# AI HARMS: TAXONOMY, DATASET AND REAL-WORLD MAP

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Artificial intelligence (AI) harms are increasingly reported worldwide, yet no global infrastructure systematically maps their distribution. In contrast, humanitarian crises are monitored through established networks that continuously collect, structure, and visualize evidence on global maps.[1] Evidence of AI harm, however, remains scattered across news reports. Existing AI-incident resources are also largely manually curated, focus on incidents instead of harms, and list cases without automatically locating or mapping them. As a result, chronic and infrastructural AI harms are largely missing, and the global distribution of AI harm remains invisible. To address that, we present the first automated **AI Harm Map**, a traceable pipeline that turns news reports into a public global map. It rests on a novel harm-centered taxonomy, anchored to 56 human-verified cases and 1,162 human-annotated harm reports. Building on this reference dataset, we scale coverage through automated retrieval, keyword-gated filtering, and LLM-based classification over a 2021–2026 news corpus, yielding 3,038 AI harm cases. To preserve the relational, temporal, and data provenance structure of these cases, We introduce PHTKG, a data provenance-aware temporal hypergraph that jointly models multi-role relationships, temporal dynamics, and evidence reliability, outperforming graph and hypergraph baselines on a chronological benchmark. On a chronological held-out benchmark of 174 events, our PHTKG outperforms the strongest of seven baselines by 4.68 AUC points. The enriched JSON output of PHTKG generates the public interactive AI Harm Map, geolocating 2,845 of 3,038 cases across 81 countries; 193 cases cannot be assigned to a single country, including regional and global cases. Our map reveals that subcategory prevalence varies sharply by geography: weapons-related harms are least US-centered (35% US vs 49%–65% other regions) and clustered in conflict-affected regions (Iran, Palestine, Russia and Ukraine), a pattern that holds when using only locations named in the reports. The dynamic AI Harm Map is available here[2]

## 1 INTRODUCTION

Humanitarian organizations monitor global crises through maps, yet no equivalent infrastructure systematically tracks, models, and visualizes real-world AI harms. As AI becomes embedded across infrastructure, governance, and labor markets, harms increasingly span interconnected socio-technical systems, while evidence remains fragmented across news sources. Existing resources primarily use model outputs or reported incidents rather than harm. DarkBench (Kran et al., 2025) examines harmful model behavior, while AIID (McGregor, 2021) and the CSET AI Harm Framework (Hoffmann & Frase, 2023) organize reported incidents. These approaches underrepresent chronic and infrastructural harms, while manual collection limits scalability, and reported cases introduce source and geographic biases (Olteanu et al., 2019; Bourgeois et al., 2018; Suresh & Guttag, 2021). Addressing these gaps requires a harm-centered detection and analysis framework, broader coverage of AI contexts, scalable human-verified collection, and a representation that captures how harms are interconnected, evolve, and recur across contexts.

[1] FEWS NET; WFP HungerMap; IPC; HDX Signals; ACAPS; UNHCR Data; IOM DTM. AI Alliance

[2] The dataset and code will be released after acceptance



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce the first automated **AI Harm Map**, a public and traceable automated framework for tracking, modeling, and analyzing real-world AI harms at *scale* (Figure 1). It combines a human-verified taxonomy and novel dataset with automated retrieval, filtering, and classification; represents cases as data provenance-aware temporal hyperedges to model harm recurrence and evolution through a data provenance-aware temporal hypergraph (PHTKG) model; and provides a public, interactive, dynamic world map (Figure 2).

Our proposed taxonomy (Figure 3) treats harm, rather than incidents, and spans 5 main categories (compute and model behavior, energy and land, labor, data, and deployment context and weapons) with 22 subcategories with acute/chronic tags. Compared with four existing frameworks, our taxonomy has 11 inherited, 5 adapted, and 6 original subcategories.

