11/12/2024

IST 692: RESPONSIBLE AI

PROJECT 1

JOVITA ANDREWS

1. Introduction

This report examines Microsoft-related incidents documented in the AIAAIC Repository, focusing on identifying recurring patterns, harms, and ethical implications associated with these incidents. By analyzing quantitative and qualitative aspects of the data, we aim to provide the company with insights into potential risks and ethical considerations when adopting AI technologies like Microsoft Copilot.

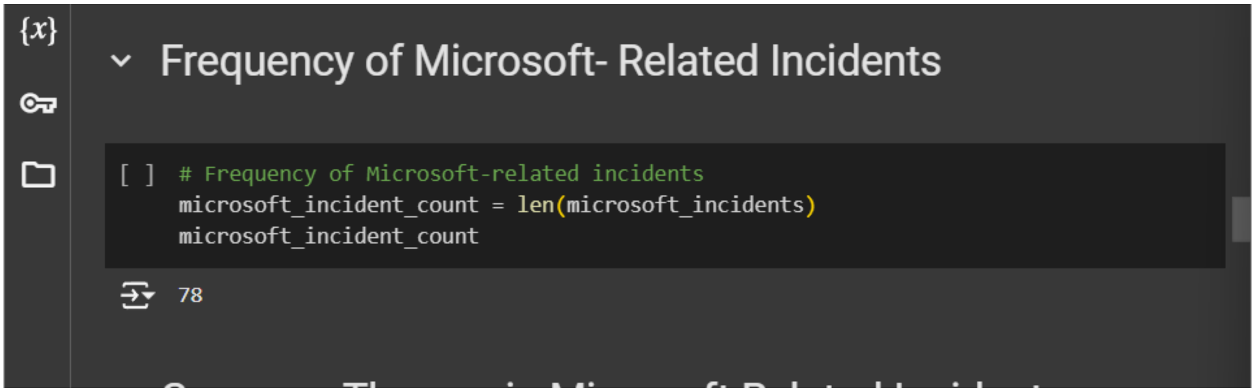
2. Data Card: Data Transparency Review of the AIAAIC Repository

The AIAAIC Repository is structured to enhance transparency in AI by documenting incidents involving ethical concerns across various AI systems and organizations.

Data Card for AIAAIC Repository:

- **Dataset Name**: AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) Repository
- **Version**: Beta (launched June 2019)
- **Purpose**: To increase accountability and awareness by cataloging AI-related incidents, especially those with ethical or societal impacts.
- **Key Metadata Fields**:
  - **Incident ID**: Unique identifier for each incident.
  - **Headline**: A summary or title of the incident.
  - **Type**: Categorizes the event as either an “Incident” or “Issue.”
  - **Country**: Geographic location impacted by the incident.
  - **Sector**: Industry affected, such as finance or healthcare.
  - **Developer/Deployer**: Organizations responsible for developing and deploying the AI system.
  - **Issues**: Ethical issues related to the incident (e.g., privacy, security, bias).
  - **External/Internal Harms**: Types of harm experienced by stakeholders within and outside the deploying organization.
  - **Legal/Regulatory Outcomes**: Any legal actions or regulatory responses related to the incident.
- **Limitations**:
  - **Coverage Bias**: Some incidents might be omitted, especially those without public reporting.
  - **Update Frequency**: Incidents are updated periodically, which may delay the inclusion of recent cases.

3. Frequency and Quantitative Analysis of Microsoft-Related Incidents

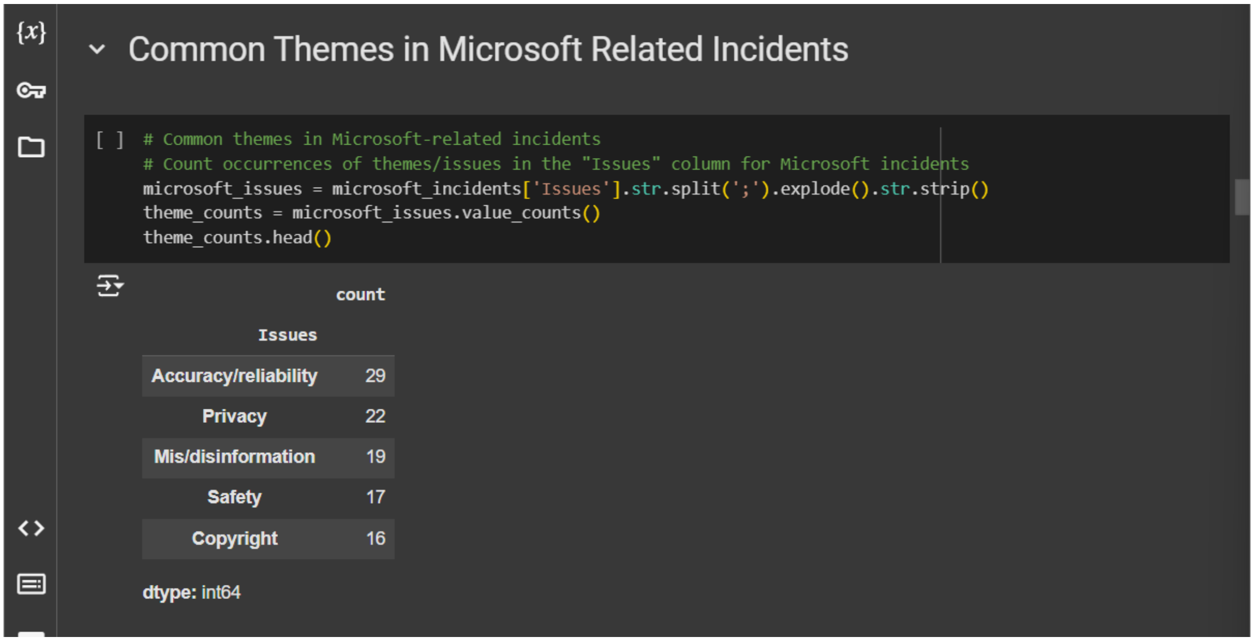


Frequency of Microsoft- Related Incidents

```
# Frequency of Microsoft-related incidents
microsoft_incident_count = len(microsoft_incidents)
microsoft_incident_count
```

```
78
```

Using the AIAAIC Repository data, we found that **78 incidents** are associated with Microsoft. This frequency is significant and indicates a need for scrutiny when deploying Microsoft AI systems.



