
Lawrence Lee, Darlena Phuong Quyen Nguyen, Iftekhar Ahmed (UC Irvine)

Year Ahead

Hiring!

- Two to four full-time positions

Impact Forecast

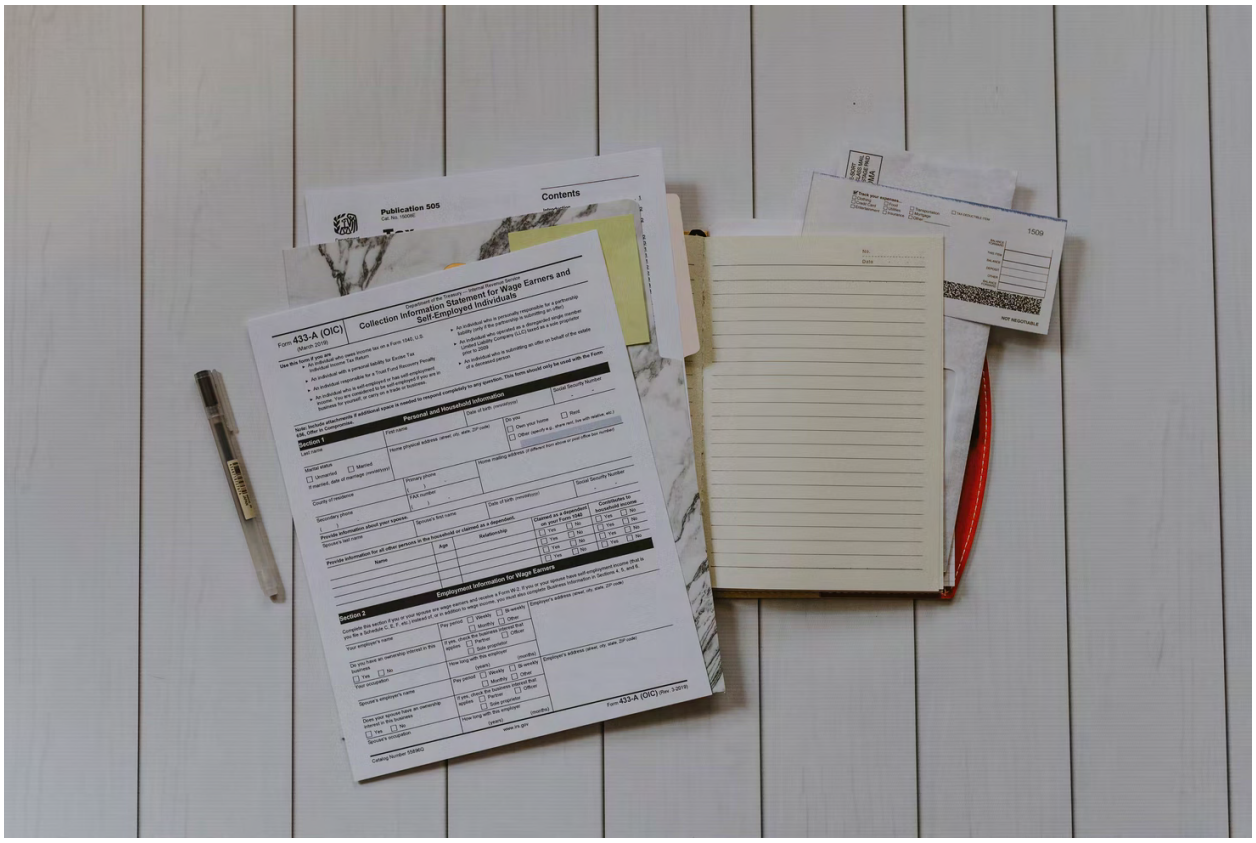
- Greater adoption of AI incident analysis as a means of making safer and more socially beneficial AI systems
- Expansion of the AI incident research and developer communities
- Expanded user base

Feature Forecast

- An additional doubling of incidents indexed in the AIID to 400+ incidents
- Translation of all incidents to and from 100+ languages
- A resources taxonomy applied to 100+ incident citation pages
- Begin NLP monitoring for incidents similar to those currently in the database
- Expanded group of credentialed editors on the database
- Programmatic integration with external systems

