# Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents

Megan Li & Wendy Bickersteth

In collaboration with Ningjing Tang, Jason Hong, Hong Shen, Hoda Heidari, and Lorrie Cranor

CyLab Carnegie Mellon University Security and Privacy Institute

# Generative AI is everywhere and has caused harm

# Generative AI is everywhere and has caused harm

| Incident 623: Google Bard Allegedly Generated Fake Legal Citations in Michael Cohen Case |
| --- |
| **Description**: Michael Cohen, former lawyer for Donald Trump, claims to have used Google Bard, an AI chatbot, to generate legal case citations. These false citations were unknowingly included in a court motion by Cohen's attorney, David M. Schwartz. The AI's misuse highlights emerging risks in legal technology, as AI-generated content increasingly infiltrates professional domains. |

# Generative AI is everywhere and has caused harm

**Incident 623: Google Bard Allegedly Generated Fake Legal Citations in Michael Cohen Case**

**Description**: Michael Cohen, former lawyer for Donald Trump, claims to have used Google Bard, an AI chatbot, to generate legal case citations. These false citations were unknowingly included in a court motion by Cohen's attorney, David M. Schwartz. The AI's misuse highlights emerging risks in legal technology, as AI-generated content increasingly infiltrates professional domains.

**Deepfake President Zelenskyy instructs Ukraine army to surrender**

A deepfake video circulated online allegedly showing Ukraine president Volodymyr Zelenskyy instructing his army to lay down their arms and surrender.

# Generative AI is everywhere and has caused harm

**Incident 623: Google Bard Allegedly Generated Fake Legal Citations in Michael Cohen Case**

**Description**: Michael Cohen, former lawyer for Donald Trump, claims to have used Google Bard, an AI chatbot, to generate legal case citations. These false citations were unknowingly included in a court motion by Cohen's attorney, David M. Schwartz. The AI's misuse highlights emerging risks in legal technology, as AI-generated content increasingly infiltrates professional domains.

**Deepfake President Zelenskyy instructs Ukraine army to surrender**

A deepfake video circulated online allegedly showing Ukraine president Volodymyr Zelenskyy instructing his army to lay down their arms and surrender.

**Incident 765: 22 Students at Richmond-Burton Community High School in Illinois Targeted by Deepfake Nudes**

**Description**: 22 students at Richmond-Burton Community High School in Illinois were targeted in the creation of deepfake nudes. One of the students, Stevie Hyder, was targeted by classmates who used deepfake technology to alter her April 2023 prom picture into nude pictures, which were then circulated on social media. Two unnamed minors were arrested in late April 2024.

# A deeper understanding of the existing risks of Generative AI

"one whose instances have been observed in the real world and caused harm to individuals, communities, organizations, or the environment"

# A deeper understanding of the existing risks of Generative AI

"one whose instances have been observed in the real world and caused harm to individuals, communities, organizations, or the environment"

**Our contribution:** a coordinated taxonomy of

- Generative AI harms,
- *how* they arise, and
- *who* they affect,

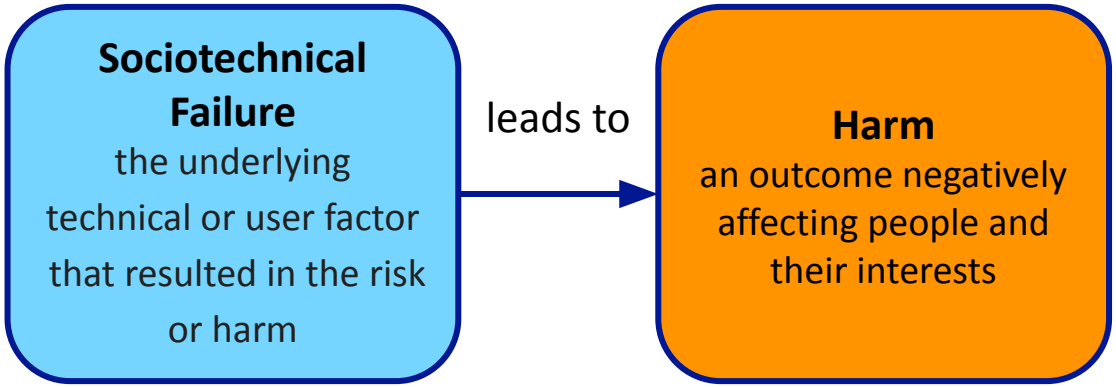
developed through a systematic analysis of 499 publicly reported incidents.

# Conceptual framework

**Sociotechnical Failure**
the underlying technical or user factor that resulted in the risk or harm

"How"

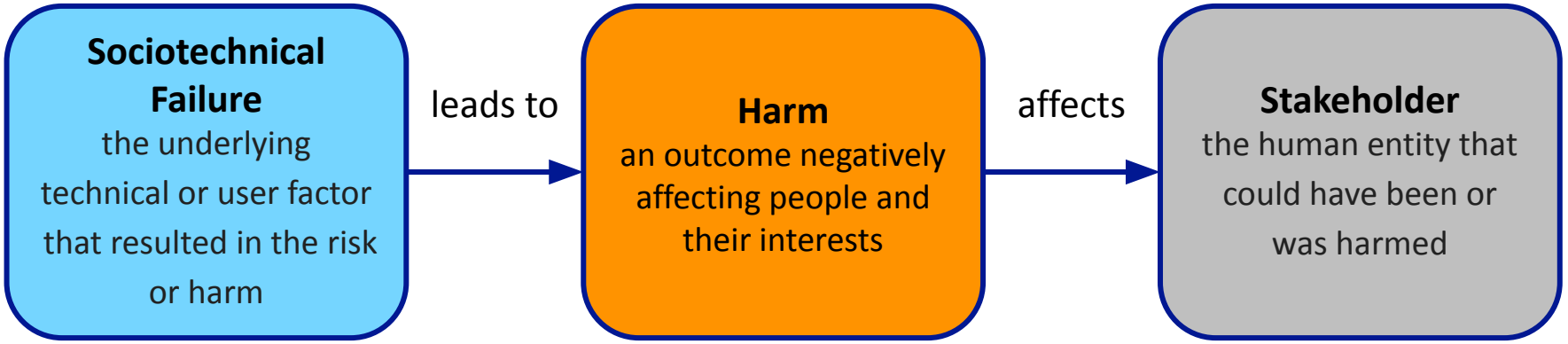
# Conceptual framework




"How"

"What"

# Conceptual framework



# Dataset of 499 Generative AI Incidents

**AI Incident Repositories**

AI, Algorithmic and Automation Incidents and Controversies Repository (AIAAIC)
https://www.aiaaic.org/aiaaic-repository

AI Incident Database (AIID)
https://incidentdatabase.ai/

# Dataset of 499 Generative AI Incidents

**AI Incident Repositories**

AI, Algorithmic and Automation Incidents and Controversies Repository (AIAAIC)
https://www.aiaaic.org/aiaaic-repository

AI Incident Database (AIID)
https://incidentdatabase.ai/

→ Identify and de-duplicate incidents involving *Generative* AI systems

# Dataset of 499 Generative AI Incidents

**AI Incident Repositories**

AI, Algorithmic and Automation Incidents and Controversies Repository (AIAAIC)
https://www.aiaaic.org/aiaaic-repository

AI Incident Database (AIID)
https://incidentdatabase.ai/

Identify and de-duplicate incidents involving *Generative* AI systems

Our dataset of 499 Generative AI incidents

# Applying our conceptual framework

**Incident 615: Colorado Lawyer Filed a Motion Citing Hallucinated ChatGPT Cases**

**Description**: A Colorado Springs attorney, Zachariah Crabill, mistakenly used hallucinated ChatGPT-generated legal cases in court documents. The AI software provided false case citations, leading to the denial of a motion and legal repercussions for Crabill, highlighting risks in using AI for legal research.

# Applying our conceptual framework

**Incident 615: Colorado Lawyer Filed a Motion Citing Hallucinated ChatGPT Cases**

**Description**: A Colorado Springs attorney, Zachariah Crabill, mistakenly used hallucinated ChatGPT-generated legal cases in court documents. The AI software provided false case citations, leading to the denial of a motion and legal repercussions for Crabill, highlighting risks in using AI for legal research.

Failure modes: *Hallucination* and *Improper use*

# Applying our conceptual framework

**Incident 615: Colorado Lawyer Filed a Motion Citing Hallucinated ChatGPT Cases**

**Description**: A Colorado Springs attorney, Zachariah Crabill, mistakenly used hallucinated ChatGPT-generated legal cases in court documents. The AI software provided false case citations, leading to the denial of a motion and legal repercussions for Crabill, highlighting risks in using AI for legal research.

Harms: *Reputational* and *Human Rights & Civil Liberties*

# Applying our conceptual framework

**Incident 615: Colorado Lawyer Filed a Motion Citing Hallucinated ChatGPT Cases**

**Description**: A Colorado Springs attorney, Zachariah Crabill, mistakenly used hallucinated ChatGPT-generated legal cases in court documents. The AI software provided false case citations, leading to the denial of a motion and legal repercussions for Crabill, highlighting risks in using AI for legal research.

Stakeholders: *End user* (Crabill) and *Individual beyond the end user* (Crabill's client)

# Findings

Our coordinated taxonomy has three parts:

| Taxonomy of sociotechnical failure modes | Taxonomy of Generative AI harms | Taxonomy of harmed stakeholders |
|---|---|---|

which we present alongside quantitative representations of how they relate to one another.