Common Themes in Microsoft Related Incidents

```
# Common themes in Microsoft-related incidents
# Count occurrences of themes/issues in the "Issues" column for Microsoft incidents
microsoft_issues = microsoft_incidents['Issues'].str.split(';').explode().str.strip()
theme_counts = microsoft_issues.value_counts()
theme_counts.head()
```

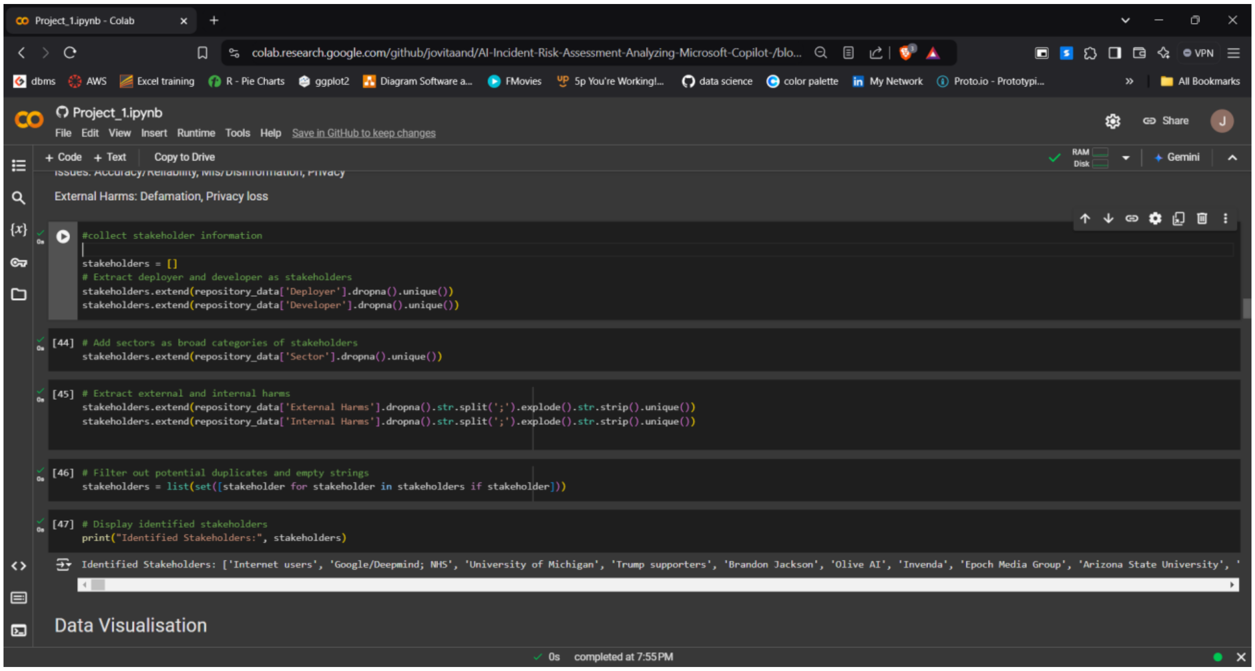
```
                       count
Issues
Accuracy/reliability      29
Privacy                   22
Mis/disinformation        19
Safety                    17
Copyright                 16

dtype: int64
```

To gain deeper insights, we examined the frequency of themes and harms within Microsoft incidents. **Top incident themes** include:

- **Accuracy/Reliability**: Present in 29 incidents, reflecting issues where Microsoft AI systems produced incorrect or unreliable results, raising concerns about the dependability of these systems.
- **Privacy Violations**: Documented in 22 incidents, primarily due to unauthorized data access or usage, indicating recurring privacy risks.
- **Misinformation**: Appearing in 19 incidents, particularly involving Microsoft's language models and tools like Copilot, which have generated inaccurate or misleading information.

- **Bias/Discrimination**: Bias issues affecting marginalized communities, are documented, highlighting the need for fairness and inclusivity in AI.
- **Environmental Concerns**: Incidents linked to Microsoft's data centers suggest environmental impacts due to high energy usage, contributing to carbon emissions.