Finances

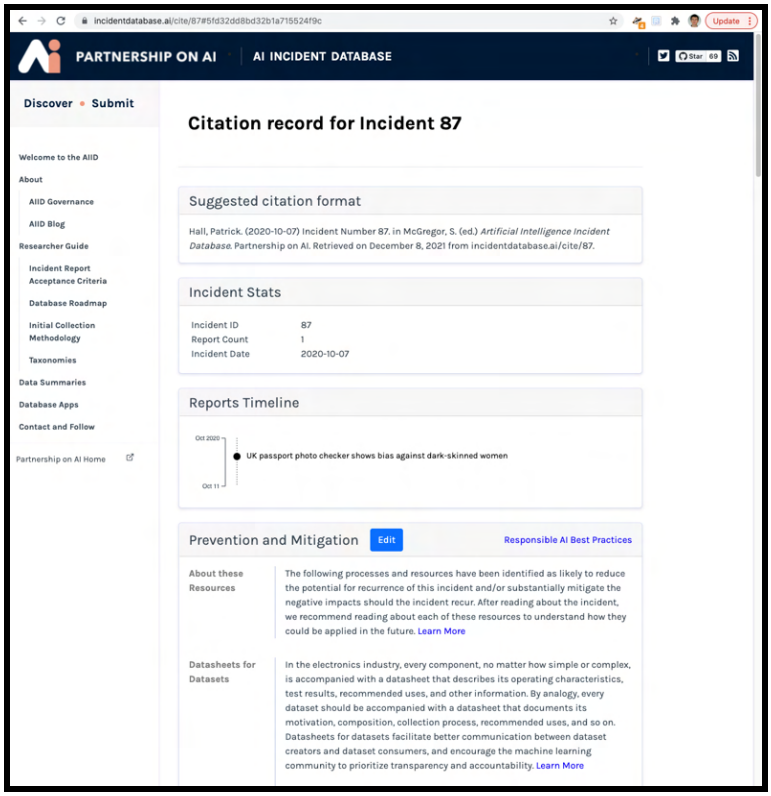
On the recommendation of Longview Philanthropy, the Waking Up Foundation has contributed a founding grant of $550k USD directed to the AI Incident Database. $33k of the total is committed to RAIC's fiscal sponsor for running the books and tax compliance for the organization. Approximately $5k is dedicated to maintaining the RAIC non-profit entity that directs the AIID programming sponsored by a fiscal sponsor. The remaining funds are all dedicated to programmatic outputs, including staff time for programming and research.



Committed Projects

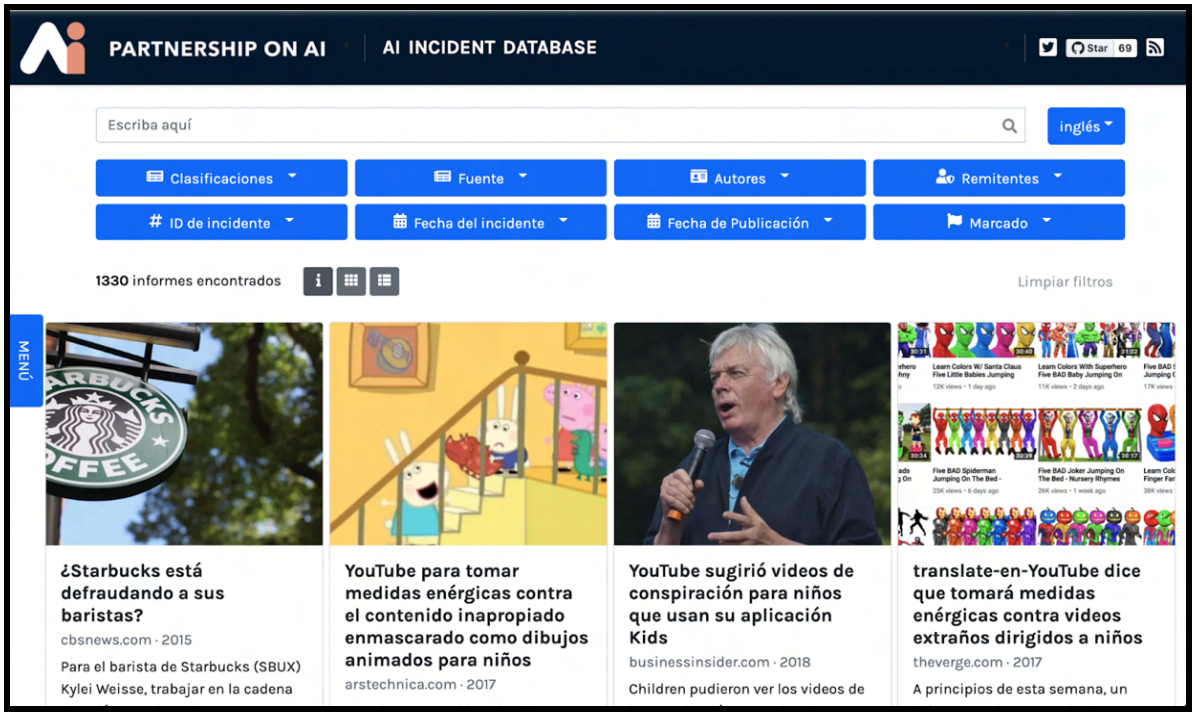
Project 1: Best Practices for Preventing Incidents

