Dangers of Locally hosted AI Foundational Models: A SWE Ethics perspective and solution proposal

A proposed research for the course ITSE414 Software Engineering Ethics
for the Student Sanad Al-Arousi 2181801442 for Fall 2024 DRAFT1

Table of Contents:

Dangers of Locally hosted AI Foundational Models: A SWE Ethics perspective and solution proposal....1
1.0 Introduction....2
1.1 Importance of the matter in SWE Ethics....2
1.2 Goal....2
2.0 Description....3
2.1 main concepts....3
2.1.1 Kinds of Models:....3
2.1.2 Flaws and Errors of Models....3
2.1.3 Attacks on Models....3
● Jailbrakingg and bypassing attacks:....4
3.0 Ethics challenges....4
3.1 list of challenges....4
3.2 impact on individuals, communities, economies and states....4
3.3 AI Ethics....4
4. Study Cases....5
4.1 Center for Security and Emerging Technology CSETv1 AI Harm Taxonomy:....5
4.1.1 Harm Distribution Basis:....5
4.1.2 Sector of Deployment:....6
4.2 Goals, Methods, and Failures (GMF):....10
Figure 4 GMF Known AI Goal Incidents chart....10
Figure 5 GMF Know AI Technical Failure chart....10
4.2.2 Known AI Technology....11
Figure 6 GMF Known AI Technology chart....11
4.2.3 Known AI Technical Failure....11
5. Proposed Solutions....12
5.1 AI Governance:....12
● 5.1.1 EU Ethics Guidelines for Trustworthy AI: which states that for an AI to be “Trustworthy”:....12
● 5.1.2 AI Governance Framework:....12
● 5.1.3 IEEE Global Initiative on Ethics of Autonomous and Intelligent systems:....12
● 5.1.4 Montreal Declaration of Responsible AI....13
● 5.1.5 NIST AI Risk Management Framework:....13
● 5.1.6 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence:....13
5.1.6.1 AI Safety and Security....13
5.1.6.2 Trustworthy AI....13
5.1.6.3 Privacy and Civil Rights....13
5.1.6.4 Supporting Innovation....14
5.1.6.5 National Security and Defense....14
5.1.6.6 Global Leadership and Collaboration....14
5.1.6.7 Workforce Development and Education....14
5.1.6.8 Environmental Sustainability....15
5.2 Enforced Rail-guards....15
5.3 Embedding the Unique key for each instance of a model....15
6. Conclusion....16
7. References....16

1.0 Introduction

The rapid advances in AI, particularly foundational AI models, have created unprecedented opportunities and challenges. While these models have the potential to revolutionize various sectors, whether by increasing the productivity of professionals or reducing the amount of repetitive work, hosting them locally—which is less expensive— and very possible with various kinds and sizes of AI FMs, raises significant ethical concerns that require careful considerations. This report aims to examine the risks of locally hosted foundational AI models from a software engineer's ethics perspective, focusing on their impact on individuals and society.

1.1 Importance of the matter in SWE Ethics

In software engineering ethics perspective, the deployment of such systems makes breaching the "Code of Ethics" very possible, untrackable, and has deep impact on various parties.
For such dangerous depths and widths of effects, the governance of AI has been the interest of many governing and compliance bodies since ever the sector and the models gained great public interest and showed strong capabilities.

1.2 Goal

This report aims to raise awareness on the ethical challenges of AI FMs, as the code of ethics isn't enough for governance, yet handling the matter from that perspective is the starting point of any legislation, framework or controlling system.

1. Increasing awareness about the harmful capabilities of AI FMs
2. Raising the awareness on the proposed and set on legislations, frameworks and initiatives
3. Analysing the various kinds of proposed governance systems
4. Proposing a solution that doesn't hinder the creativity, academic research and entrepreneurship efforts by small companies and limited funding bodies.

2.0 Description

2.1 main concepts

2.1.1 Kinds of Models:

2.1.1.1. Foundational Artificial Intelligence Models: wide scope of dataset and capabilities not aimed for a specific topic or sector

2.1.1.2. Local Hosting of Models: deployment of these systems on private Servers or PCs with above average capabilities, owned or managed by private, government or anonymous entities

2.1.1.3. Closed Source FM: General purpose model with its customization characteristics unavailable for modification by none other than the developer

2.1.1.4. Open Source FM: General purpose model either partially or fully customizable
2.1.1.5. Instruct-FM: a model that is interactable through a UI to usually do Text generation
2.1.1.6. Fine-Tuned FM: general purpose model that is trained on more specific domain data-set, such as Accounting models or coding models, in an effort to improve its accuracy in that specific domain or topic.
2.1.1.7. AI/ML Cloud Service: A cloud service that provides Models, hosting, running or development environments for various parties.