4. Stakeholder Analysis



Issues: Accuracy/Reliability, Mis/Disinformation, Privacy

External Harms: Defamation, Privacy loss

```
#collect stakeholder information

stakeholders = []
# Extract deployer and developer as stakeholders
stakeholders.extend(repository_data['Deployer'].dropna().unique())
stakeholders.extend(repository_data['Developer'].dropna().unique())
```

```
[44] # Add sectors as broad categories of stakeholders
     stakeholders.extend(repository_data['Sector'].dropna().unique())
```

```
[45] # Extract external and internal harms
     stakeholders.extend(repository_data['External Harms'].dropna().str.split(';').explode().str.strip().unique())
     stakeholders.extend(repository_data['Internal Harms'].dropna().str.split(';').explode().str.strip().unique())
```

```
[46] # Filter out potential duplicates and empty strings
     stakeholders = list(set([stakeholder for stakeholder in stakeholders if stakeholder]))
```

```
[47] # Display identified stakeholders
     print("Identified Stakeholders:", stakeholders)
```

```
Identified Stakeholders: ['Internet users', 'Google/Deepmind; NHS', 'University of Michigan', 'Trump supporters', 'Brandon Jackson', 'Olive AI', 'Invenda', 'Epoch Media Group', 'Arizona State University', '
```

Data Visualisation

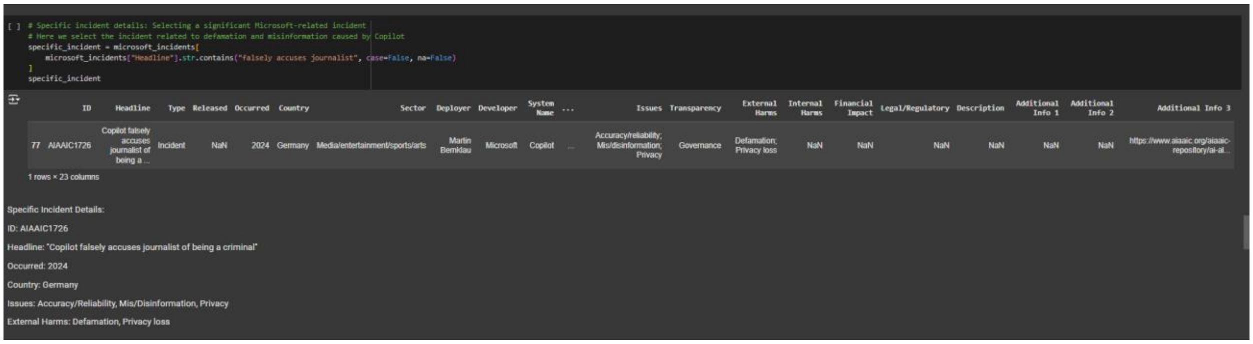
✓ 0s completed at 7:55 PM

Analyzing incidents with a stakeholder-focused approach reveals a range of impacted parties:

- **Individuals**: End-users often face privacy breaches, reputational harm, and exposure to misinformation.
- **Communities**: Community stakeholders may also face indirect environmental harms from AI operations.
- **Businesses**: Legal and regulatory repercussions may arise if AI deployments violate data protection laws.
- **Environment**: Data center energy consumption impacts the environment, contributing to carbon emissions, which indirectly affects society and future generations.

This stakeholder analysis underscores the importance of sociotechnical considerations in Microsoft AI deployments, revealing that technical issues in AI can have broad social ramifications.

5. Qualitative Analysis of a Significant Incident



```
[ ] # Specific incident details: Selecting a significant Microsoft-related incident
    # Here we select the incident related to defamation and misinformation caused by Copilot
    specific_incident = microsoft_incidents[
        microsoft_incidents["Headline"].str.contains("falsely accuses journalist", case=False, na=False)
    ]
    specific_incident
```

	ID	Headline	Type	Released	Occurred	Country	Sector	Deployer	Developer	System Name	...	Issues	Transparency	External Harms	Internal Harms	Financial Impact	Legal/Regulatory	Description	Additional Info 1	Additional Info 2	Additional Info 3
77	AIAAIC1726	Copilot falsely accuses journalist of being a ...	Incident	NaN	2024	Germany	Media/entertainment/sports/arts	Martin Bernklau	Microsoft	Copilot	...	Accuracy/reliability; Mis/disinformation; Privacy	Governance	Defamation; Privacy loss	NaN	NaN	NaN	NaN	NaN	NaN	https://www.aiaaic.org/aiaaic-repository/ai-al...

1 rows × 23 columns

```
Specific Incident Details:
ID: AIAAIC1726
Headline: "Copilot falsely accuses journalist of being a criminal"
Occurred: 2024
Country: Germany
Issues: Accuracy/Reliability, Mis/Disinformation, Privacy
External Harms: Defamation, Privacy loss
```

Incident Example: Microsoft Copilot Misinformation Incident

- **ID**: AIAAIC1726
- **Headline**: "Copilot falsely accuses journalist of being a criminal"
- **Year**: 2024
- **Country**: Germany
- **Incident Description**: Microsoft's Copilot, an AI-powered content generation tool, produced content that inaccurately linked a journalist to criminal activities. This incident was widely reported, bringing public attention to the reliability of AI-generated information.
- **Stakeholders Impacted**: The journalist (individual stakeholder) faced reputational harm and potential career setbacks. Broader societal stakeholders were affected due to increased distrust in AI-generated content.
- **Company Response**: Microsoft acknowledged the incident, explaining Copilot's limitations but did not implement specific corrective actions, leaving unresolved concerns about the tool's reliability.

This incident highlights the potential harm of misinformation and raises questions about Microsoft's accountability. Without concrete corrective actions, similar incidents may continue to occur.

6. Ethical Implications Using Frameworks

- **Consequentialism**: This framework evaluates the outcomes of actions. Although Microsoft Copilot aims to enhance productivity, the risk of misinformation resulting in reputational harm must be weighed against productivity gains.
- **Deontology**: Deontological ethics focuses on upholding duties and respecting rights. Microsoft has a duty to ensure accuracy and fairness, especially when deploying tools that can impact personal reputations.
- **Virtue Ethics**: This approach emphasizes moral character and virtues, such as honesty and accountability. By acknowledging but not addressing Copilot's limitations, Microsoft may be seen as lacking in accountability.