1. **User Story**
    a. "As a company shipping an AI product, I want to know how I can prevent this incident from happening with our new product so we will not be {embarrassed, sued, bankrupted}"
2. **Engineering Activity**
    a. Front end improvements to the taxonomy feature specific to the research activity.
    b. Develop pages detailing responsible AI resources
3. **Research Activity**
    a. Identify collection of actionable responsible AI resources that are appropriately associated with collected incidents.
    b. Write resource detail pages.
    c. Classify all incidents according to their most relevant resources.
4. **Success Measure:** Track view and click counts on resources
5. **Status**
    a. Principal software engineering complete, needs content and application of resources to incident citation pages



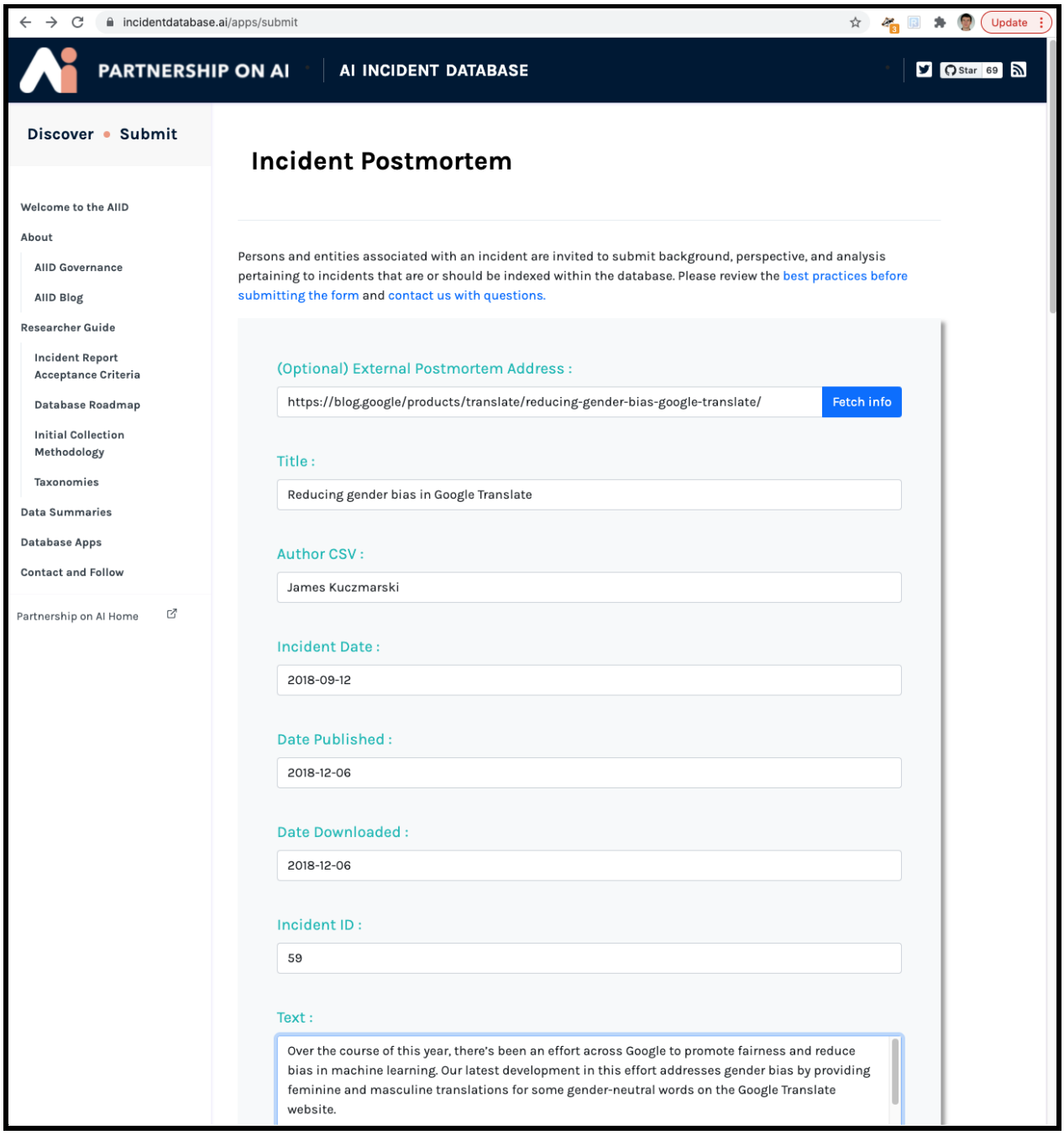
Project 2: Translate from/to 100+ Languages