2.1.2 Flaws and Errors of Models

- Categorization errors
- False Positives and True Negatives
- Overgeneralization
- Over-Focus on the details of the training dataset
- Focusing on metrics and variables that aren't important or of an impact
- Sensitivity to specific inputs
- Performance Plunge in some real life matters

which is usually rare in deployed Foundational Models, and are noticed in development stage models.
Once these Flaws and Errors are managed by the Engineers, Attacks on Models generates more risk and might make the model commit the said Flaws and Errors.

2.1.3 Attacks on Models

- Hostile goal input
- Poisoning: infecting dataset or inputs of a prompt with defected, incorrect or illegal/unethical contents.
- Data Retrieval attacks
- Data Confirmation attacks
- Training weights attacks
- Denial of service through exhaustion

- Jailbrakingg and bypassing attacks:
    - Code injection
    - Noise
    - Manipulation through context
    - Manipulation through sequence of prompts
    - Manipulation through
- Model misinformation
- Targeted weaknesses attacks

3.0 Ethics challenges

Ethical challenges in the AI Models Field is very complicated, and leans more to compliance and controlling frameworks rather than approaches that rely on the ethical endeavour of the host, developer or user of the model.
as the industry has reached a level of deployability that a normal user with no technical skills can download it, host it on a local server or environment and run it, yielding any direct controlling mechanisms inefficient or null.
Once we lose the control of hosts, the only direct influence on the models left is the closure of its weights, and customization properties, and while this approach is throttling the dangers of the models by limiting its capabilities to the goals and to the scope of its development, it hinders academic researchers, entrepreneurs and developers ability to utilize that model in their work, life or applications.

3.1 list of challenges

- 3.1.1 Bias and prejudice against a set of people
- 3.1.2 Accountability
- 3.1.3 Misuse
- 3.1.4 Misinformation
- 3.1.5 Human or User harm

3.2 impact on individuals, communities, economies and states

- 3.2.1 Physical Harm
- 3.2.2 Mental and Psychological harm
- 3.2.3 Fraud
- 3.2.4 Prejudice

3.3 AI Ethics

an ecosystem of ethical standards and guardrails throughout all phases of an AI system's life cycle. Whilst it is close and almost similar to 'SWE Code of Ethics', yet it is based on different kind of incidents, impacts and scope.

4. Study Cases

"The AI Incident Database is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. "

The incident database, first founded by 'Sean McGregor', and is now managed in a participatory manner by persons and organizations contributing code, research, and broader impacts, offers great information, Taxonomies, and a full list, 850 till Dec 2024, of incidents about this matter and the sector of AI Ethics in general.

4.1 Center for Security and Emerging Technology CSETv1 AI Harm Taxonomy:

Characterizes AI incidents and classifies harms of relevance to the public policy community. AI harm has four elements which, once appropriately defined, enable the identification of AI harm. These key components serve to distinguish harm from non-harm and AI harm from non-AI harm. To be an AI harm, there must be:

1) an entity that experienced
2) a harm event or harm issue that
3) can be directly linked to a consequence of the behavior of
4) an AI system.