# Taxonomy of harmed stakeholders

**Interacting stakeholders**

- End user (individual)
- End user (organization)
- Developer/deployer

# Taxonomy of harmed stakeholders

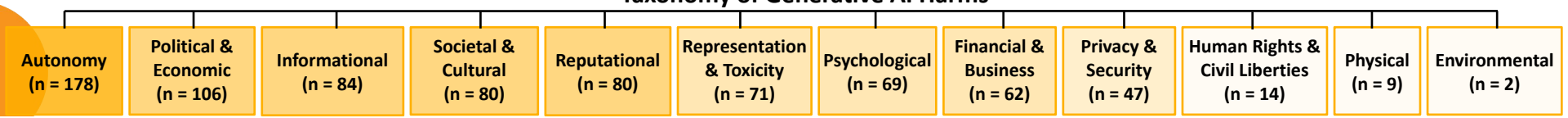
**Interacting stakeholders**

- End user (individual)
- End user (organization)
- Developer/deployer

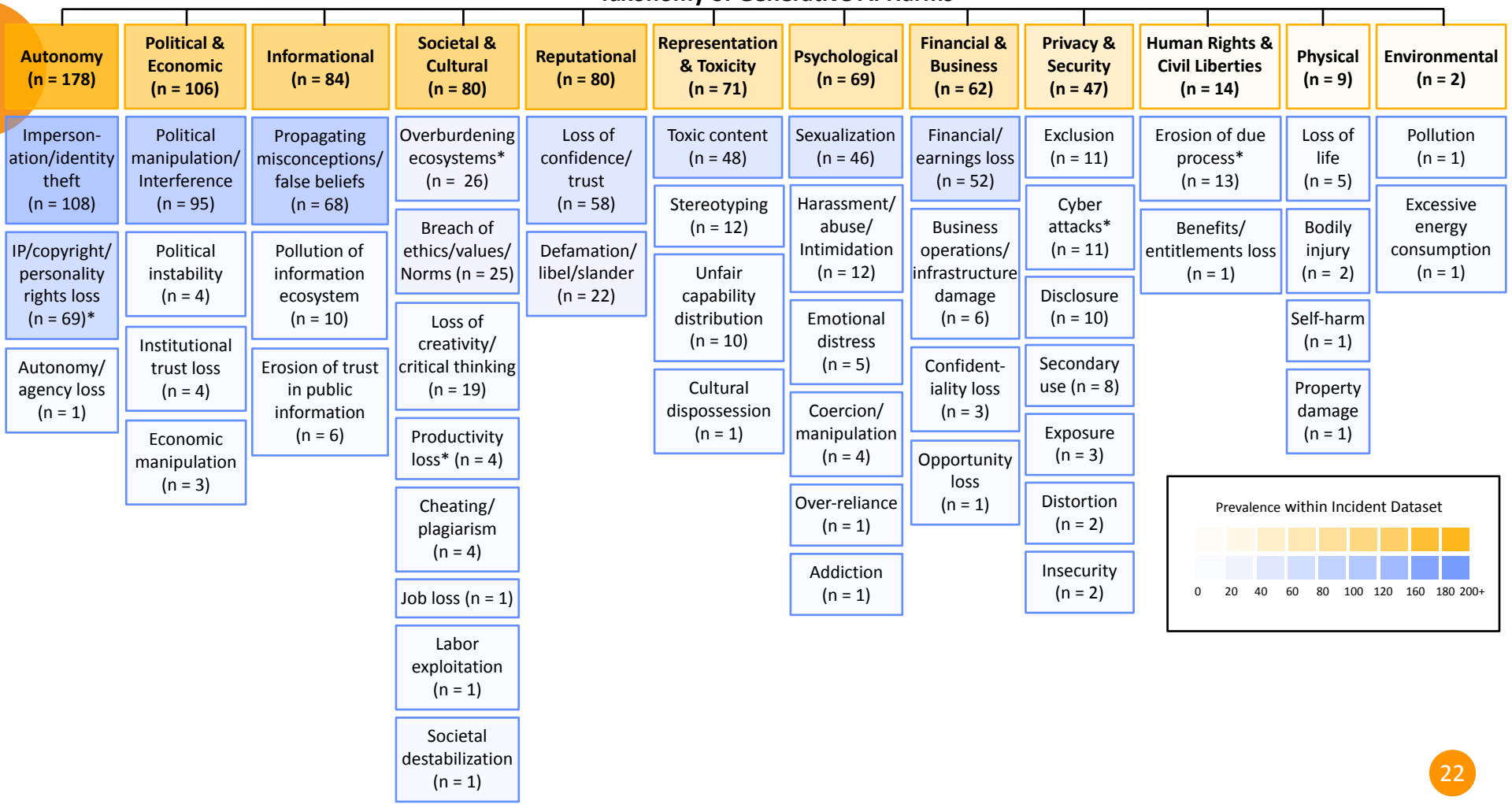
**Non-interacting stakeholders**

- Individual(s) beyond the end user
- Organization(s) beyond the end user
- Community
- Society