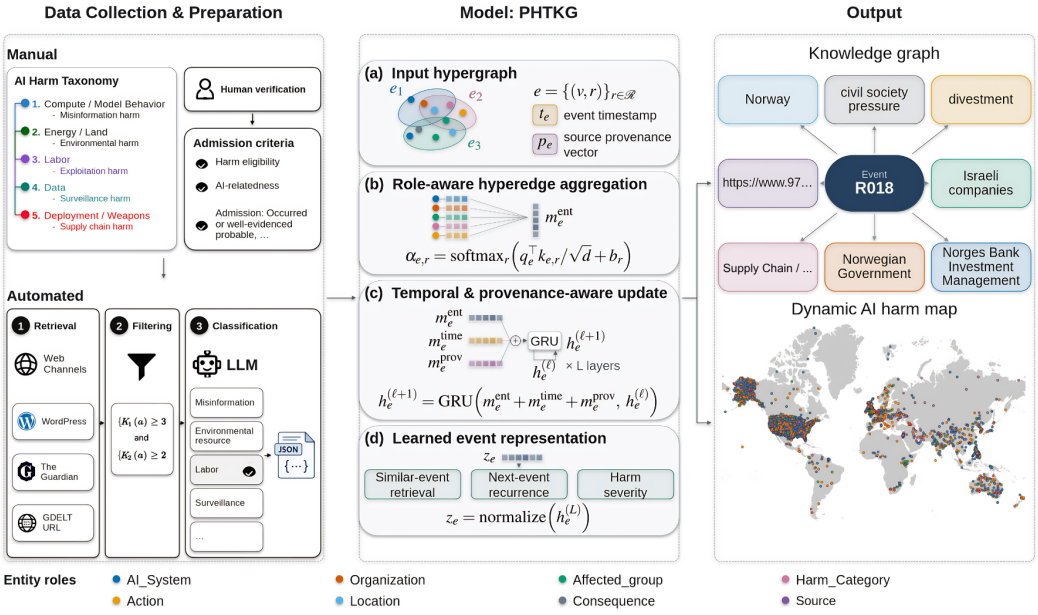



Figure 1: AI Harm Map pipeline: data collection and verification, PHTKG modeling, and the resulting knowledge graph and dynamic harm map.

Our contributions are (1) **a novel AI harm taxonomy** with 5 categories and 22 subcategories, treating harm rather than incident; (2) a dataset of 3,038 real-world AI harm cases from a 2021–2026 news corpus, anchored to 56 human-verified seeds; (3) **PHTKG**, a data provenance-aware temporal hypergraph model that represents AI harm events as role-aware seven-participant hyperedges and jointly models their relational structure, temporal evolution, evidence provenance, and recurrence; and (4) **a public, dynamic, automated AI Harm Map** geolocating 2,845 cases across 81 countries with source links. The dynamic map is available here.

## 2 Related Work

**AI harm resources** Several resources document real-world AI harm. AIID (McGregor, 2021), AIAAIC (AIAAIC), the OECD AI incidents and Hazards Monitor (OECD.AI, 2024), CSET's AI Harm Framework (Hoffmann & Frase, 2023), and the MIT AI Risk Repository (Slattery et al., 2024) collect cases through human submission and editorial review, following verification practices developed for conflict and human-rights monitoring. This process produces verified entries but bounds how many cases can be recorded and skews coverage toward incidents that receive English-language press. GDELT (Leetaru & Schrodt, 2013) extracts events at web scale, but it uses generic event categories and does not verify individual cases. Existing taxonomies (Weidinger et al., 2022; Shelby et al., 2023; Hoffmann & Frase, 2023; Slattery et al., 2024) organize AI risks from different angles, but they describe anticipated risks rather than verified harms, and they do not name labor, environmental (Crawford, 2021), or military harms, because their definition of AI is narrower than the one we adopt. Our taxonomy treats harm rather than incident and covers the 5 infrastructural categories of AI. The detailed AI harm resources are in Appendix A.