4.1.1 Harm Distribution Basis:

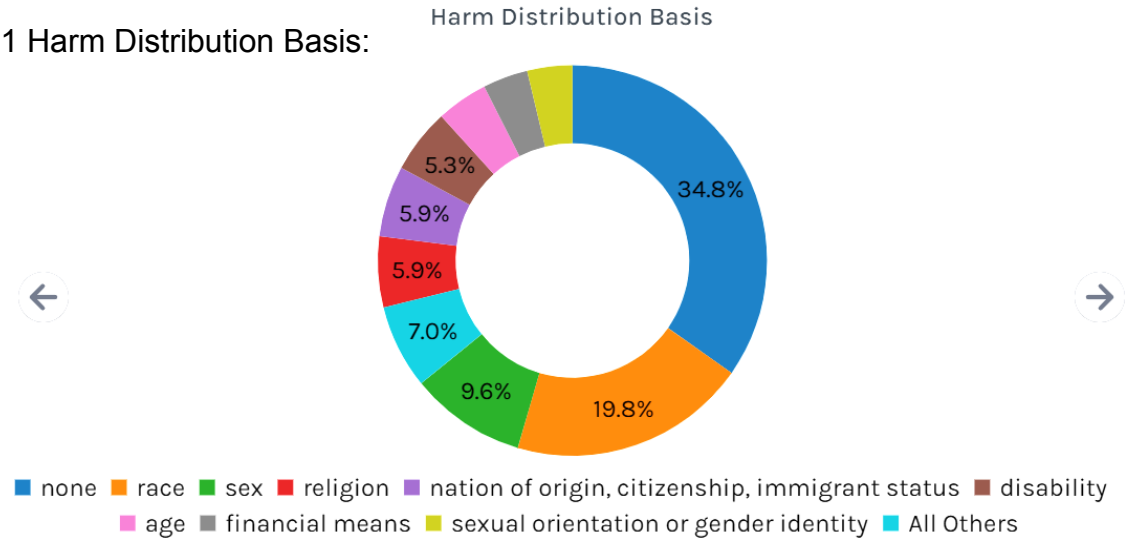


Figure 1 CSETv1 AI Harm Distribution 'differential treatment basis' Chart

Category	Count
race	37
sex	18
religion	11
nation of origin, citizenship, immigrant status	10
disability	10
sexual orientation or gender identity	7
age	7
financial means	6
geography	5
ideology	2
none	1
familial status (e.g., having or not having children) or pregnancy	1
other	
unclear	

Figure 2 AI Harm Distribution 'differential treatment basis' by numbers

4.1.2 Sector of Deployment:

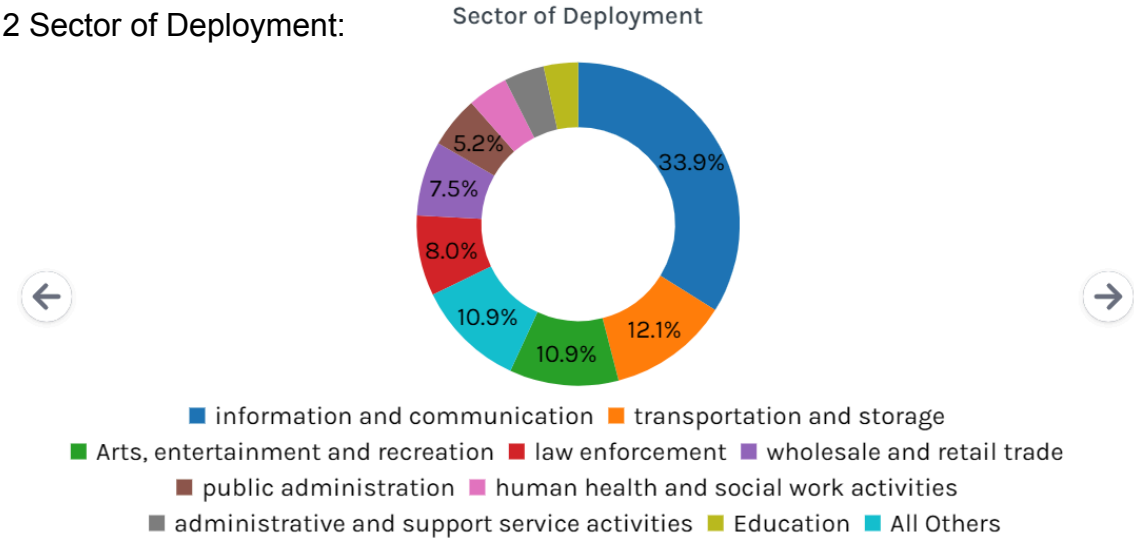


Figure 3 CSETv1 AI Harm incident Sector of Deployment chart

Category	Count
information and communication	59
transportation and storage	21
Arts, entertainment and recreation	19
law enforcement	14
wholesale and retail trade	13
public administration	9
human health and social work activities	7
administrative and support service activities	7
Education	6
accommodation and food service activities	5
professional, scientific and technical activities	4
financial and insurance activities	4
other	2
defense	1
real estate activities	1
other service activities	1
unclear	1
manufacturing	

Figure 4 CSETv1 AI Harm incident Sector of Deployment by numbers

From the data showcased above, 34.8% of AI Incidents that caused harm, the harm was not based on a specific reason, but mainly on Race, Sex and origine/race of individuals.
Incident from the DB based on Race:

1. **Incident ID**: 16 **Title**: Images of Black People Labeled as Gorillas **Date**: 2015-06-03 **Alleged Developer**: google, **Alleged Deployer**:google, **Group**: black-people
    a. **Description:**Google Photos image processing software mistakenly labelled a black couple as "gorillas."
    b. **Significance:**Google's image processing software mislabeled a black couple as "gorillas," demonstrating a clear failure in addressing **3.1.1 Bias and Prejudice Against a Set of People**. The inadequate dataset used in the algorithm caused significant harm to black communities.
    c. **Ethical Challenges:**
        - **3.1.1 Bias and Prejudice Against a Set of People:**
            - The training data lacked diversity, leading to offensive mislabeling.
        - **3.1.2 Accountability:**
            - Google's failure to conduct sufficient testing on diverse datasets highlights a lack of accountability in ensuring the system's accuracy and fairness.
    d. **Impacts 3.2:**
        - **3.2.2 Mental and Psychological Harm:**
            - The mislabeling caused emotional harm to individuals and groups, undermining trust in AI systems.