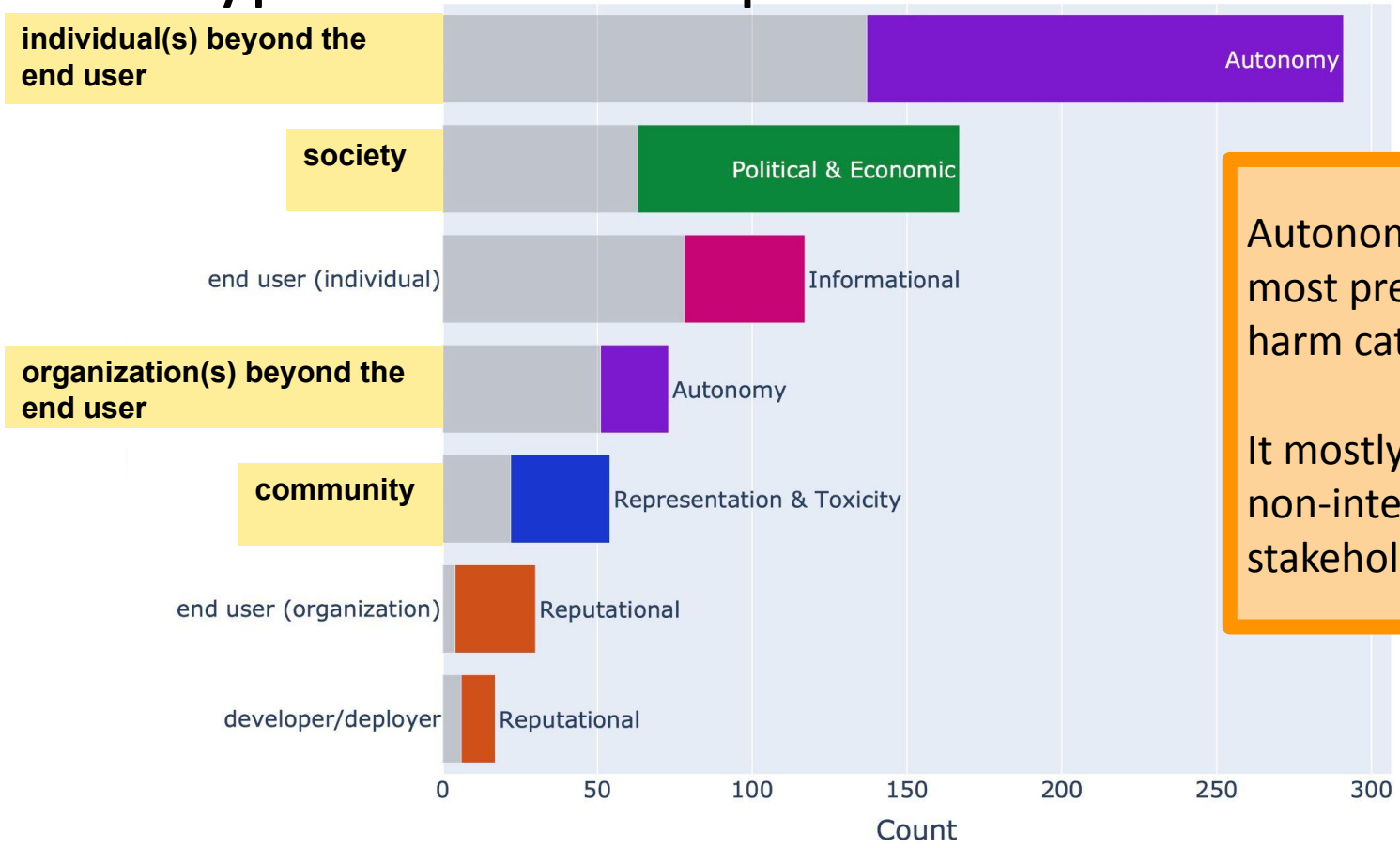
# Taxonomy of Generative AI Harms



# Taxonomy of Generative AI Harms



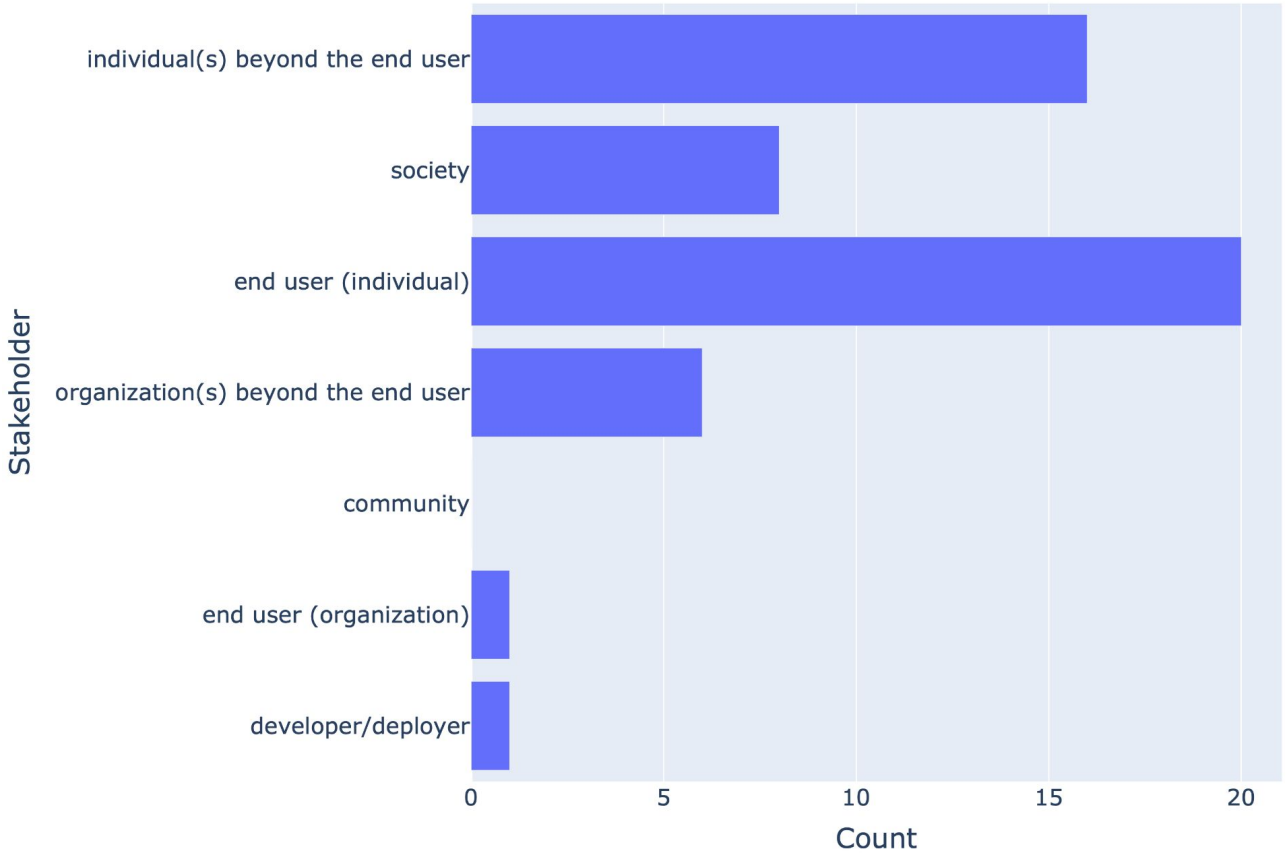
# What types of harm impact each stakeholder?



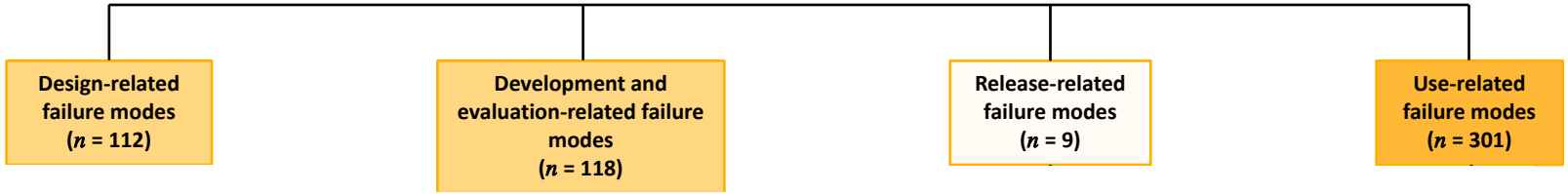

Autonomy is the most prevalent harm category

It mostly impacts non-interacting stakeholders

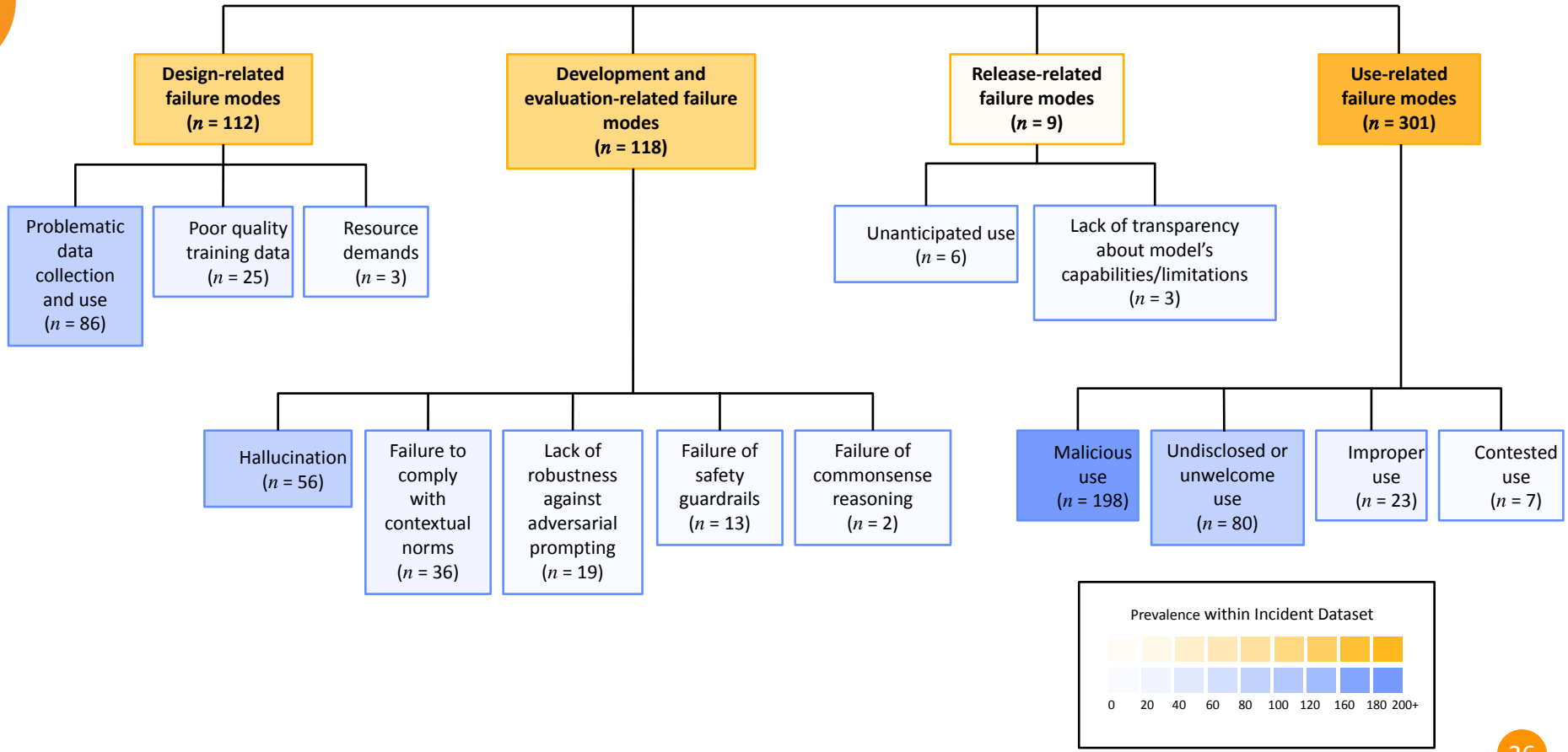
# Who is impacted by privacy harms?