- **Sociotechnical Perspective**: AI incidents show that ethical AI is both a technical and social issue. Microsoft's AI tools affect not just the technical systems in which they operate but also the social contexts and communities interacting with these tools.

Recommendations

Based on a thorough analysis of Microsoft-related incidents, we recommend the following strategies for the company to ensure ethical and responsible AI adoption:

1. **Establish Clear Privacy and Data Usage Policies**: With privacy being a recurring issue, it's critical to set and enforce data privacy policies. Ensure compliance with relevant data protection regulations, and clearly communicate data usage practices to users.
2. **Implement Routine AI Audits**: Regular audits should be conducted to monitor and mitigate issues such as misinformation and bias. This can help identify and resolve emerging issues before they impact stakeholders.
3. **Conduct Bias Assessments**: To prevent discriminatory outcomes, incorporate regular bias assessments in AI deployment. Ensure models are designed and trained with fairness and inclusivity in mind.
4. **Promote Transparency and Accountability**: Develop a clear process for acknowledging and addressing AI incidents. Transparent communication about AI limitations can help manage user expectations and mitigate reputational harm.
5. **Consider Environmental Impacts**: If data centers are part of CNY Solutions' operations, prioritize energy-efficient practices to reduce the environmental impact, supporting sustainable AI practices.

Some visualizations were useful to make this deduction is as follows:

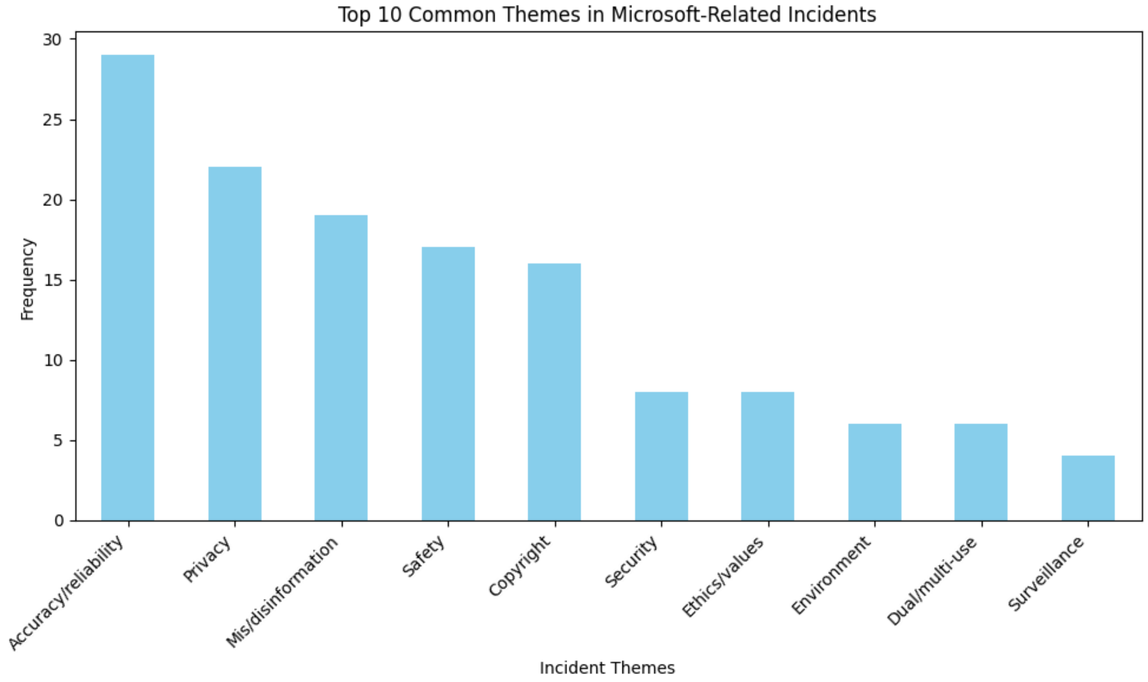


Fig: Top 10 Common Themes in Microsoft – Related Incidents

This bar chart shows the most frequent issues in incidents: Accuracy/reliability, Privacy, Misinformation, Safety, Copyright, Security, Ethics/ Values, Environment, Dual and Surveillance.

