Developing a framework for AI incident reporting, and an AI Incidents Monitor (AIM)

Audrey Plonk
Head of the Digital Economy Policy Division
AI incidents come in many shapes!

Picture on bus mistaken for human

AIID Incident 36: Picture of woman on bus billboard led to woman shamed for jaywalking and credit score dropped in China

Human mistaken for car

AIID Incident 171: Traffic camera read text on pedestrian T-shirt and interpreted it as a license plate and fined the license plate owner in the U.K.
Why monitor AI incidents?

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. ...

Source: AI Incidents database (AIID) incident 6
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

Source: AI Incidents database (AIID) incident 106
Why monitor and report AI incidents?

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

–George Santayana, *The Life of Reason*

AI systems can cause real-world harm to people, organisations and the environment.

In an initial dataset compiled by the OECD from media reports, the rates of incidents climbs rapidly.
Why is a *common framework* for AI incident reporting needed now?

- AI risks are materialising into incidents
- Treating/mitigating AI risks requires learning from evidence on past AI incidents, which calls for **global consistency and interoperability** in incident reporting
- Risks and incidents can then be linked to AI system *characteristics*
  - To inform policy and regulation
  - To enable risk treatment / mitigation
- Concepts of AI incident reporting are already forming, but coordination is required to ensure interoperability
Experts involved to date come from

- the OECD
- the European Commission
- the AI Incident Database (AIID)
- the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENCELEC)
- the Centre for European Policy Studies (CEPS) in Brussels
- the National Institute of Standards and Technology (NIST) of the US
- the Center for Security and Emerging Technology (CSET) at Georgetown
- the XPRIZE Foundation
- the Jozef Stefan Institute (Ljubljana, Slovenia)
- the Infocomm Media Development Authority (IMDA) of Singapore
- And more...
A common framework for AI incident reporting starts with *defining* an “AI incident” and related concepts.

Characteristics of successful definitions related to AI incidents:

- Clear and operational
- Actionable and useful
- Modular and flexible
- Aligned with other incident reporting regimes
- Forward-looking
Distinguishing different types of risks caused by AI systems

Severity of harm

« Serious »
EUAIA: person death; serious damage to person health, to property or the environment; serious and irreversible disruption of critical infrastructure

Less « serious »

- Hazard (Has not occurred)
- Near miss (Almost occurred)
- AI Incident (Did occur)

Serious AI problem
(« Serious AI incident » in EUAIA)
Illustrating different types of risks caused by AI systems

Severity of harm

« Serious »

Less « serious »

Has not occurred

Almost occurred

Did occur

Materialisation of risk

Serious AI problem (« serious AI incident » in EUAIA)

e.g. Autonomous vehicle severely injures pedestrian or almost injures pedestrian.

Hazard
e.g. deployment of content recommendation system known to demonstrate certain social biases

Near miss
e.g. Autonomous vehicle fails to stop at a red light – narrowly missing pedestrian.

AI Incident
e.g. Facial recognition system is deployed in workplace without notice to employees.
### Current working definitions

<table>
<thead>
<tr>
<th>Hazard</th>
<th>A situation where the risk posed by an AI system, or the perception thereof, is relevant to a possible harm to person(s), property, or the environment that has yet to occur, including an infringement upon human rights, such as privacy and non-discrimination.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near miss</td>
<td>An event where the development or use of an AI system [allegedly] would have caused harm to person(s), property, or the environment, were it not for external circumstances.</td>
</tr>
<tr>
<td>AI incident</td>
<td>An event where the development or use of an AI system [allegedly] caused harm to person(s), property, or the environment.</td>
</tr>
<tr>
<td>Serious AI problem (« Serious AI incident » in EUAIA)</td>
<td>“Incident that directly or indirectly leads, might have led or might lead to: (a) the death of a person or serious damage to a person’s health, to property or the environment, (b) a serious and irreversible disruption of the management and operation of critical infrastructure”. (EUAIA)</td>
</tr>
</tbody>
</table>
Next Steps

- Thematic workshops with invited experts (every 4-6 weeks)
- Discussion on concept note (AIGO meeting of November 2022)
- Proposed OECD AI incident definition (AIGO meeting of April 2023)
- Building and testing the incident reporting framework (2023)

*AI Incidents Monitor project (in parallel)*
Overview of AI Incidents Monitor (AIM)
Using AI incidents to inform policy

• **Goal:** Through AI incidents, build evidence base to inform
  • Incident reporting framework
  • AI risk assessments
  • AI foresight work
  • Regulatory choices

• **Approach:**
  • Start by leveraging publicly-available news articles on AI incidents and hazards from reputable sources, recognising limitations
  • At later stage, enable direct submissions (e.g. from individuals, organisations or governments)
  • Collaborate with relevant partners e.g. CSET, RAIC’s AI Incidents Database, Event Registry, Jozef Stefan Institute
AI Incidents Monitor (AIM)