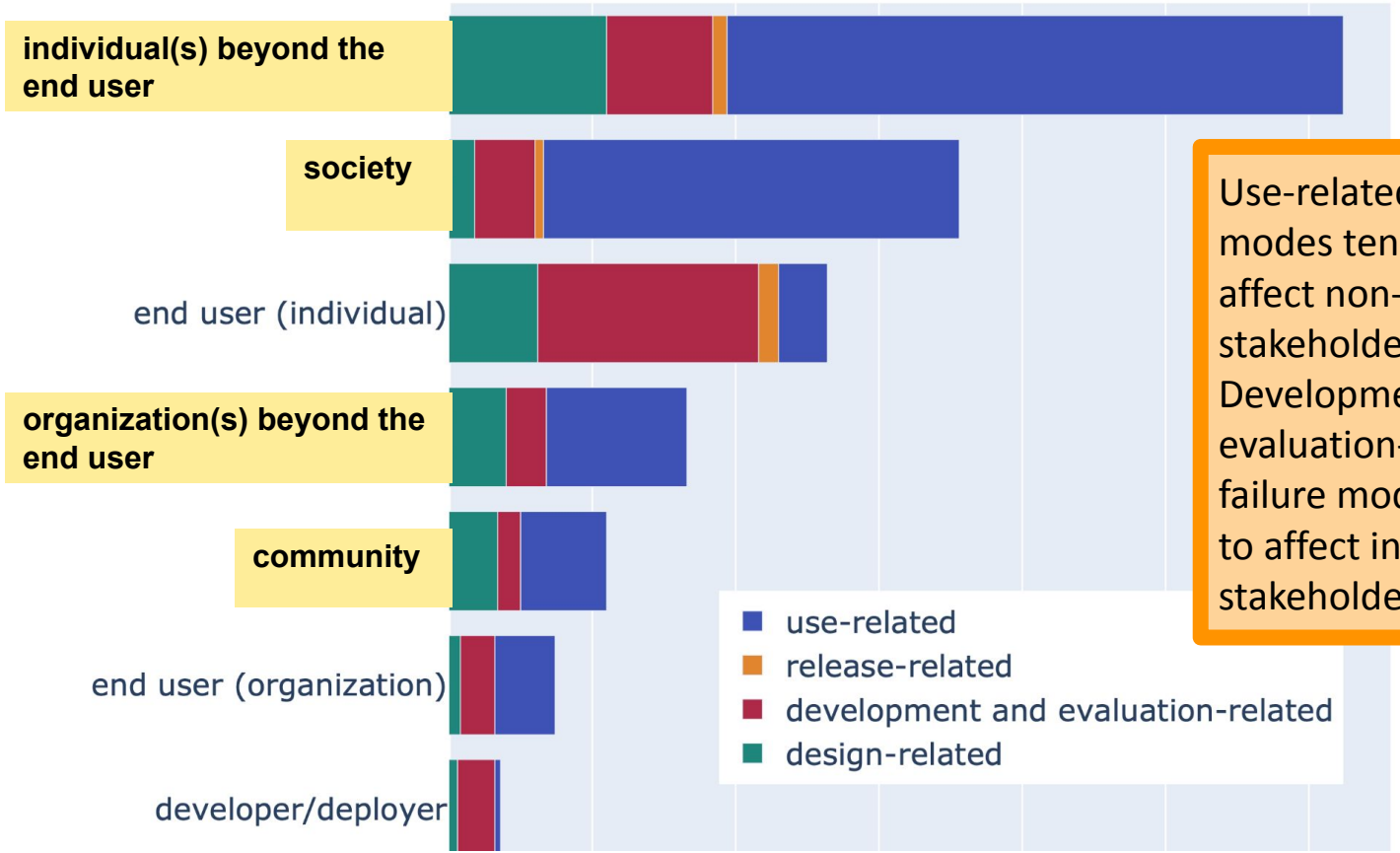
# Taxonomy of Sociotechnical Failure Modes



# Taxonomy of Sociotechnical Failure Modes

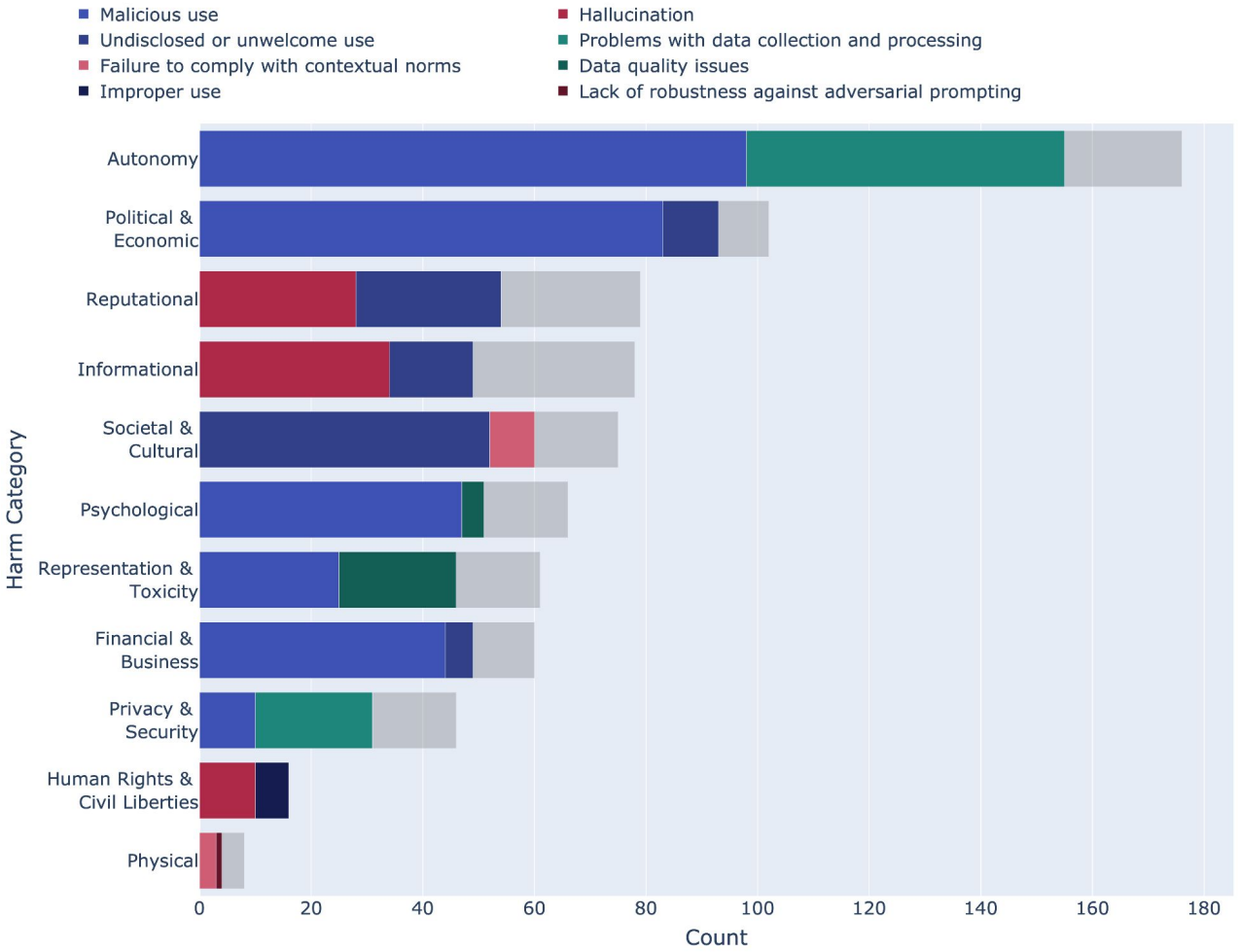


# What failure modes affect each stakeholder?




Use-related failure modes tended to affect non-interacting stakeholders while Development and evaluation-related failure modes tended to affect interacting stakeholders.

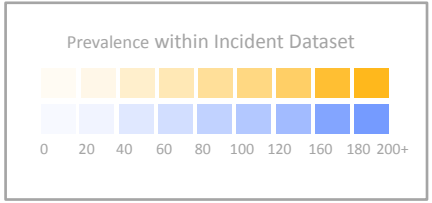
# What failure modes are associated with each type of harm?