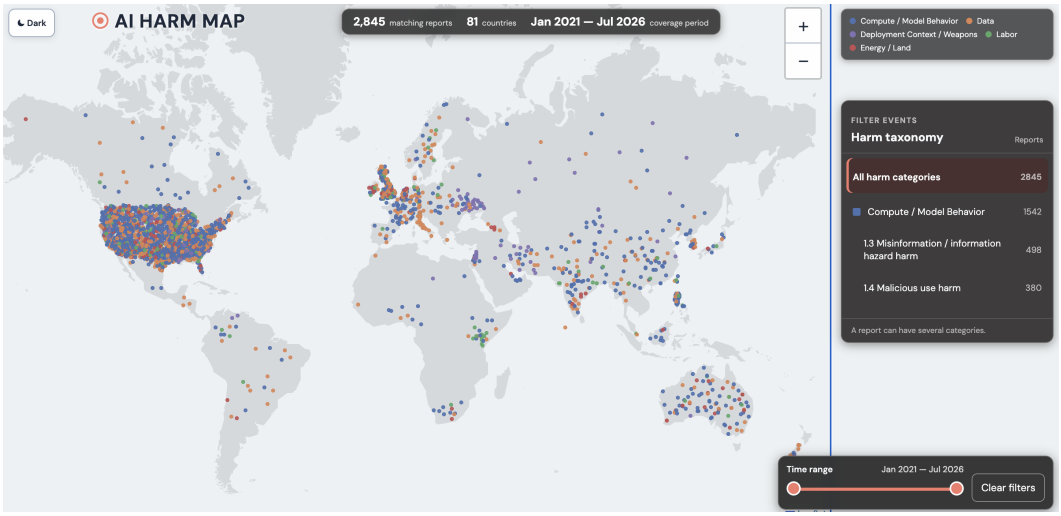



Figure 2: The interactive AI Harm Map: 2,845 cases across 81 countries, colored by category (dots randomized within countries). The legend filters by category or subcategory; selecting a dot shows its evidence, labels, date, and source link. Per-category static view: Appendix Figure 10. See the supplementary video for an interactive demonstration here.

**Knowledge representation and link prediction** Early work focused on learned embeddings on binary relations. DistMult (Yang et al., 2015) used a diagonal bilinear form, while ComplEx (Trouillon et al., 2016) extended it to complex space to better model asymmetric relations. RotatE (Sun et al., 2019) modeled relations as rotations in complex space, capturing composition patterns beyond bilinear models. These models are inherently transductive, assigning unique embeddings to entities and relations, which limits generalization to unseen ones. They also decompose each event into independent triples. Temporal knowledge graph models, including Know-Evolve (Trivedi et al., 2017), RE-Net (Jin et al., 2020), and DHyper (Tang et al., 2024), effectively capture evolving facts. However, they do not model extraction confidence or recurrence. Conversely, hypergraph and set-based models, including Deepsets-HG (Kim et al., 2021), SetTransformer-HG, Temporal-SetTransformer (Xue et al., 2026), and HYPER (Huang et al., 2026), preserve joint structure but differ in how they model time, confidence, and recurrence. Recent systems such as AutoSchemaKG (Bai et al., 2026) and SHARP (Ma et al., 2026) automate binary event extraction but do not model multi-participant-events or extraction confidence. Appendix A details the baselines, scoring functions, and limitations.

The **dynamic AI Harm Map** addresses these limitations through a novel harm-centered taxonomy, automated retrieval with human verification, and a provenance-aware temporal hypergraph, providing the first public map of geolocated AI harms.

## 3 Methodology

We design a pipeline that converts news reports into structured, temporal, and geolocated AI harm events (Figure 1). First, we build and validate a harm-centered taxonomy (Section 3.1). Second, automated retrieval, keyword filtering, and LLM-based classification produce 3,038 structured events from a 2021–2026 news corpus (Section 3.2). Third, we represent AI harm events as role-aware temporal hyperedges, train PHTKG on 1,162 human-annotated events with seven semantic roles, and store them in Neo4j (Section 3.3). Finally, we place the AI harm events on a public, dynamic AI Harm Map (Section 3.4); Of these, 2,845 are assigned to a country, while 193 cannot be mapped to a single country, including regional and global cases. Each stage provides the input to the next.

### 3.1 AI Harm Taxonomy and Dataset

First, we establish a taxonomy and a human-verified reference dataset that anchor subsequent automated collection and evaluation.