1. **User Story**
    a. "As a person in {the United Kingdom, Indonesia}, I want to know how AI has failed in {the United Kingdom, Indonesia} so I will not {develop, use} systems that are harmful to people or society in {the United Kingdom, Indonesia}."
2. **Engineering Activity**
    a. Systems required for producing and maintaining 100+ translations of incidents and classifications within the database.
    b. Localization of the user interface.
3. **Research Activity**
    a. Audit the efficacy of the machine translations
    b. Coordinate with NGOs operating outside English-speaking countries to promote the submission and knowledge of AI incidents outside the English speaking world.
4. **Success Measures**
    a. Track user counts according to source geography and languages viewed
    b. Total incident reports submitted from non-English languages
5. **Status**
    a. Proof of concept on indexing in Spanish complete. Need to deploy for all languages. Need to also localize for all languages and ingest from non-English languages.



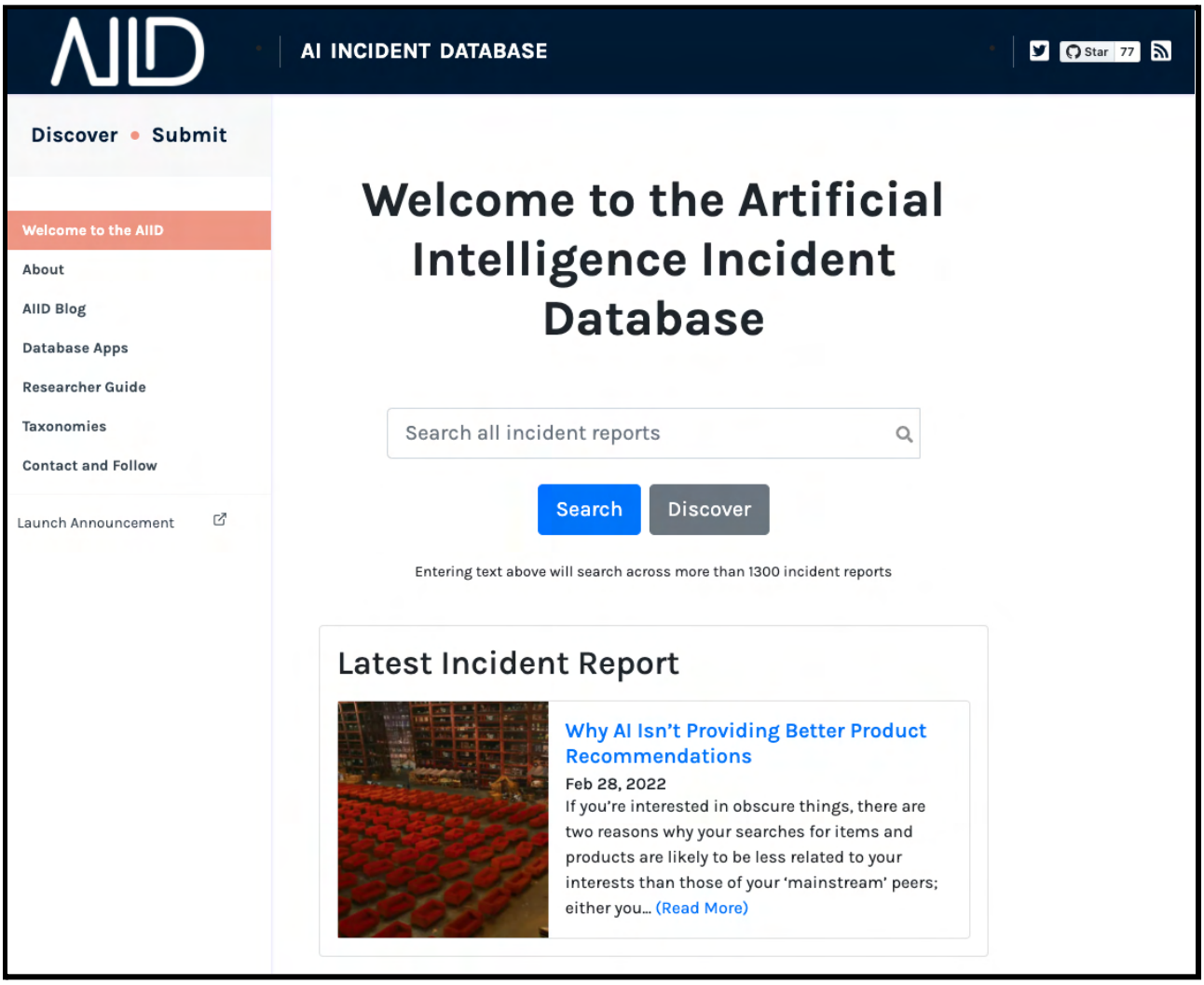
Project 3: Post-Mortem Reporting from Corporate Incidents

1. **User Stories**
    a. "As a company that has shipped a product *now in the AI Incident Database*, I want to publish our perspective on what happened, why it happened, what we could have done to prevent it, and what we will do in the future so everyone can produce better systems for the world." (positive view)
    b. "As a company that has shipped a product that has caused an incident *that is not in the AI Incident Database*, I want to publish our perspective on what happened, why it happened, what we could have done to prevent it, and what we will do in the future so we can get in front of the bad press that will land when the world learns of this." (negative view)
    c. "As a researcher of AI Incidents, I would like to know the perspective of companies about their incidents so I can identify current and potential future risks of AI systems."