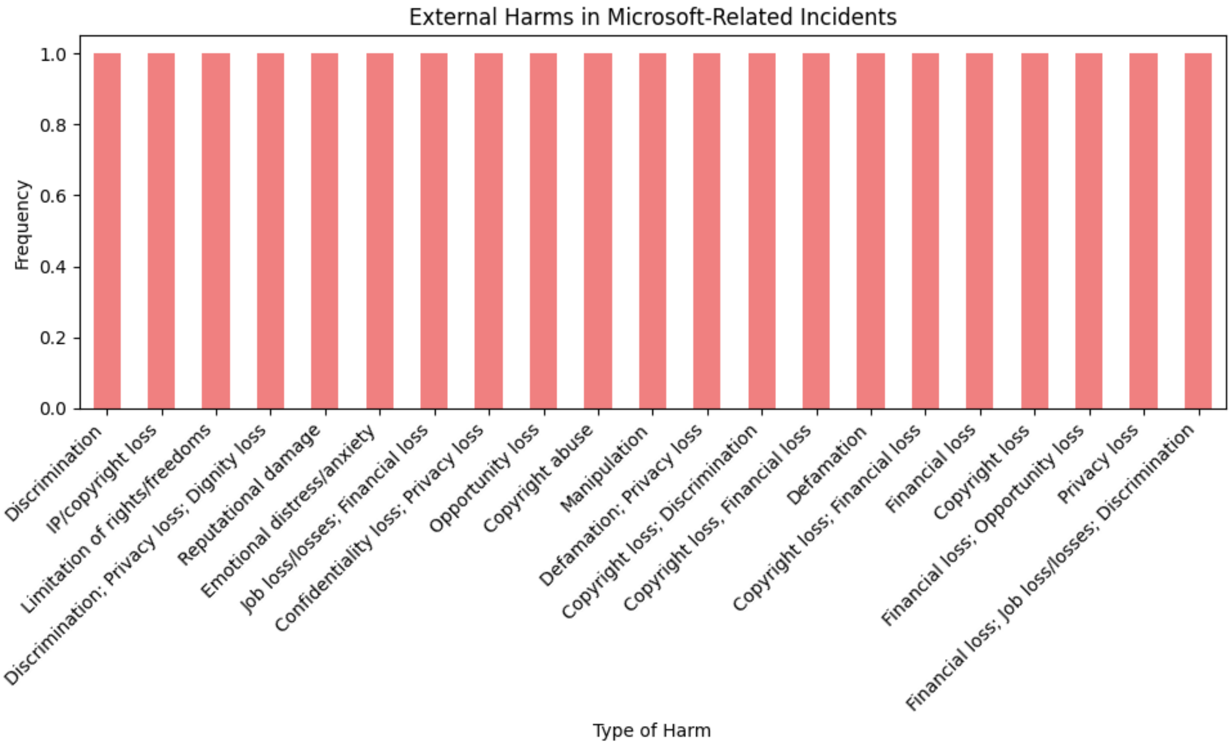


Fig: Type of Harms

This visualization displays the type of external harm that are caused my Microsoft- co pilots has caused. The list varies from Discrimination, Copyright issues, Limitation of rights, Reputational losses etc. that must be addressed by the company.

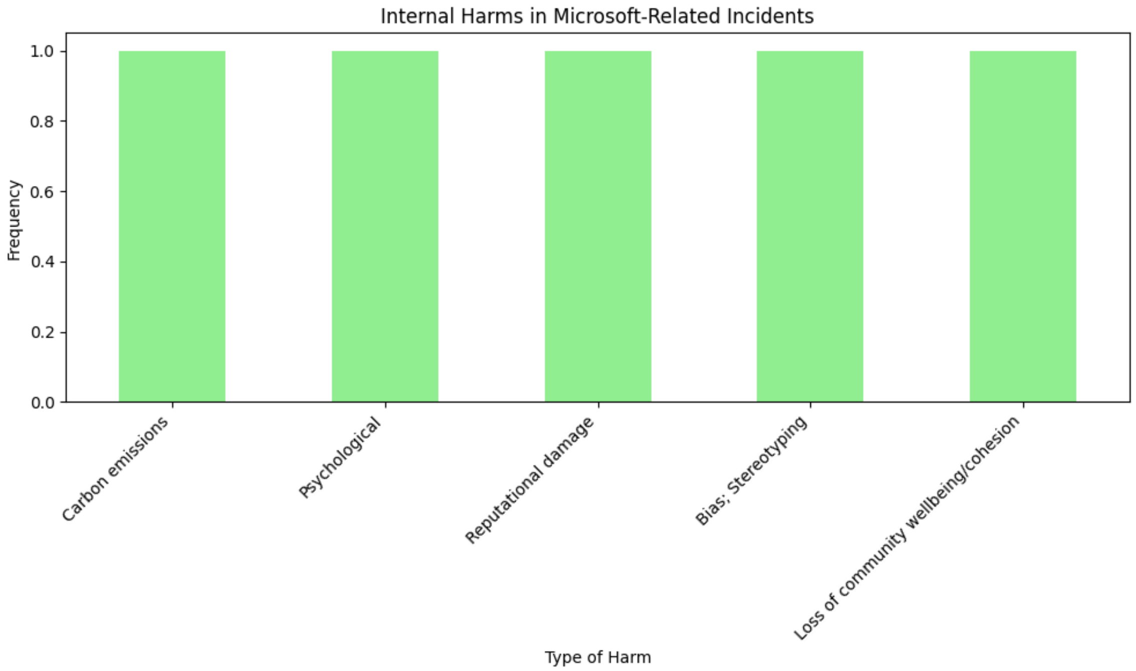


Fig: Type of Harms

This bar graph is for internal harms in Microsoft co pilot incidents. Reputational damage and psychological impact are the most reported, reflecting the organizational risks of deploying AI without robust safeguard.