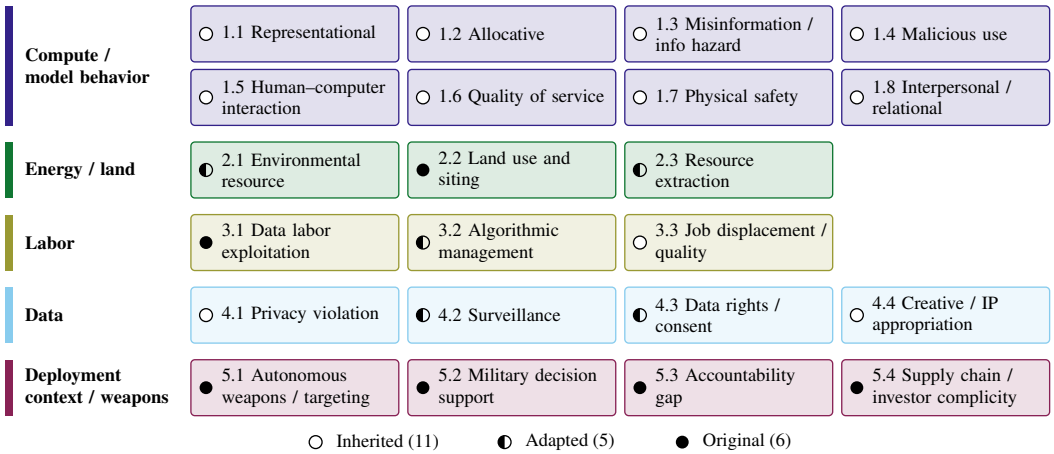



Figure 3: The AI Harm Map taxonomy: 5 categories and 22 subcategories. Colors indicate categories; markers show each subcategory's verdict: *Inherited*, *Adapted*, or *Original*.

**AI harm taxonomy construction** The AI Harm taxonomy is grounded in an infrastructural definition of AI covering compute, energy, land, labor, and data (Crawford, 2021). A facility, labor arrangements, and supply relationships are considered AI-related when dedicated to AI training, inference, or compute; shared infrastructure is included only when its AI-related contribution is identifiable, with unresolved cases neither forced into inclusion nor exclusion. This definition yields 5 categories comprising 22 subcategories. These categories are Compute/Model Behavior, Energy/Land, Labor, Data, and Deployment/Weapons. The 22 subcategories were compared with CSET (Hoffmann & Frase, 2023), the MIT AI Risk Repository (Slattery et al., 2024), Weidinger et al. (2022), Shelby et al. (2023), and relevant humanitarian law and arms-control literature, and classified as Inherited (11), Adapted (5), or Original (6). Appendix B.1 has the full taxonomy and comparison.

Following Feinberg (1987), harm is defined as a documented adverse effect on a protected interest, directly or foreseeably caused by an AI system. Each case receives an occurrence status: occurred, probable, implied, or speculative; only occurred and well-evidenced probable harms are included. Taking harm as the unit allows both acute and chronic harms to be represented. Geolocation is especially important for data-center and weapons-related cases because their impacts depend strongly on local context. The weapons category integrates AI-harm research with international humanitarian law and arms-control literature, incorporating four weapons-related subcategories to represent documented harms and accountability gaps.

### 3.2 Automated AI Harms Collection and Classification

The manual protocol establishes a rigorous 56-case ground truth, while real-world AI harms accumulate faster than manual review can track. Therefore, automated collection extends coverage at scale while preserving the same inclusion criteria and taxonomy. It retrieves, filters, and classifies news from a corpus spanning January 2021 to July 2026.

**AI harm sources, retrieval, and filtering** Automated retrieval uses 3 channels (Appendix C.1) across 27 English-language domains spanning general interest, technology and security, and regional outlets in Africa, South and Southeast Asia, Latin America, and Eastern Europe, reflecting the uneven distribution of AI-harm reporting across these categories (Brennen et al., 2018). The 3 channels are public WordPress REST APIs, the Guardian Content API, and direct fetching of GDELT-discovered article URLs. All channels produce records under a common schema (Appendix C.2). The detailed retrieval procedures and source counts are given in Appendix C.1 (Table A3).

We filter one retrieved article $a$ by the co-occurrence of two keyword lists (Appendix C.3) within one article: the AI keyword list $K_1$ (231 entries) and the Harm keyword list $K_2$ (588 entries), both built from the taxonomy. An article is retained only if $|K_1(a)| \geq 3 \quad \text{and} \quad |K_2(a)| \geq 2$, counting distinct matched entries rather than occurrences; the joint and asymmetric thresholds enforce genuine co-

occurrence of AI- and harm-related terms while reducing incidental matches. The filtering rule is in Appendix C.4, with pre- and post-filter article counts in Table A6 (Appendix C.5).