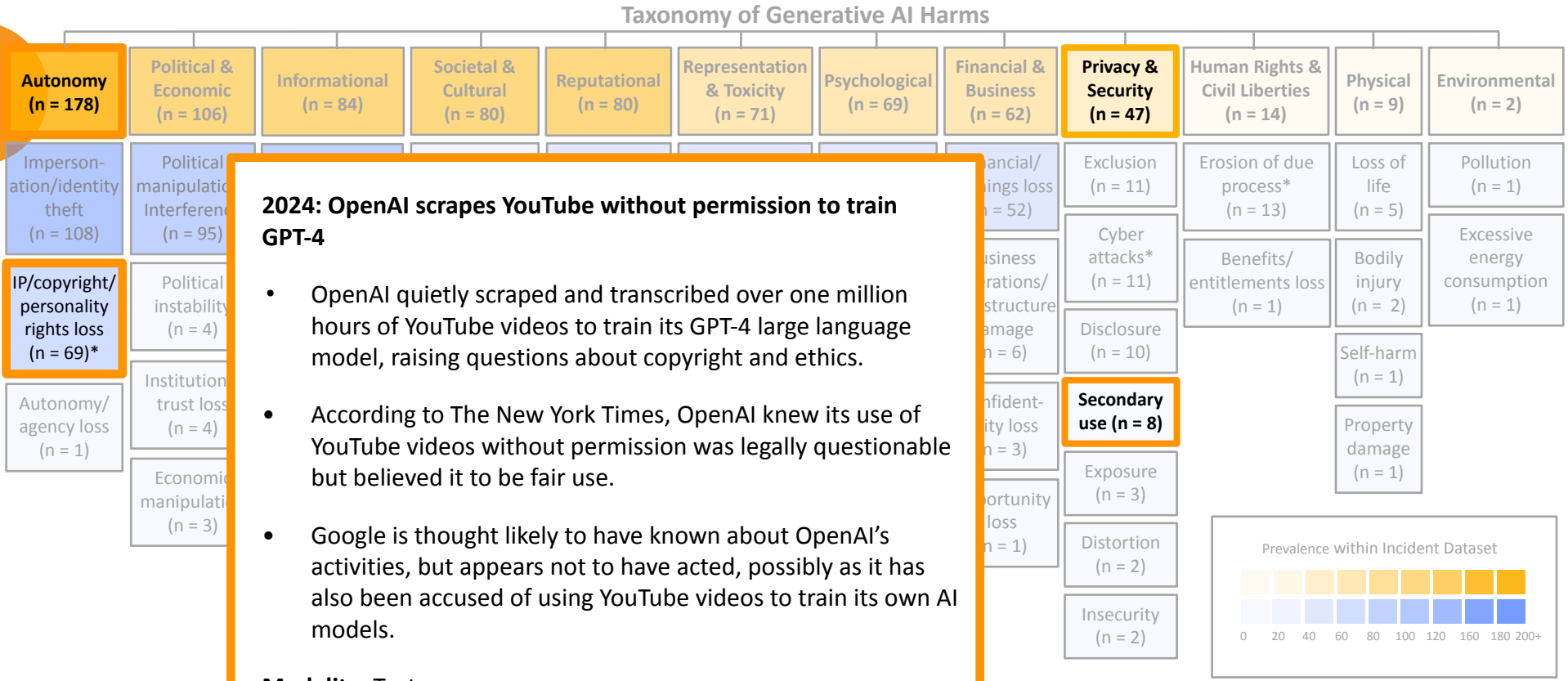


# Time for some examples!

# Taxonomy of Generative AI Harms

- **Autonomy (n = 178)**
  - Imperson-ation/identity theft (n = 108)
  - IP/copyright/ personality rights loss (n = 69)*
  - Autonomy/ agency loss (n = 1)
- **Political & Economic (n = 106)**
  - Political manipulation/ Interference (n = 95)
  - Political instability (n = 4)
  - Institutional trust loss (n = 4)
  - Economic manipulation (n = 3)
- **Informational (n = 84)**
  - Propagating misconceptions/ false beliefs (n = 68)
  - Pollution of information ecosystem (n = 10)
  - Erosion of trust in public information (n = 6)
- **Societal & Cultural (n = 80)**
  - Overburdening ecosystems* (n = 26)
  - Breach of ethics/values/ Norms (n = 25)
  - Loss of creativity/ critical thinking (n = 19)
  - Productivity loss* (n = 4)
  - Cheating/ plagiarism (n = 4)
  - Job loss (n = 1)
  - Labor exploitation (n = 1)
  - Societal destabilization (n = 1)
- **Reputational (n = 80)**
  - Loss of confidence/ trust (n = 58)
  - Defamation/ libel/slander (n = 22)
- **Representation & Toxicity (n = 71)**
  - Toxic content (n = 48)
  - Stereotyping (n = 12)
  - Unfair capability distribution (n = 10)
  - Cultural dispossession (n = 1)
- **Psychological (n = 69)**
  - Sexualization (n = 46)
  - Harassment/ abuse/ Intimidation (n = 12)
  - Emotional distress (n = 5)
  - Coercion/ manipulation (n = 4)
  - Over-reliance (n = 1)
  - Addiction (n = 1)
- **Financial & Business (n = 62)**
  - Financial/ earnings loss (n = 52)
  - Business operations/ infrastructure damage (n = 6)
  - Confident-iality loss (n = 3)
  - Opportunity loss (n = 1)
- **Privacy & Security (n = 47)**
  - Exclusion (n = 11)
  - Cyber attacks* (n = 11)
  - Disclosure (n = 10)
  - **Secondary use (n = 8)**
  - Exposure (n = 3)
  - Distortion (n = 2)
  - Insecurity (n = 2)
- **Human Rights & Civil Liberties (n = 14)**
  - Erosion of due process* (n = 13)
  - Benefits/ entitlements loss (n = 1)
- **Physical (n = 9)**
  - Loss of life (n = 5)
  - Bodily injury (n = 2)
  - Self-harm (n = 1)
  - Property damage (n = 1)
- **Environmental (n = 2)**
  - Pollution (n = 1)
  - Excessive energy consumption (n = 1)






**2024: OpenAI scrapes YouTube without permission to train GPT-4**

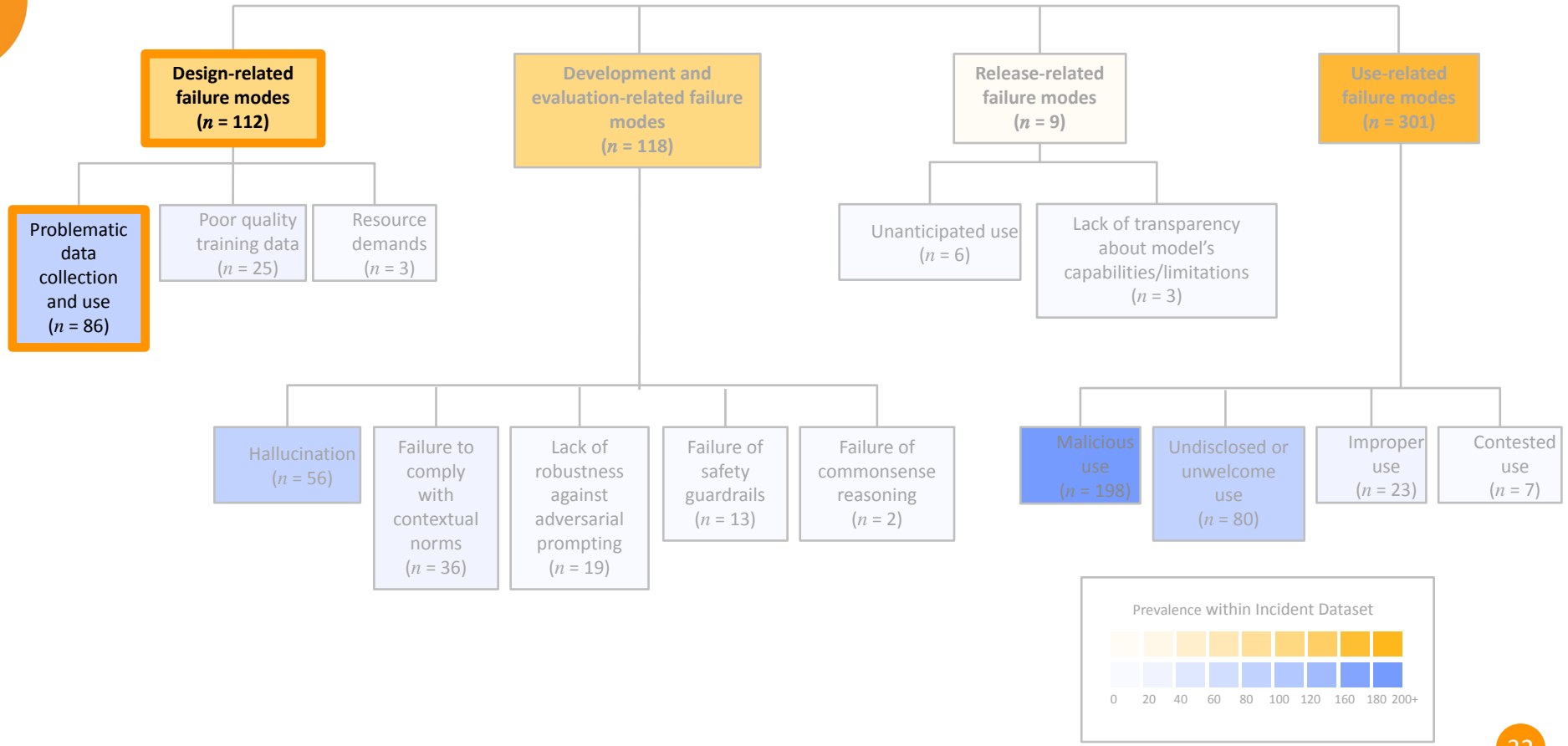
- OpenAI quietly scraped and transcribed over one million hours of YouTube videos to train its GPT-4 large language model, raising questions about copyright and ethics.
- According to The New York Times, OpenAI knew its use of YouTube videos without permission was legally questionable but believed it to be fair use.
- Google is thought likely to have known about OpenAI's activities, but appears not to have acted, possibly as it has also been accused of using YouTube videos to train its own AI models.