        - **3.2.4 Prejudice:**
            - Reinforced racial stereotypes, perpetuating societal discrimination against black people.
2. **Incident ID**: 49 **Title**: AI Beauty Judge Did Not Like Dark Ski, **Date**: 2016-09-05 **Alleged Deployer**: youth-laboratories, **Alleged Developer**: youth-laboratories, **Group**: people-with-dark-skin, Women
    a. **Description:** in 2016, after artificial intelligence software Beauty.AI judged an international beauty contest and declared a majority of winners to be white, researchers found that Beauty.AI was racially biased in determining beauty.
    b. **Significance:**Beauty.AI judged an international beauty contest, favoring white participants and marginalizing individuals with dark skin tones. This reflects the system's failure to address **3.1.1 Bias and Prejudice Against a Set of People** due to racially skewed training data.
    c. **Ethical Challenges:**
        - **3.1.1 Bias and Prejudice Against a Set of People:**
            - The AI system favored Eurocentric beauty standards, marginalizing people with darker skin tones.
        - **3.1.2 Accountability:**
            - The developers failed to ensure that the AI was trained on diverse datasets representing global beauty standards.
    d. **Impacts 3.2:**
        - **3.2.2 Mental and Psychological Harm:**
            - The outcome damaged the self-esteem of individuals with darker skin tones.
        - **3.2.4 Prejudice:**
            - Reinforced racial and cultural biases in defining beauty.
3. **Incident ID**: 118 **Date**: 2020-08-06 **Title**: OpenAI's GPT-3 Associated Muslims with Violence" **Alleged Deployer**: openai **Alleged Developer**: openai **Group**: muslims
    a. **Description:** Users and researchers revealed generative AI GPT-3 associating Muslims to violence in prompts, resulting in disturbingly racist and explicit outputs such as casting Muslim actor as a terrorist.
    b. **Significance:**GPT-3 linked Muslims to violence in its responses, reflecting **3.1.1 Bias and Prejudice Against a Set of People** embedded in its training data. This caused harm to the Muslim community and highlighted the risks of deploying AI without rigorous oversight.
    c. **Ethical Challenges:**
        - **3.1.1 Bias and Prejudice Against a Set of People:**
            - The model reproduced harmful stereotypes due to biased training data.
        - **3.1.2 Accountability:**
            - OpenAI's lack of robust safeguards to prevent such outputs points to insufficient accountability in monitoring and refining the system.
        - **3.1.5 Misuse:**
            - The ability for users to prompt the AI into generating harmful content raises concerns about its potential misuse.
    d. **Impacts (3.2):**
        - **3.2.1 Physical Harm:**

        - The content could be used maliciously, putting the Muslim community at risk of targeted violence.
    - **3.2.2 Mental and Psychological Harm:**
        - Harmful outputs caused emotional distress and perpetuated feelings of exclusion.
    - **3.2.4 Prejudice:**
        - Reinforced societal biases and systemic discrimination against Muslims.
4. **Incident ID**: 40 **Date**: 2016-05-23 **Title**: COMPAS Algorithm Performs Poorly in Crime Recidivism Prediction **Alleged Developer**: equivant **Alleged Deployer**: equivant **Group:** accused-people
    a. **Description:** Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a recidivism risk-assessment algorithmic tool used in the judicial system to assess likelihood of defendants' recidivism, is found to be less accurate than random untrained human evaluators.
    b. **Significance:**COMPAS, a recidivism prediction tool, demonstrated **3.1.1 Bias and Prejudice Against a Set of People** by disproportionately labeling black defendants as high-risk for reoffending, even when controlling for other factors. This highlights the dangers of relying on biased AI in high-stakes environments like the criminal justice system.
    c. **Ethical Challenges:**
        - **3.1.1 Bias and Prejudice Against a Set of People:**
            - The tool perpetuated racial bias due to flawed training data and unaddressed systemic bias in its design.
        - **3.1.2 Accountability:**
            - Developers failed to validate and monitor the algorithm adequately before deployment.
        - **3.1.4 Human or User Harm:**
            - The tool's erroneous predictions negatively affected the lives and legal outcomes of defendants.
    d. **Impact 3.2**:
        - **3.2.2 Mental and Psychological Harm:**
            - Black defendants experienced stigma and distress due to unjust labeling.
        - **3.2.4 Prejudice:**
            - Reinforced systemic racism within the judicial system.
        - **3.2.3 Fraud:**
            - Unreliable predictions misled legal professionals, undermining trust in AI.