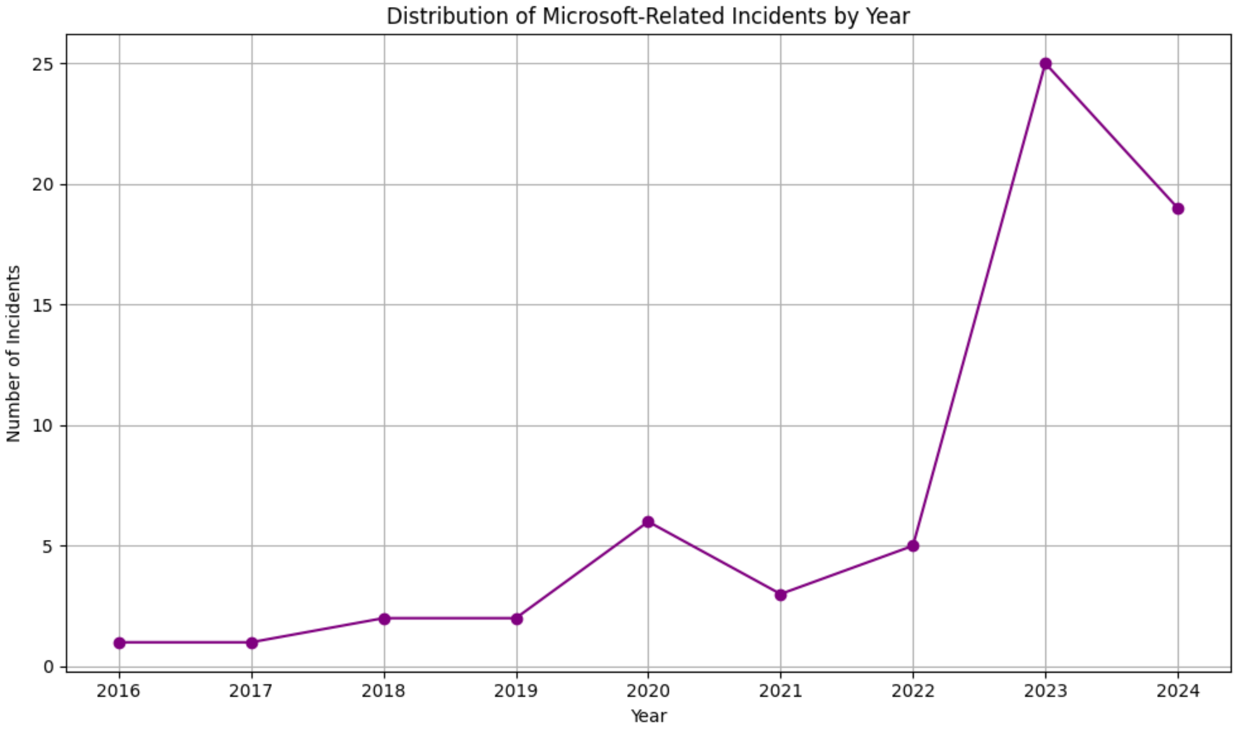


Fig: Distribution of Microsoft – Related Incidents by Year

This line chart shows the trend of Microsoft-related incidents over the years. Notable increases in recent years suggest a growing number of incidents as AI adoption expands, emphasizing the importance of updated ethical frameworks and safety protocols.

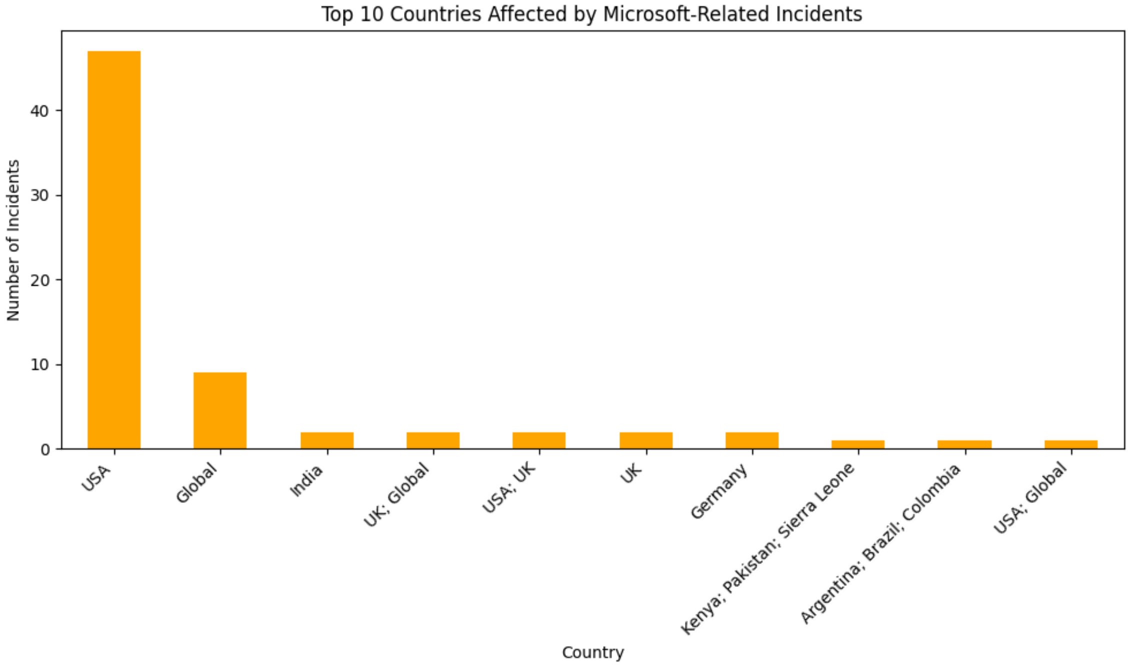


Fig: Top 10 Countries Affected by Microsoft – Related Incidents

This bar chart illustrates the geographic distribution of incidents, with certain countries experiencing higher numbers of issues. It reflects the need for location-specific policies and guidelines tailored to regional data protection and privacy laws.