2. **Engineering Activity**
    a. Develop reporting form, identity verification, backend functions for postmortem collection and approval, and continuous integration systems.
3. **Research Activity**
    a. Design the incident postmortem form and process
    b. Produce a research paper on postmortem practices and promote it to large companies that typically produce multiple incidents per year
    c. Run a workshop with the risk managers of large tech companies to discuss incident postmortem practices and promote the adoption of such disclosures as an industry standard
4. **Success Measures**
    a. Number of postmortems submitted to AI Incident Database
5. **Status**
    a. No progress yet.


incidentdatabase.ai/apps/submit
Update
PARTNERSHIP ON AI
AI INCIDENT DATABASE
Star 69
Discover • Submit
Welcome to the AIID
About
AIID Governance
AIID Blog
Researcher Guide
Incident Report Acceptance Criteria
Database Roadmap
Initial Collection Methodology
Taxonomies
Data Summaries
Database Apps
Contact and Follow
Partnership on AI Home
Incident Postmortem
Persons and entities associated with an incident are invited to submit background, perspective, and analysis pertaining to incidents that are or should be indexed within the database. Please review the best practices before submitting the form and contact us with questions.
(Optional) External Postmortem Address :
https://blog.google/products/translate/reducing-gender-bias-google-translate/
Fetch info
Title :
Reducing gender bias in Google Translate
Author CSV :
James Kuczmarski
Incident Date :
2018-09-12
Date Published :
2018-12-06
Date Downloaded :
2018-12-06
Incident ID :
59
Text :
Over the course of this year, there's been an effort across Google to promote fairness and reduce bias in machine learning. Our latest development in this effort addresses gender bias by providing feminine and masculine translations for some gender-neutral words on the Google Translate website.