5. **Incident ID**: 583 **-** 624
Date: 2023-06-07 **/** 2023-12-20
Title 1: Instagram Algorithms Allegedly Promote Accounts Facilitating Child Sex Abuse Content
Title 2: Child Sexual Abuse Material Taints Image Generators
Alleged Developer: Meta, Instagram **-** Various People, Various Organizations
Alleged Deployer: Meta, Instagram **-** 'Laion + Stable Diffusion + Anderegg of Holmen'
Group: Children, General Public, Minors, Teenagers

a. **Description**: An investigation disclosed that Instagram's recommendation algorithms are promoting accounts that facilitate and sell child sexual abuse material (CSAM). The study, conducted by The Wall Street Journal and researchers at Stanford University and the University of Massachusetts Amherst, indicates that Instagram's algorithms not only allow for the discovery of such accounts through keyword searches but also actively recommend them to users within the network. The issue is especially concerning given Instagram's popularity among teenagers.

   a person was arrested by the FBI with 13K pictures of CSAM generated by a model that he hosted on his machine. The LAION-5B dataset (a commonly used dataset with more than 5 billion image-description pairs) was found by researchers to contain child sexual abuse material (CSAM), and that Stable Diffusion version 1.5 was trained on it. And the model only required 4GB of VRAM and no more than 20GB of Storage to run.

b. **Significance**: The study highlights the dangers of unregulated and unmonitored AI systems in social media, particularly in platforms popular among minors. It also underscores the need for stricter oversight and accountability in the deployment of such algorithms.

   The incident highlights the critical importance of rigorous data curation and the potential for significant harm when large datasets are used without adequate vetting for harmful content.

c. **Ethical Challenges**:

   Bias and prejudice against a set of people: The tool perpetuated harm to children and minors through flawed algorithms, leading to the spread of harmful content.

   - **Accountability**:Meta and Instagram failed to adequately validate and monitor the algorithm before deployment, leading to significant harm.
   - **Human or user harm**: The tool's predictions and recommendations negatively affected the lives and safety of children and minors.

d. **Impact**:
   - **Mental and psychological harm**: Children and minors experienced trauma and distress due to exposure to harmful content.
   - **Prejudice**: Reinforced systemic harm and neglect within the platform.
   - **Fraud**: Misled users and organizations, undermining trust in the platform.

4.2 Goals, Methods, and Failures (GMF):

The Goals, Methods, and Failures (GMF) taxonomy is a failure cause analysis taxonomy for AI systems in the real world, interrelating the goals of the system deployment, the system's methods, and likely technical causal factors for the observed failure events. 4.2.1 AI Goal incident

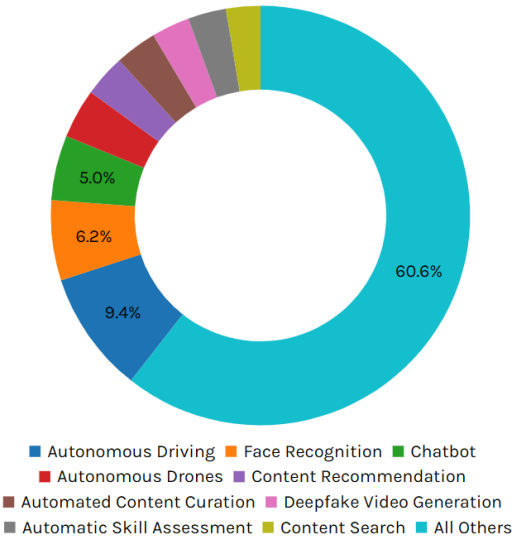
Known AI Goal

Searchable in Discover App

Discover:

- Autonomous Driving — 32 Incidents
- Face Recognition — 21 Incidents
- Chatbot — 17 Incidents
- Autonomous Drones — 13 Incidents
- Content Recommendation — 11 Incidents