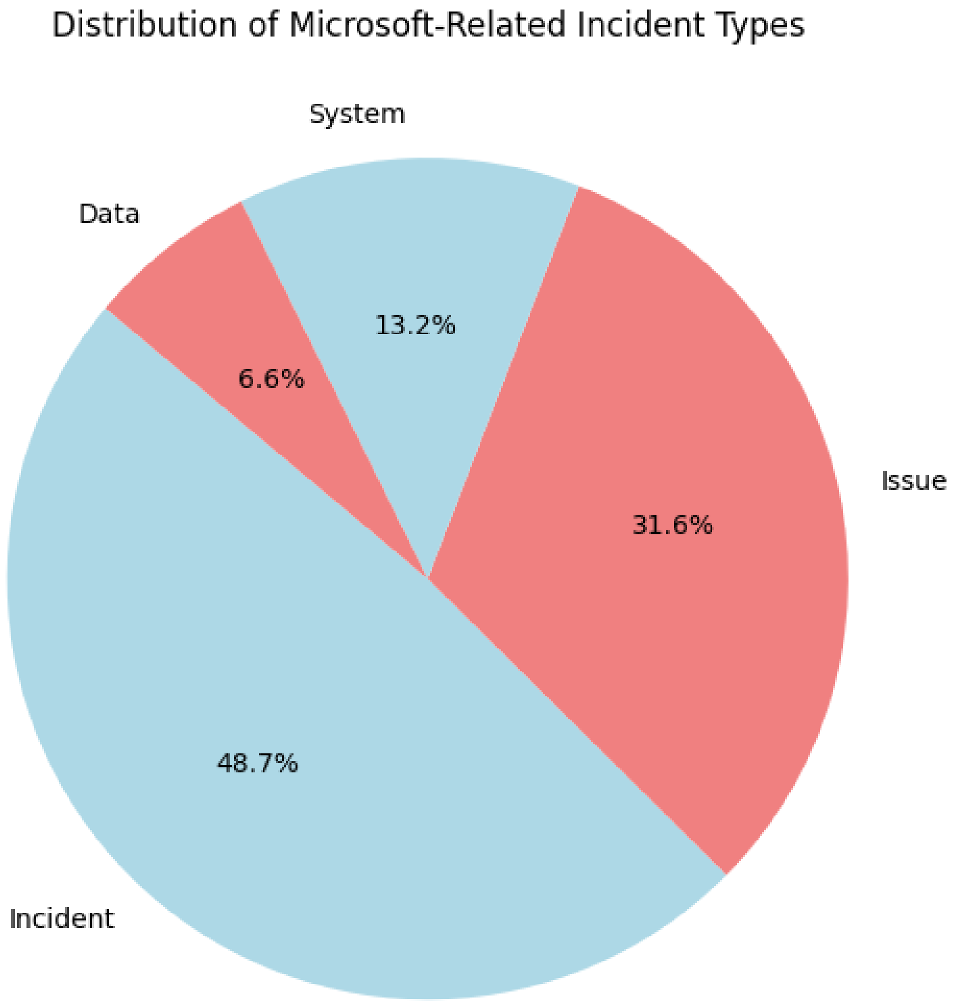


Fig: Distribution of Microsoft – Related Incident Types

This pie chart categorizes incidents by type, showing the balance between Incidents and Issues. It provides insight into the nature of reported cases, helping stakeholders understand whether these involve persistent issues (like bias or reliability) or singular incidents (like specific security breaches).