**Modality:** Text

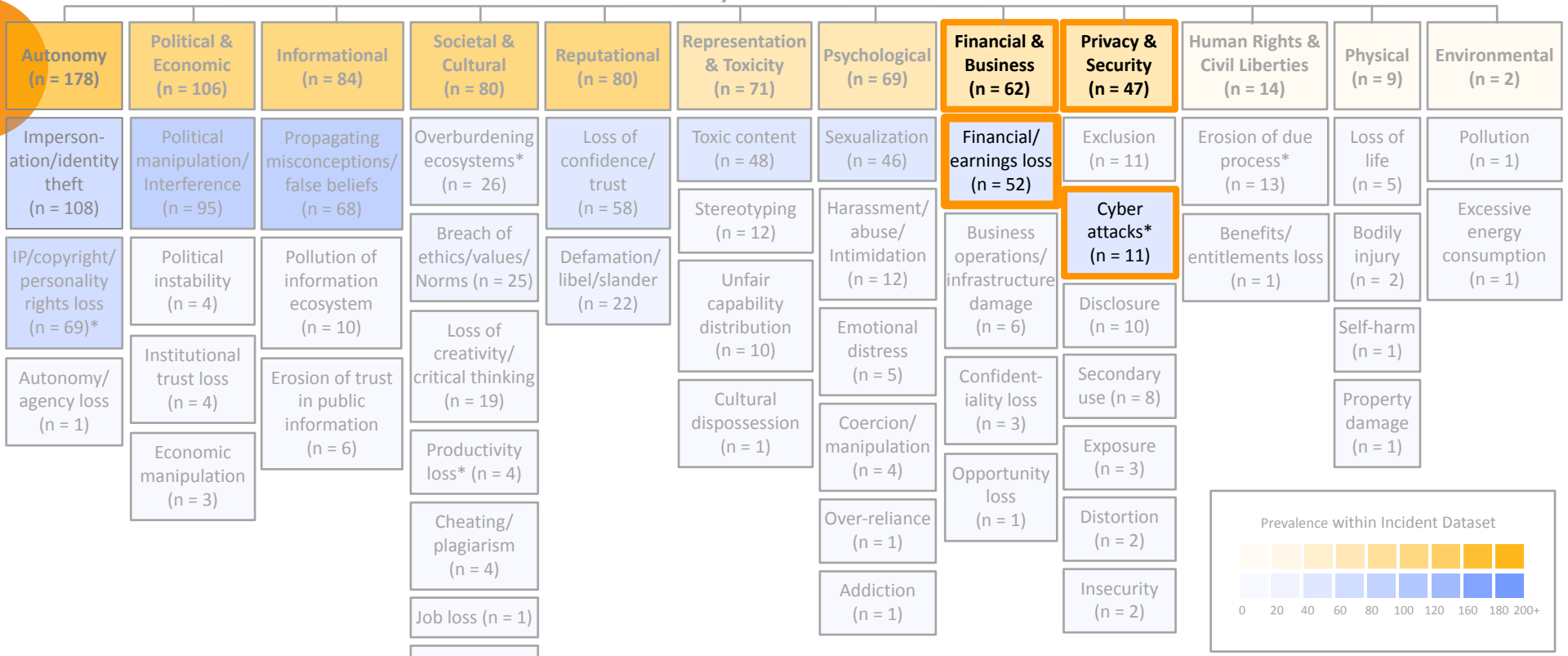
**Affected stakeholder:** Individual beyond the end user

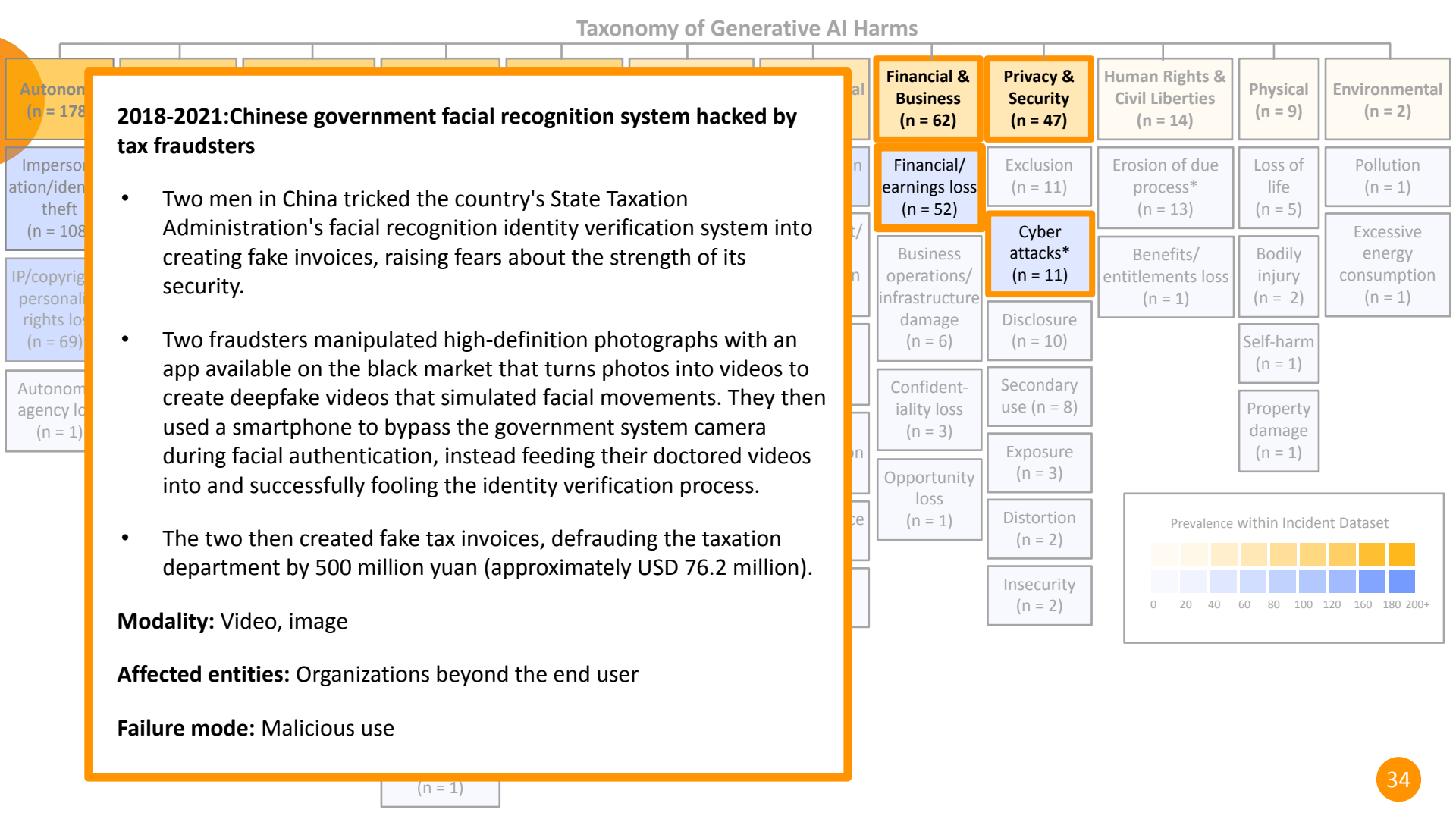
**Failure mode:** Problems with data collection and processing

# Taxonomy of Sociotechnical Failure Modes



# Taxonomy of Generative AI Harms






**2018-2021:Chinese government facial recognition system hacked by tax fraudsters**

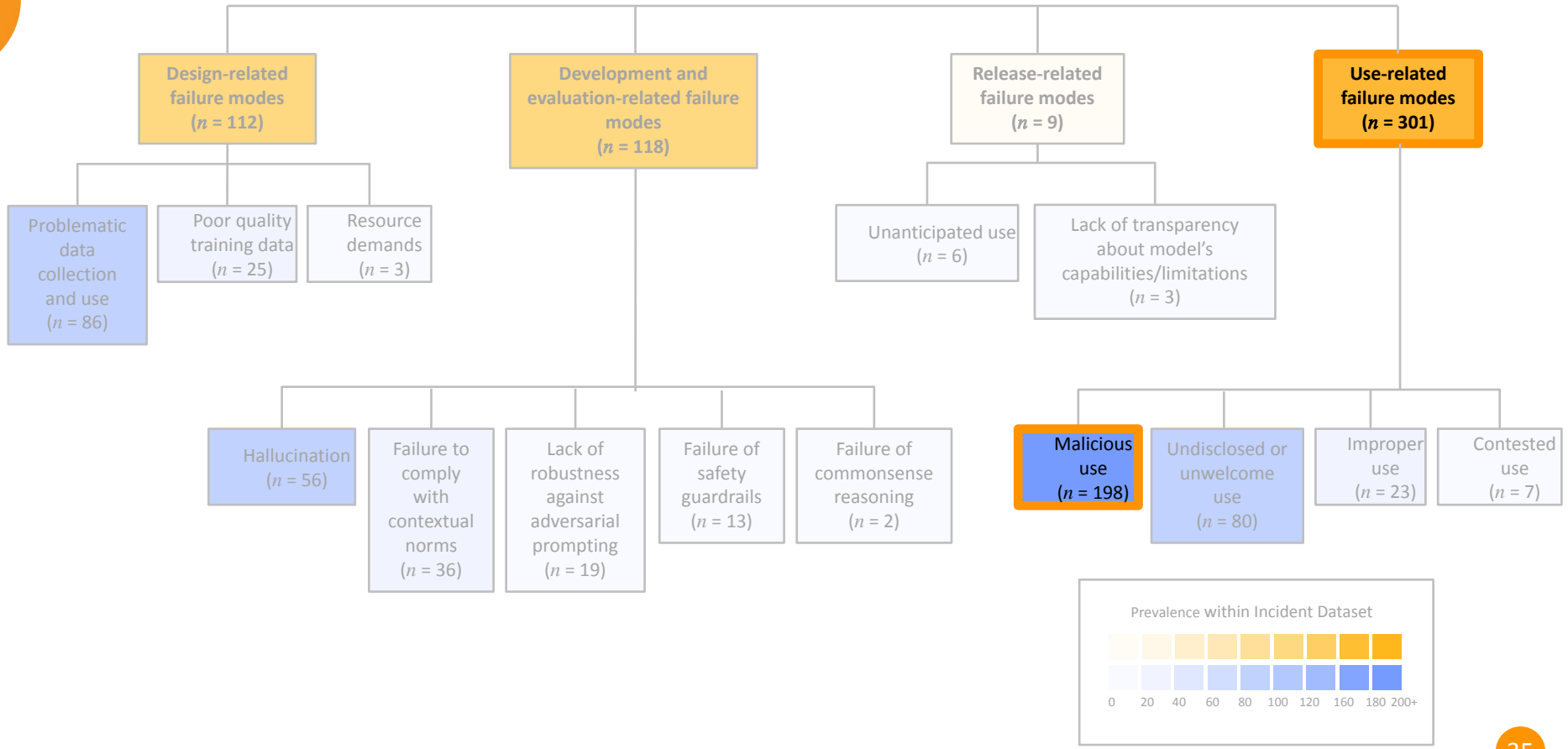
- Two men in China tricked the country's State Taxation Administration's facial recognition identity verification system into creating fake invoices, raising fears about the strength of its security.
- Two fraudsters manipulated high-definition photographs with an app available on the black market that turns photos into videos to create deepfake videos that simulated facial movements. They then used a smartphone to bypass the government system camera during facial authentication, instead feeding their doctored videos into and successfully fooling the identity verification process.
- The two then created fake tax invoices, defrauding the taxation department by 500 million yuan (approximately USD 76.2 million).