**Manual collection and verification of AI Harm dataset** Using the taxonomy and inclusion criteria, a human reviewer manually collected and fully verified 56 cases to establish the reference dataset. For each case, the reviewer (i) verifies the source as primary or well-corroborated; (ii) checks eligibility against the AI-relatedness and harm criteria; (iii) assigns the occurrence status and applies the inclusion rule above; and (iv) assigns the applicable taxonomy labels, acute/chronic status, and geolocation where applicable. Two examples are provided in Appendix B.2.

**AI harm classification** Keyword co-occurrence does not guarantee that AI and harm terms are related, so we introduce a classification stage that enforces precision. We prompt an LLM with a fixed protocol that encodes the taxonomy and inclusion criteria (Appendix C.6): it first strips boilerplate and data-quality artifacts, then reviews every remaining article. For each selected article we assign all applicable subcategories and one primary label by severity, and we mark articles that fit no subcategory as UNMAPPED. Of the 11,628 articles retrieved through WordPress and GDELT, we select 2,337 (20%); adding 352 Guardian articles, we obtain a classified set of 2,689 articles carrying 3,052 classifier labels. The resulting dataset contains 3,038 cases, each a structured AI harm record; a case can carry more than one subcategory label, and 56 cases fall under more than one category.

### 3.3 PHTKG (PROVENANCE-AWARE HYPERGRAPH TEMPORAL KNOWLEDGE GRAPH)

We introduce PHTKG, a data provenance-aware temporal hypergraph model for representing and dynamically updating AI harm events over time. Each AI harm event is represented as a single role-aware temporal hyperedge connecting seven semantic roles: organization, AI system, affected group, action, harm category, consequence, and location. Event and entity states evolve chronologically as new AI harm events enter the graph, with role-aware aggregation, learned temporal decay, and provenance weighting controlling how past evidence contributes to later states. PHTKG is jointly trained for event compatibility and next-year entity recurrence. To separate representation quality from extraction noise, we evaluate the same architecture on a gold-event graph derived from human annotations and a predicted-event graph derived from automatically extracted events. This design preserves multi-role event structure while modeling how AI harm events recur, evolve, and connect across time. The learned representations also support similarity-based retrieval of structurally related AI harm events. Appendix D.1 provides an example.

**Structured event extraction and verification** We use 1,162 human-annotated AI harm reports with structured event fields and process them with a frozen `Qwen/Qwen2.5-3B-Instruct` model (Qwen et al., 2025), given the original-language report, its English translation, source metadata, and taxonomy information. For each report, the model extracts one central AI harm event as JSON: event_type, the seven event roles, the event date, and verbatim original and translated evidence spans; missing information is recorded as `"Not specified in report"`. Each extracted event then passes three verification stages: schema verification (valid JSON, a single event, required fields), natural language inference (NLI) verification against the source report and evidence spans, and harm-category verification against the taxonomy definition and inclusion criterion. A single pass yields valid predictions for 1,159 of the 1,162 reports, which we evaluate field-wise against the human annotations using exact match and token-level F1. Each predicted event receives a provenance score $p \in [0, 1]$ from the extractor's self-reported confidence; since it is produced during extraction, it is available at inference time without human annotations, and PHTKG uses it to weight predicted events during message passing. Appendix D.1 details extraction, verification, and evaluation.

**PHTKG construction** After extraction and verification, we represent each AI harm event $e$ as a role-aware hyperedge $e = (V, \tau, p)$, where $V = \{(v_1, r_1), \ldots, (v_7, r_7)\}$ lists the seven participants with their roles, $\tau$ is the event year, and $p$ is the provenance score (Figure 1, panel (a)). We keep the source URL as metadata so that every event remains traceable, but it is not a graph role. We build two graphs with the same seven-role structure. The *gold-event graph* contains the 1,162 human-annotated AI harm events with $p = 1$, while the *predicted-event graph* contains the corresponding automatically extracted AI harm events with their provenance scores. Comparing these paired graphs allows us to assess how extraction errors affect graph modeling while keeping the graph architecture unchanged. PHTKG processes AI harm events in chronological yearly steps. First, we