Show more stats



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

Figure 4 GMF Known AI Goal Incidents chart

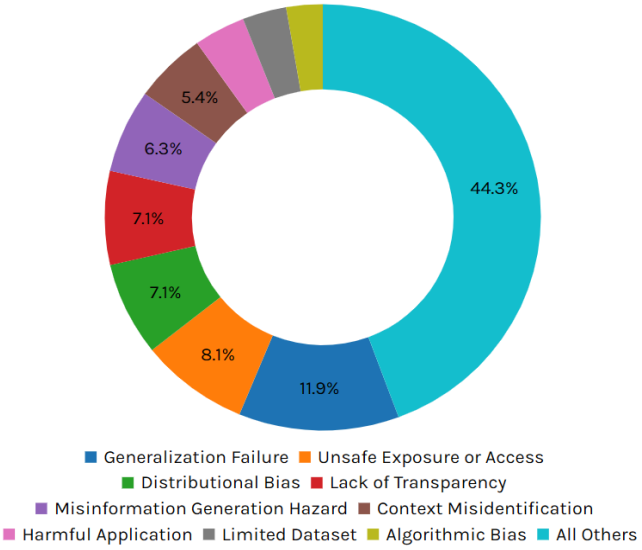
Known AI Technical Failure

Searchable in Discover App

Discover:

- Generalization Failure — 62 Incidents
- Unsafe Exposure or Access — 42 Incidents
- Distributional Bias — 37 Incidents
- Lack of Transparency — 37 Incidents
- Misinformation Generation Hazard — 33 Incidents

Show more stats



Definition: An AI Technical Failure which almost certainly contributes to the AI system failure referenced in the incident.

Figure 5 GMF Know AI Technical Failure chart

4.2.2 Known AI Technology

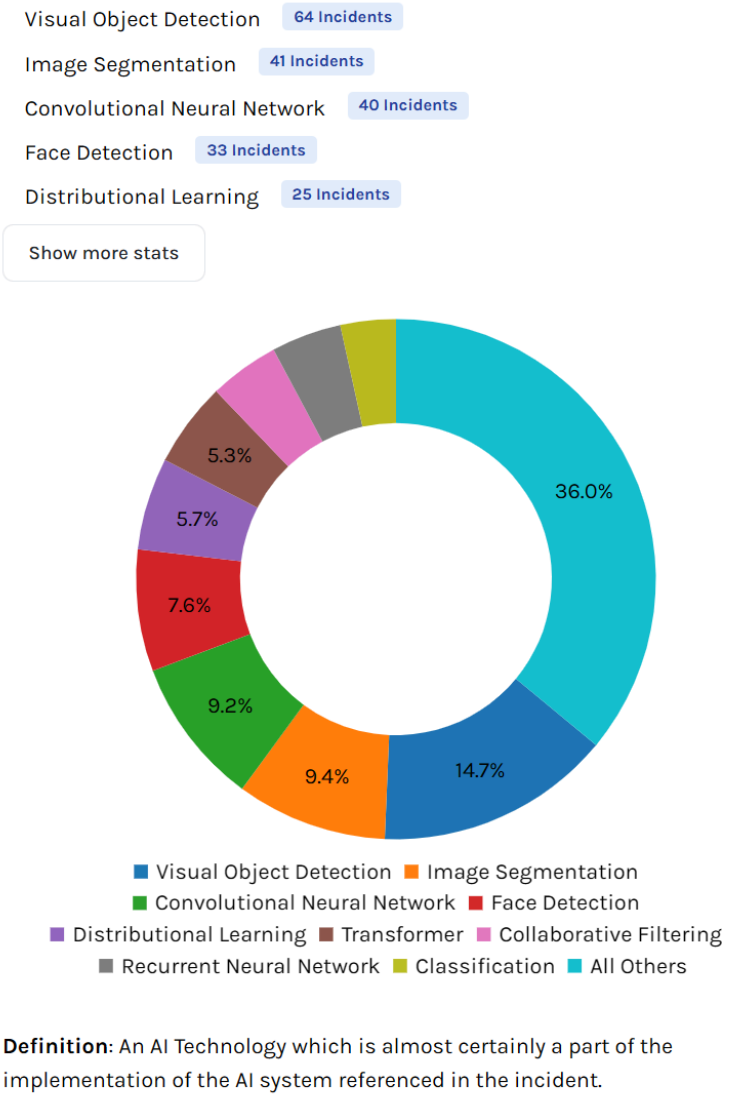


Figure 6 GMF Known AI Technology chart

4.2.3 Known AI Technical Failure

Taxonomies are contributed to the AI Incident Database by persons and organizations working to structure the data and provide views into the data. Each taxonomy must be of sufficient quality and completeness to be included in the AI Incident Database.