**Modality:** Video, image

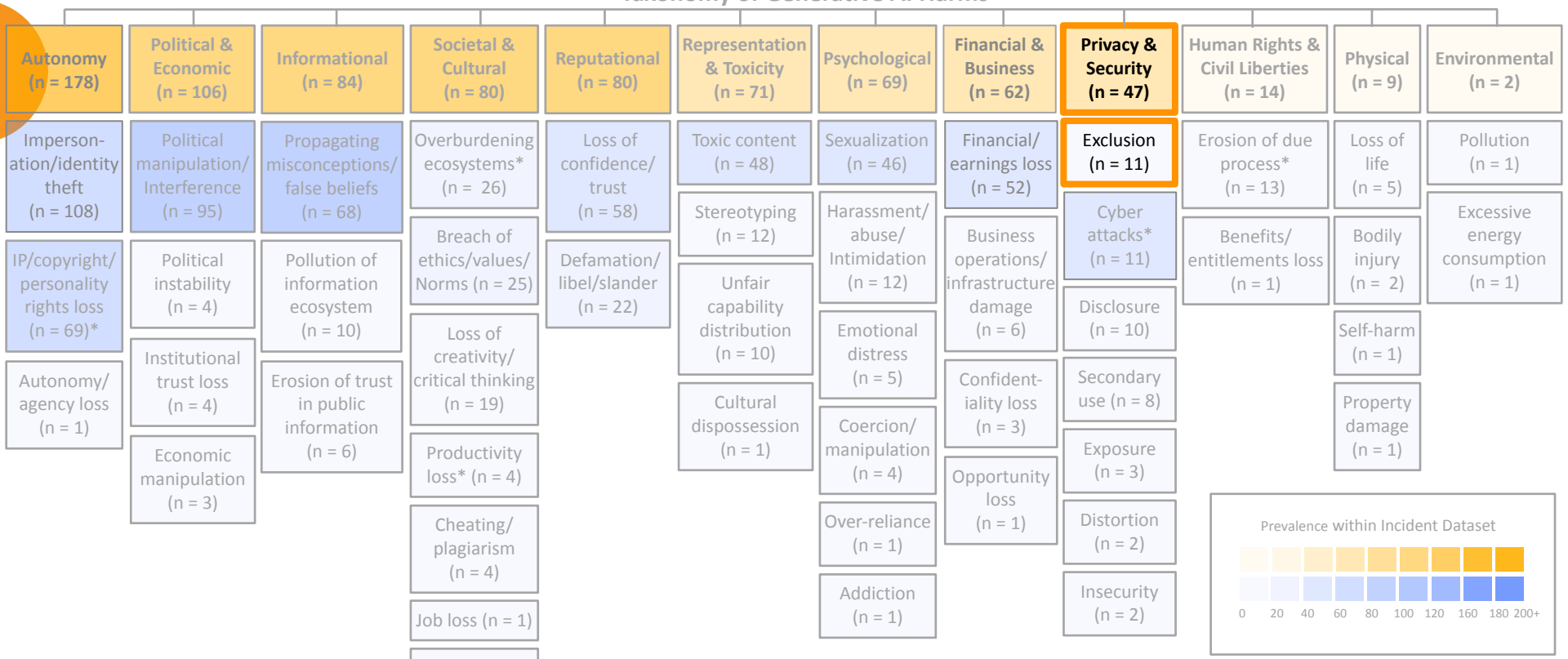
**Affected entities:** Organizations beyond the end user