• Global & multilingual monitoring of AI incidents and hazards
• Automatic identification of AI incidents from news articles

Step 1 – Collect training data [completed]: collected a sample of over 900 manually identified AI incidents or hazards since 2009 to illustrate trends and help train automated system

Step 2 – Identify AI incidents in real time [in progress]: automating AI incidents identification and classification in real time into the criteria of the OECD classification framework, using natural language processing
Step 1: Collection of training data

Illustrative findings from manually identified news articles on AI incidents and hazards
*Caveat*: significant sampling bias (not the whole story)

### Sectors covered by news articles about AI incidents

- Government, security and defence
- Media, social platforms, marketing
- Healthcare, drugs and biotechnology
- Education and training
- Business processes and support services
- Mobility and autonomous vehicles
- Robots, sensors, IT hardware
- Travel, leisure and hospitality
- IT infrastructure and hosting
- Financial and insurance services

Step 1: Collection of training data

Illustrative findings from manually identified news articles on AI incidents and hazards
Caveat: significant sampling bias (not the whole story)

Main issues identified in the AI incidents

- Bias/Discrimination
- Safety
- Robustness/Security
- Privacy/Data governance
- Accountability
- Other human rights
- Transparency
Step 2: Automatic identification of incidents

Examples of automating AI incidents identification in real time using news articles

- **Probability: 0.7305 AI INCIDENT**
- **Probability: 0.3705 NOT AI INCIDENT**
AI Incidents Monitor (AIM) - Prototype

Search for a concept or keyword
Enter topics or keywords

Type of search AND

Country Select countries

SEARCH INCIDENTS

Evolution of incidents* by country

Evolution of articles by country

*An incident is a collection of one or more news articles covering the same event.
After crashes, can Tesla reassure consumers?

Date of first report: 2016-06-26
N.Articles: 101

Summary: Airlines, for instance, still need to reassure customers of their safety decades after jet engines and computers made flying the safest way to travel. But while plane passengers have to put their faith in the crew, self-driving cars require people for the first time to surrender control of their movement to machines, says Nidhi Kaira, a senior information scientist at the RAND Corporation. "When people are in a situation when they don't have control or they don't understand as much how a machin..."
Microsoft's A.I. bot Tay just wants to chat

Date of first report: 2016-03-22
N. Articles: 646

Summary: A new A.I.-powered chat bot called Tay.ai is helping researchers learn more about humans' ability to converse, reports Tech Crunch. Microsoft unveiled the bot, which uses various applications such as Twitter, Kik and GroupMe to assess conversation and social interaction of 18 to 24 year olds. Users can engage in conversation with Tay — as it is known for short — as well as ask a joke, play games and tell stories. The more users play with Tay, the more interactive it becomes. Tay was put toge...
Microsoft's A.I. bot Tay just wants to chat

Date of first report: 2016-03-22

N. Articles: 646

Summary: A new A.I.-powered chatbot called Tay.ai is helping researchers learn more about humans' ability to converse, reports Tech Crunch. Microsoft unveiled the bot, which uses various applications such as Twitter, Kik and GroupMe to assess conversation and social interaction of 18 to 24 year olds. Users can engage in conversation with Tay -- as it is known for short -- as well as ask a joke, play games and tell stories. The more users play with Tay, the more interactive it becomes. Tay was put toge...

FaceApp removes racial filters after instant backlash

Date of first report: 2017-08-08

N. Articles: 37

Summary: Provided by Huffington Post. You would've thought that after a series of similar PR fails (see Snapchat's Bob Marley filter), companies would have learned by now not to try and change a person's race using technology. While we're all keen to see how selfie-editing apps can make us look like an elderly pensioner or with fluorescent pink hair, we're definitely not here for racial filters. And that is exactly what popular FaceApp, has been accused of doing after releasing a software update on We...

Facebook is trying to eliminate bias by getting rid of humans

Date of first report: 2016-08-25

N. Articles: 51

Summary: Facebook will no longer employ humans to write descriptions for items in its Trending section, which attracted controversy over allegations of political bias in May. Topics appearing in the Trending section will now appear solely as a short phrase...
FaceApp, the popular photo-filter application, said it is removing new "ethnicity filters" after angry users condemned the update as racist.

The feature prompted users to alter selfies with "black," "Indian" and "Asian" filters. It gained widespread attention this week as social media lit up with comments that called the feature racist and harmful.

FaceApp’s removal of the filters comes amid a broader conversation about the use of ethnicity-based filters in technology. Such tools have raised concerns about perpetuating stereotypes and reinforcing biases.
AI Incidents Monitor (AIM)

Next steps

• Automate AI incidents classification into the criteria of the OECD classification framework, using natural language processing

• Differentiate between incidents (including serious incidents) and other hazards (including near misses) to align with incident definition work

• Develop interactive visualisations to show trends by AI Principle, industry, severity, etc.

• Increase usability of the platform and integrate into OECD.AI

• Develop process to enable direct submissions