Observation:
It's clear that AI has the potential to revolutionize countless industries and improve our lives in countless ways. However, as we've seen, there's a dark side to this technology. Locally hosted AI models, while offering flexibility, can also introduce significant risks if not handled carefully.

- **Bias and Discrimination:** AI systems can perpetuate societal biases if not trained on diverse and representative data.
- **Misinformation and Disinformation:** AI-generated content can be used to spread false information, leading to harmful consequences.
- **Security Vulnerabilities:** Locally hosted models may be susceptible to cyberattacks, potentially compromising sensitive data.
- **Lack of Accountability:** It can be difficult to hold individuals or organizations accountable for the negative impacts of their AI systems.

Are the main concerns around AI Ethics.

5. Proposed Solutions

5.1 AI Governance:

- 5.1.1 EU Ethics Guidelines for Trustworthy AI: which states that for an AI to be "Trustworthy":
  - Lawful: respects all regulations
  - Ethical:Respects Ethical principals and values
  - Robust: both from a technical perspective
  - The guidlines are:
    - Human Agency and oversight
    - Technical Robustness and safety
    - Privacy and data governance
    - Diversity, non-discrimination and fairness
    - Societal and environmental well-being
    - Accountability
- 5.1.2 AI Governance Framework:
  - which presents 67 tasks all through the lifecycle phases, Phase 1 split into: 1a Planning and Design, 1b Data Activities. 1c Model Building and interpretation Phase 2 V&V. Phase 3 Deployment decision, Phase 4 Operation and Monitoring
- 5.1.3 IEEE Global Initiative on Ethics of Autonomous and Intelligent systems:
  - The framework is a consortium of different standards, including specific documents, e.g. regarding system design, certification, and bias. This framework generally consists of eight principles:
    - transparency
    - accountability
    - awareness of limitation
    - safety and well-being
    - reliability and dependability
    - equity
    - inclusivity
    - privacy protection

  It also includes a set of metrics to assess the extent to which AI systems adhere to these principles. These metrics are designed to provide a standardized way of

evaluating the ethical and responsible use of AI across different industries and applications.

- 5.1.4 Montreal Declaration of Responsible AI
    - 10 principles
    - ecological responsibility
    - democratic participation
    - respect for autonomy
    - as prudence during development
- 5.1.5 NIST AI Risk Management Framework:
    - Four core functions each split into categories and subcategories containing specific actions and outcomes:
        - governance
        - map
        - measure
        - manage AI risks
- 5.1.6 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence:

    5.1.6.1 AI Safety and Security

    - Mandates testing of AI models for safety, robustness, and reliability before deployment in critical areas like healthcare, defense, and infrastructure.
    - Calls for continuous monitoring to prevent malfunctions or misuse.
    - Directs agencies to ensure that AI systems meet security standards to guard against cybersecurity threats.
    - IEEE: "Avoid injuring others, their property, reputation, or employment by false or malicious action."
    - SWE Ethics: "Ensure that products meet the highest professional standards possible." Rigorous testing upholds these standards.

    5.1.6.2 Trustworthy AI

    - Requires transparency in AI operations and decision-making.
    - Sets measures to identify and mitigate algorithmic biases, ensuring fair and equitable outcomes.
    - Addresses the spread of AI-generated misinformation, including deepfakes, through standards and monitoring systems.
    - IEEE: "Avoid conflicts of interest and ensure integrity in professional practices." Transparency ensures accountability and fosters public trust.
    - SWE Ethics: "Act in a manner that is in the best interests of the client and public." Combating bias and misinformation aligns with the obligation to prioritize societal well-being.

    5.1.6.3 Privacy and Civil Rights

    - Protects individuals from intrusive surveillance, discrimination, or misuse of personal data by AI.

- Enhances privacy-preserving technologies and enforces data protection regulations.
- IEEE: "Respect the privacy of others." This aligns directly with the commitment to safeguard personal information.
- SWE Ethics: "Maintain the confidentiality of any information entrusted to them." Upholding privacy rights is a core responsibility of ethical engineering practice.

5.1.6.4 Supporting Innovation

- Promotes research and development of ethical, beneficial AI systems.
- Encourages public-private partnerships to advance innovation in line with safety and security standards.
- Allocates funding to AI initiatives that address public needs, such as education and healthcare.
- IEEE: "Improve the understanding by individuals and society of the capabilities and societal implications of conventional and emerging technologies." Fostering innovation while educating society fulfills this ethical mandate.
- SWE Ethics: "Advocate for the responsible use of computer systems and software." Developing beneficial AI aligns with the broader goal of ethical advocacy.