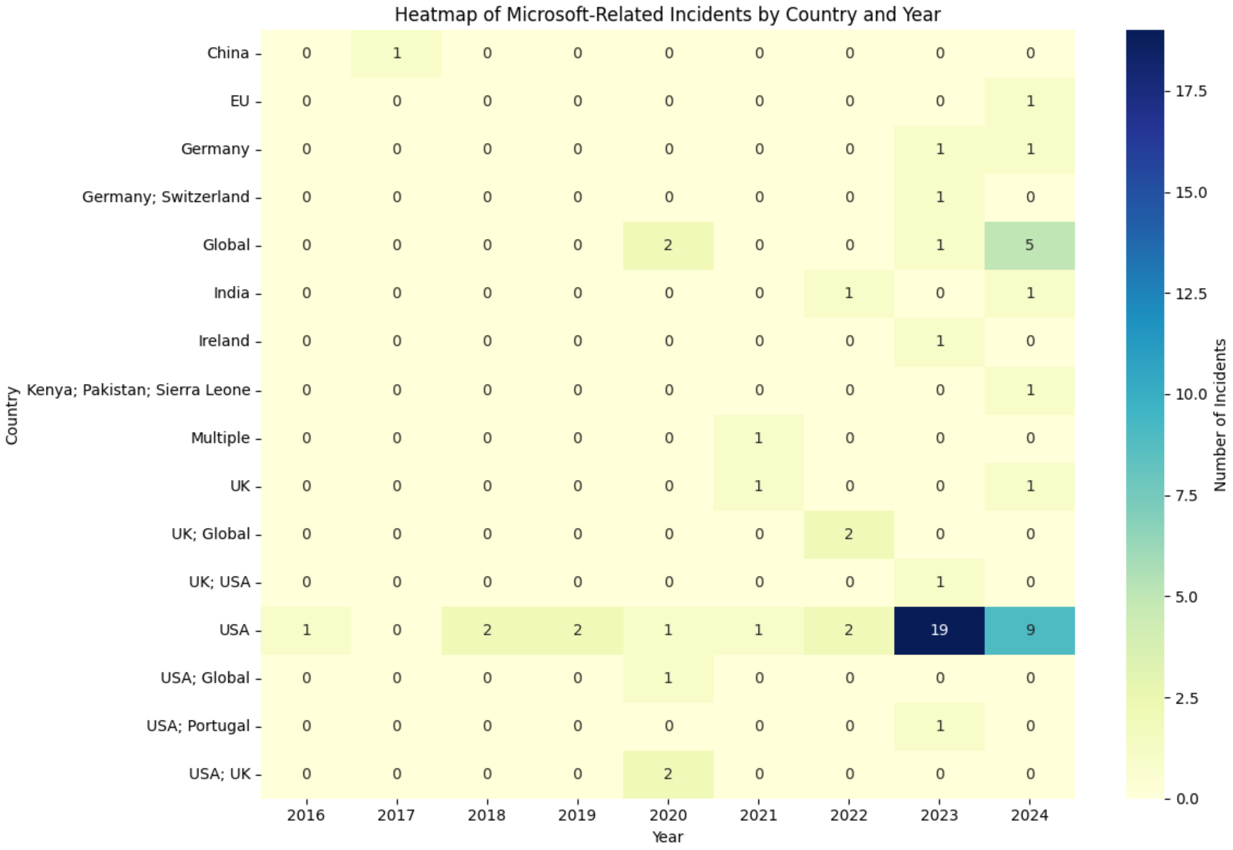


Fig: Heatmap of Microsoft – Related Incidents by Country and Year

This is a heatmap of all the Microsoft copilot related incidents based on country and year. The visualization and temporal distribution of Microsoft – related incidents, showing which regions and years experienced higher activity. There is a noticeable increase in incidents from 2020 onwards, particularly for the US. Most countries have few to no recorded incidents per year, suggesting that Microsoft-related incidents are concentrated in certain regions or that reporting may less frequent outside major areas.

Data Modeling: Model Performance Report

1. Logistic Regression

Before Feature Engineering

- **Accuracy:** ~81%
- **Class 0 (Human-Generated):**
  - **Precision:** 0.82
  - **Recall:** 1.00
  - **F1-Score:** 0.90
- **Class 1 (AI-Generated):**

    - **Precision:** 1.00
    - **Recall:** 0.23
    - **F1-Score:** 0.37
- **Key Insight:** Performed well for human-generated content but struggled to identify AI-generated content due to low recall.

After Feature Engineering

- **Accuracy:** ~83.57%
- **Class 0 (Human-Generated):**
    - **Precision:** 0.91
    - **Recall:** 0.87
    - **F1-Score:** 0.89
- **Class 1 (AI-Generated):**
    - **Precision:** 0.62
    - **Recall:** 0.70
    - **F1-Score:** 0.65
- **Key Insight:** Adding metadata features improved the model's ability to detect AI-generated content, with notable gains in recall for Class 1.

2. Support Vector Machine (SVM)

Before Feature Engineering

- **Accuracy:** ~84.12%
- **Class 0 (Human-Generated):**
    - **Precision:** 0.89
    - **Recall:** 0.91
    - **F1-Score:** 0.90
- **Class 1 (AI-Generated):**
    - **Precision:** 0.66
    - **Recall:** 0.59
    - **F1-Score:** 0.62
- **Key Insight:** SVM provided a balanced performance but still missed many AI-generated instances.

After Feature Engineering

- **Accuracy:** ~85.5%
- **Class 0 (Human-Generated):**
  - **Precision:** 0.91
  - **Recall:** 0.90
  - **F1-Score:** 0.90
- **Class 1 (AI-Generated):**
  - **Precision:** 0.71
  - **Recall:** 0.67
  - **F1-Score:** 0.69
- **Key Insight:** The metadata features further improved SVM's ability to identify AI-generated content, enhancing both precision and recall for Class 1.

3. Random Forest

Before Feature Engineering

- **Accuracy:** ~86.63%
- **Class 0 (Human-Generated):**
  - **Precision:** 0.86
  - **Recall:** 0.98
  - **F1-Score:** 0.92
- **Class 1 (AI-Generated):**
  - **Precision:** 0.88
  - **Recall:** 0.46
  - **F1-Score:** 0.61
- **Key Insight:** High precision but low recall for Class 1 indicated a conservative approach toward labeling AI-generated content.

After Feature Engineering

- **Accuracy:** ~88.2%
- **Class 0 (Human-Generated):**
  - **Precision:** 0.90

    - **Recall:** 0.97
    - **F1-Score:** 0.93
- **Class 1 (AI-Generated):**
    - **Precision:** 0.80
    - **Recall:** 0.58
    - **F1-Score:** 0.67
- **Key Insight:** Improved recall and balanced performance, showing the model became slightly more assertive in labeling AI-generated content while retaining high accuracy for human-generated content.

4. Ensemble Model (SVM + Random Forest)

Before Feature Engineering

- **Accuracy:** ~85.79%
- **Class 0 (Human-Generated):**
    - **Precision:** 0.87
    - **Recall**: 0.96
    - **F1-Score:** 0.91
- **Class 1 (AI-Generated):**
    - **Precision:** 0.78
    - **Recall:** 0.50
    - **F1-Score**: 0.61
- **Key Insight:** Balanced performance but missed some AI-generated instances due to moderate recall for Class 1.

After Feature Engineering

- **Accuracy:** ~88.5%
- **Class 0 (Human-Generated):**
    - **Precision:** 0.92
    - **Recall:** 0.94
    - **F1-Score:** 0.93
- **Class 1 (AI-Generated):**
    - **Precision:** 0.75

    - **Recall:** 0.64
    - **F1-Score**: 0.69
- **Key Insight:** The ensemble model combined the strengths of both SVM and Random Forest, yielding a robust balance between precision and recall for both classes. The added features improved recall for AI-generated content, while maintaining high performance for human-generated content.

Further Analysis:

For further analysis I would:

- Perform grid search on hyperparameters like regularization strength to further refine model's performance.
- Experiment with other models to see if they yield better results,
- Use feature importance or SHAP values to identify the metadata features that are contributing most to the model's performance.

Conclusion:

This refined report provided a comprehensive exploration of Microsoft-related incidents in the AIAAIC database. By examining themes, harms, stakeholders, and ethical implications, we identified key areas of risk and provided actionable recommendations. Through careful adherence to privacy, transparency, and bias mitigation practices, the company can adopt AI responsibly, aligning with ethical standards and minimizing potential harm to stakeholders.