Project 4: Incident collection and NLP incident detection tooling: Modern natural language processing (NLP) methods have the ability to scrape the entire internet on a daily basis for candidate incidents that have some similarity to incident reports already in the database. We would like to use this capability to monitor all indexed world news for candidate incidents and ingest them on mass on a daily basis.

- **Success Measures**
  - Number of ingested incident reports
  - Total number news sources monitored
- **Status**
  - Undergraduate capstone team will complete a proof of concept before the end of the school year. From there we need to systematize.



Prospective Projects

Project 5: Pilot Federation: This is highly speculative because we have not selected a partner with which to pilot a federated database standard. Stay tuned for more information in the future
Success measure: adoption of the AIID codebase by additional organizations or countries.

Project 6: Technical failure taxonomy:
The AI Incident Database does not offer one canonical source of truth regarding AI incidents. Indeed, reasonable parties will have well-founded reasons for why an incident should be reported or classified differently. Consequently, the Database supports multiple perspectives on incidents both by ingesting multiple reports (to date, 1199 authors from 547 publications), and by supporting multiple *taxonomies*. The AIID taxonomies are flexible collections of classifications managed by expert individuals and organizations. The taxonomies are the means by which society collectively works to understand both individual incidents, as well as the population-level statistics for these classifications. Well structured and rigorously applied AI incident taxonomies have the ability to inform research and policy making priorities for safer AI, as well as help engineers understand the vulnerabilities and problems produced by increasingly complex intelligent systems.

The first taxonomy built for the AIID is managed by the Georgetown Center for Security and Emerging Technology (CSET). The CSET taxonomy is a general taxonomy of AI incidents involving several stages of classification review and audit to ensure consistency across annotators. Among the classifications are elements related to safety, fairness, industry, geography, timing, and cost. The intention behind the CSET taxonomy is to inform policy makers of impacts and it has been used for this purpose in both the United States and Europe. Even with the success of the CSET taxonomy for policy makers, the AIID still lacks a rigorous technical taxonomy linking experienced harms to their technical elements. We would like to close this gap by hiring machine learning engineers according to the following execution plan,

- **Execution Plan**
    - Hire a research engineer in machine learning.
        - Develop or adopt a taxonomy for the components involved and failure modes experienced. Prioritize technologies implicated by the already-collected incidents.
        - Develop a questionnaire for engineers with insider knowledge of incidents to submit relevant technical information regarding the incidents.
    - Hire a community-centered researcher to work with the partner community to self-disclose the technical details of their incidents. Additionally, leverage the open source technical community to apply classifications to incidents using publicly available reports (i.e., black box classification).

        - Develop collaborations and relationships with research organizations in the application of black box classification.
        - Contribute to the development of mandatory reporting standards now being considered by various governments.
- **Success Measures**
    - Number of incidents with black box taxonomy classifications
    - Number of incidents with self-disclosure taxonomy classifications
    - Publication of a research paper based on the taxonomy
    - Number of citations of the associated research paper

Project 7: Incident *data* collection (sensors, test datasets, etc.): Many incident types produce incident data that is analyzed by companies and used to avoid its recurrence internally. A research associate could be tasked with promoting the voluntary disclosure of incident data, which could then be associated with incident records. In the future, this type of disclosure is likely to be mandatory in some jurisdictions and the AIID should be there to support it.

- **Success Measures**
    - Number of incidents with primary source data.

Project 8: Your project here:
The AI Incident Database is governed by the Responsible AI Collaborative (RAIC). You should consider this an open invitation to collaborate on projects that match the impact thesis of the AI Incident Database.