5.1.6.5 National Security and Defense

- Directs AI applications to strengthen national security and defense capabilities.
- Implements protocols to secure critical infrastructure, ensuring resilience against AI-enabled attacks or adversarial uses.
- IEEE: "Seek, accept, and offer honest criticism of technical work." Transparency and accountability in AI for defense foster trust and adherence to ethical standards.
- SWE Ethics: "Be fair and avoid harm to others." While ensuring security, measures must balance the ethical consideration of not overstepping civil liberties.

5.1.6.6 Global Leadership and Collaboration

- Positions the U.S. as a global leader in setting norms for responsible AI development.
- Promotes international cooperation to address shared challenges like cybersecurity and ethical AI use.
- IEEE: "Support colleagues and co-workers in their professional development and encourage them to adhere to this Code of Ethics." Collaborative efforts resonate with this ethical call for mutual growth and adherence to standards.
- SWE Ethics: "Participate in lifelong learning regarding the practice of their profession." International coordination enhances collective knowledge and responsible practice.

5.1.6.7 Workforce Development and Education

- Funds initiatives to train workers in AI-related skills.

- Expands STEM and AI education programs to prepare the workforce for emerging technologies.
- Ensures inclusivity in access to education and training opportunities.
- IEEE: "Assist colleagues and co-workers in their professional development." Workforce development aligns with fostering knowledge and capability in the tech community.
- SWE Ethics: "Ensure that colleagues are supported in their work and professional growth." Supporting the workforce aligns with the ethical duty to mentor and educate.

5.1.6.8 Environmental Sustainability

- Evaluates and reduces the environmental impacts of AI, including energy use and carbon footprint.
- Encourages the development of sustainable AI technologies.
- IEEE: "Improve the environment by ensuring responsible design and application of technology." Addressing sustainability aligns with the ethical imperative to reduce environmental harm.
- SWE Ethics: "Consider environmental impact and sustainability in software engineering projects." A focus on AI's ecological effects reflects this principle.

All above frameworks share similar core principles to define the large field of AI governance. NIST and AIGA try to offer more detailed tasks to be integrated into AI development. Yet, they all lack clear guidance on implementing a safe and feasible government process for organizations.

5.2 Enforced Rail-guards

Rail-Guards are a model that controls and measures the behaviour of another model, supervises it, and protects it from abuse in either direction. One example of railguards is the answer to the question "Are you AI?" then the railguard model answers with yes.
Enforcing these models might decrease the crime rate and abuse of Locally hosted Models by beginners, but for experts whom are more acknowledged with the filed, scraping off the Railguard Model can be done.
Also Enforcing the rail-guards on all the models hinders the scientific community research, which activity and use for AI has been regulated in 'THE BELMONT REPORT', Thus this solution isn't ideal and can't prevent the crimes that occured before from happening again.

5.3 Embedding the Unique key for each instance of a model

Having a Unique Key for each instance of a model, with a variance suffix for each decision or content it produces or makes. This makes defected Models trackable, and enforces accountability on the model, its developer and its hoster. This has various technical challenges including but not limited to:

6. Conclusion

While the challenges are real, we must not let fear stifle innovation. Instead, we should embrace AI's potential while taking proactive steps to mitigate its risks.

- **Ethical AI Development:** Prioritize fairness, transparency, and accountability in AI development companies, groups and solo developers, where a serious documentation phase is followed of each step.
- **Teach AI Governance Regulations:** Teaching the formerly mentioned regulations can raise awareness and promote the development approaches centered around Human and Community safety.
- **Public Education:** Educate the public about AI's capabilities and limitations to foster informed decision-making when it is time to vote for policy projects from politics.
- **Invest in AI Safety Research:** Support research into AI safety, including techniques for detecting and mitigating biases, ensuring robustness, and preventing malicious use.
- **Foster Collaboration:** All the community should put the effort in collaborating for the greater good when it comes to testing, deployment and tracking of AI Models, As following a selfish mindset of capitalism can bring unmitigatable risks upon us.

7. References

6.1 IEEE Std 7010-2020: IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being

6.2 IEEE Std 7009-2024: IEEE Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

6.3 IEEE Std 7007-2021: IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems

6.4 IEEE Std 7001-2021: IEEE Standard for Transparency of Autonomous Systems

6.5 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence on October 30, 2023

6.6 MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE 2018

6.7 NIST AI Risk Management Framework Second Draft, 2022

6.8 Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University: On the Opportunities and Risks of Foundation Models

6.9 Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: US Whitehosue Executive Order

6.10 THE BELMONT REPORT, Ethical Principles and Guidelines for the Protection of Human Subjects of Research

6.11 AI Incident Database CSETv1 Taxonomies