add a learned role embedding to each entity embedding, so the same entity can have different representations in different roles, and role-aware attention combines the seven participants into a single event message (Figure 1, panel (b)). Second, a recurrent update combines this message with temporal and provenance information to form the event state (Figure 1, panel (c)). Third, we propagate the updated event state back to its participants through role-specific transformations. Each message is scaled by a learned temporal decay and by $\mathrm{sigmoid}(p_e)$, giving recent and higher-confidence AI harm events greater influence while allowing each role to retain information at its own rate. Finally, we carry entity states from one year to the next and update only entities that appear in the current year, so each new AI harm event contributes to existing histories rather than rebuilding them. Full equations are provided in Appendix D.2.

**Training PHTKG** We train PHTKG on two tasks jointly (panel (d)). The first task teaches the model to distinguish observed AI harm events from corrupted ones. For each observed event, we create corrupted versions by swapping in incorrect participants and ensure that these corrupted events do not reuse information from the validation or test years (Appendix D.3). A scoring function $s(e) = \mathrm{MLP}(\mathrm{normalize}(h_e))$ applied to the final event state $h_e$ assigns higher scores to observed events. We train this scorer using binary cross-entropy and a margin ranking loss. The second task is an auxiliary next-year recurrence task that predicts whether an entity seen by the end of year $t$ appears again in year $t+1$. The input is $z_{v,t} = [h_v^t; T(t+1)]$, where $h_v^t$ is the state of entity $v$ at year $t$ and $T(t+1)$ encodes the next year. The model predicts $P(v \text{ appears at } t{+}1) = \mathrm{sigmoid}\big(\mathrm{next_year_head}(z_{v,t})\big)$ using a dedicated MLP trained with binary cross-entropy on entities that do and do not reappear. This task gives the model a direct learning signal for temporal recurrence in AI harm events. Loss weights, warmup, negative sampling, and optimization details are in Appendix D.3. After training, each AI harm event has a learned representation. We retrieve similar events using cosine similarity, $\mathrm{sim}(e_i, e_j) = \frac{h_i^\top h_j}{\|h_i\| \, \|h_j\|}$, where $h_i$ and $h_j$ are the learned representations of events $e_i$ and $e_j$. We inspect retrieved events by event ID, similarity, harm category, AI system, and year, allowing structurally related AI harm events to be studied together.

**Neo4j graph database** We also store the AI harm events in a Neo4j graph database where each AI harm event is represented as an Event node linked to its seven role entities, its source, and its most similar events. We standardize entity names to reduce duplicates while retaining the original extracted text for traceability. The database supports structured queries across AI harm events, and we generate the dynamic AI Harm Map (Section 3.4) directly from the enriched JSON output. Details of the Neo4j schema and graph construction are provided in Appendix D.4.

### 3.4 AI Harm Map Visualization

News reports describe harm locations in prose rather than coordinates, so each case is assigned at most one country using a fixed priority ladder: (1) country or U.S. state named in the location field, (2) major city named there, (3) headquarters country of the responsible organization, or (4) operating country of a named platform. The first applicable rule and earliest relevant mention determine the placement. Coordinates are then sampled uniformly within the assigned country's Natural Earth 1:50m polygon using a fixed seed; thus, map points indicate the country concerned, not the precise incident location. The same seeded layout backs both the interactive world map (Figure 2) and its static per-category view (Appendix Figure 10).

The ladder places 94% of cases: 2,276 by explicitly named country, 27 by city, 510 by organizational headquarters, and 32 by platform. The remaining 193 cases (6.4%) are left unplaced when no single territorial locus can be established, avoiding arbitrary assignment for global, transnational, or inherently non-local harms. The potential geographic bias introduced by the headquarters rule is examined in Section 4. Each point is filterable by category or subcategory and links to the case evidence, labels, source outlet, and original article. The complete placement rules and additional analyses are provided in Appendix G. Unplaced cases are retained in the released dataset with their labels, location text, and source links, and are listed separately in the interactive map rather than drawn as points; Appendix G.4 analyzes them.

# 4 Experimental Results and Discussion

## 4.1 Taxonomy and dataset results

The 3,038 cases are distributed across Inherited (2,139; 70%), Adapted (630; 21%), and Original (269; 9%) subcategories. Inherited subcategories dominate, led by Misinformation/information hazard ($n = 497$) and Malicious use ($n = 425$), while the six original subcategories account for 269 cases not represented in the reviewed frameworks (Appendix E.1).