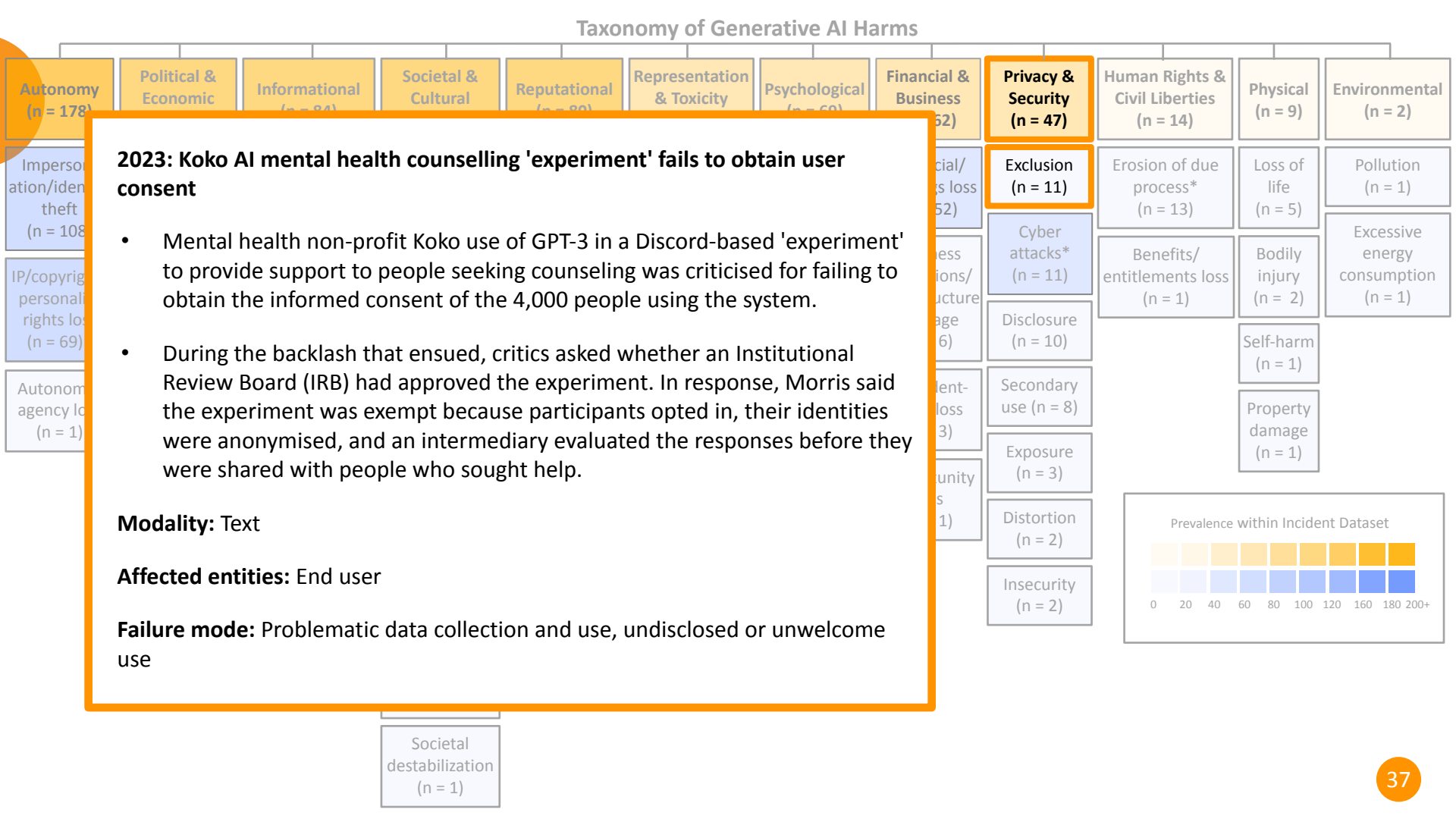
**Failure mode:** Malicious use

# Taxonomy of Sociotechnical Failure Modes



# Taxonomy of Generative AI Harms






**2023: Koko AI mental health counselling 'experiment' fails to obtain user consent**

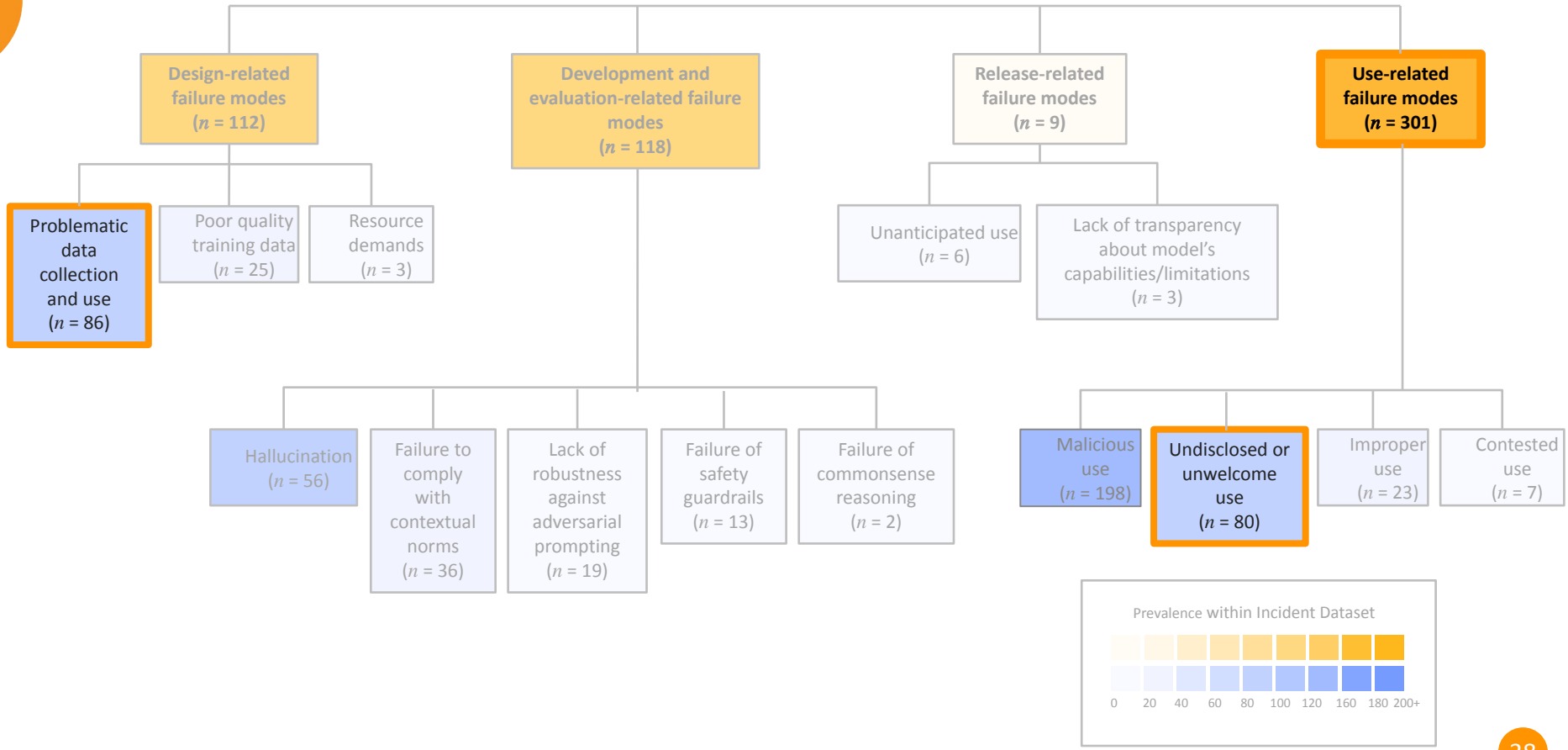
- Mental health non-profit Koko use of GPT-3 in a Discord-based 'experiment' to provide support to people seeking counseling was criticised for failing to obtain the informed consent of the 4,000 people using the system.
- During the backlash that ensued, critics asked whether an Institutional Review Board (IRB) had approved the experiment. In response, Morris said the experiment was exempt because participants opted in, their identities were anonymised, and an intermediary evaluated the responses before they were shared with people who sought help.

**Modality:** Text

**Affected entities:** End user

**Failure mode:** Problematic data collection and use, undisclosed or unwelcome use

# Taxonomy of Sociotechnical Failure Modes



# Limitations

- Some harms can have a cumulative nature

# Limitations

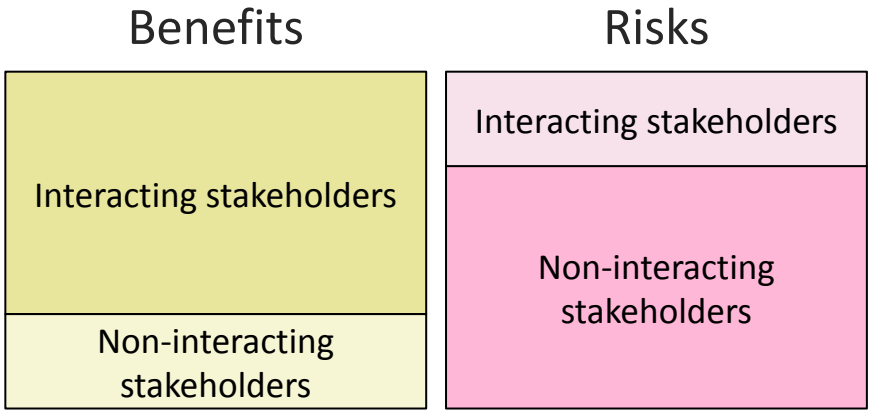
- Some harms can have a cumulative nature
- Harms may be over- or underreported due to reporting bias
  - Incidents that are sensational (e.g., deepfake impersonations) or affect well-resourced groups (e.g., celebrities) may be overreported

# Limitations

- Some harms can have a cumulative nature
- Harms may be over- or underreported due to reporting bias
  - Incidents that are sensational (e.g., deepfake impersonations) or affect well-resourced groups (e.g., celebrities) may be overreported
  - Incidents dealing with the design or development of Generative AI (e.g., data acquisition, implementation of safety guardrails) may be underreported
    - Privacy-related incidents are probably underreported

# Key Takeaway 1: The distribution of benefits and harms of Generative AI is uneven across stakeholders.

- This erodes one pathway of accountability
- What lessons can we take from data privacy?



# Key Takeaway 2: Generative AI changes the landscape of AI harms.

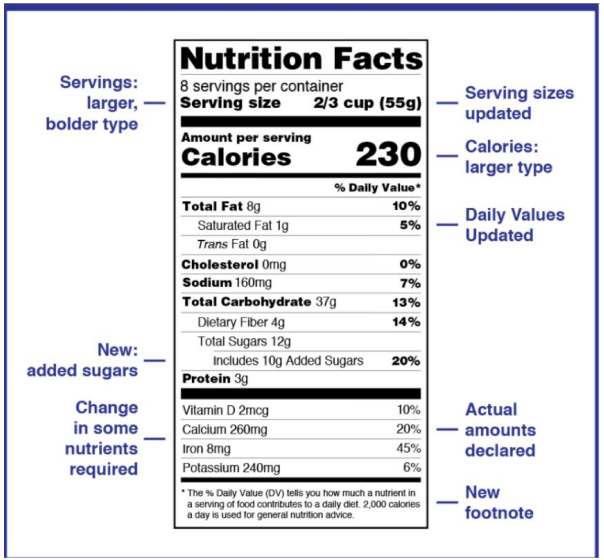
Some of our key findings are different from prior work analyzing AI harms more broadly.

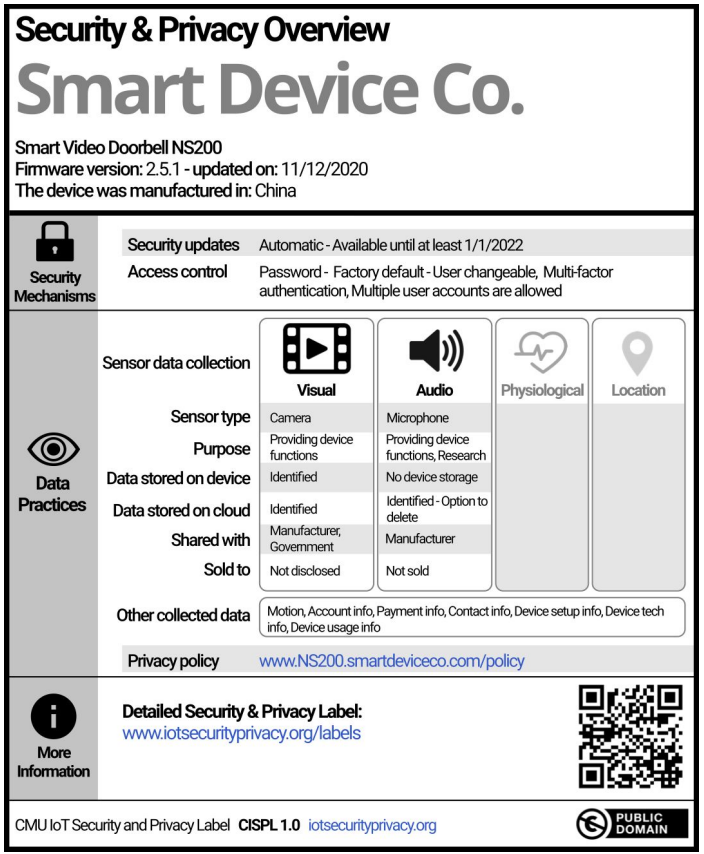
- Prior work has not found malicious use to be a common cause of harm while we found the opposite
- Prior work found the direct users of AI to experience most harm while we found the opposite

# Recommendation 1: Basic AI literacy can reduce certain risks of harm.

# Recommendation 1: Basic AI literacy can reduce certain risks of harm.

One idea: standardized **safety labels** for Generative AI applications



# Recommendation 2: Work is needed to address the limitations of this study.

- To address cumulative harms, longitudinal studies are critical
- To address reporting bias, confidential incident repositories such as that being developed by MITRE (with parallels in aviation and cybersecurity)
  - How do we encourage end users to engage in incident reporting?

# Thanks and check out our paper!

https://arxiv.org/abs/2505.22073

Megan Li • meggymuggy.github.io/

Wendy Bickersteth • wbickers@andrew.cmu.edu • cups.cs.cmu.edu

CyLab Carnegie Mellon University Security and Privacy Institute