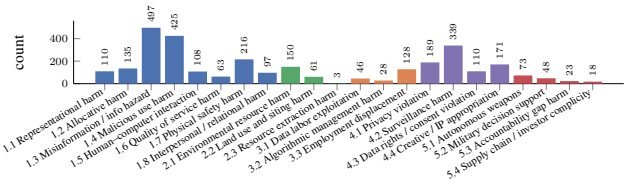



Figure 4: Subcategory-level case counts for the AI Harms.

At the category level, Compute / Model Behavior is largest ($n = 1{,}651$), followed by Data ($n = 809$), with Energy / Land ($n = 214$), Labor ($n = 202$), and Deployment Context / Weapons ($n = 162$) substantially smaller. The least populated subcategories are Resource extraction ($n = 3$), Supply chain / investor complicity ($n = 18$), Accountability gap ($n = 23$), and Algorithmic management ($n = 28$); the complete breakdown is shown in Figure 4. These frequencies reflect documented reporting under the inclusion criteria rather than incidence rates (Appendices E.1 and G.1).

By temporal type, 1,900 cases are acute (63%), 1,054 chronic (35%), and 84 both (3%). Since event type is assigned by subcategory, the distribution reflects the composition of the corpus. Chronic harms may be underrepresented because cumulative impacts such as resource consumption or labor-market changes lack a discrete triggering event suitable for article-level retrieval; therefore, the observed proportion should not be interpreted as the true incidence of chronic harm. The per-category distribution is provided in Appendix E.2.

## 4.2 PHTKG results

Table 1: PHTKG benchmark and ablation results. Left: chronological predicted-event benchmark ($n = 174$ held-out AI harm events); $\Delta$ is the AUC difference from HYPER in percentage points. Right: PHTKG component ablations at seed $= 42$. Higher is better.

**(a) Benchmark comparison**

| Method | AUC | Δ | AUPRC | F1 | MCC |
|---|---|---|---|---|---|
| ComplEx (Trouillon et al., 2016) | $49.34 \pm 0.69$ | $-34.01$ | $24.34 \pm 0.54$ | 40.24 | 3.92 |
| SetTransformer-HG | $52.36 \pm 1.13$ | $-34.99$ | $27.51 \pm 0.92$ | 37.07 | 1.84 |
| DistMult (Yang et al., 2015) | $53.75 \pm 1.38$ | $-33.60$ | $27.55 \pm 0.84$ | 37.67 | 2.94 |
| RotatE (Sun et al., 2019) | $53.81 \pm 0.83$ | $-33.54$ | $26.05 \pm 0.53$ | 38.30 | 8.17 |
| Temporal-SetTransformer (Xue et al., 2026) | $54.19 \pm 0.35$ | $-33.16$ | $28.81 \pm 0.94$ | 36.16 | 2.71 |
| DeepSets-HG (Kim et al., 2021) | $62.16 \pm 1.31$ | $-25.19$ | $31.81 \pm 1.33$ | 44.16 | 18.85 |
| HYPER (Huang et al., 2026) | $87.35 \pm 0.40$ | 0.00 | $41.60 \pm 9.00$ | 44.46 | 47.19 |
| **PHTKG (Ours)** | $\mathbf{92.03 \pm 3.30}$ | **+4.68** | $\mathbf{64.32 \pm 4.08}$ | **79.38** | **72.74** |

**(b) Component ablations**

| Variant | AUC | AUPRC | F1 | MCC |
|---|---|---|---|---|
| w/o state carryover | 44.41 | 22.41 | 40.22 | 3.71 |
| w/o explicit role parameters | 51.83 | 25.66 | 41.69 | 11.09 |
| w/o time encoding/decay | 57.10 | 30.84 | 38.41 | 0.78 |
| w/o recurrence objective | 57.54 | 28.74 | 39.53 | 4.05 |
| w/o provenance | 62.00 | 35.84 | 45.41 | 0.90 |
| Triple decomposition | 58.77 | 23.45 | 40.75 | 0.6 |

We evaluate PHTKG on a chronological held-out test set of 174 predicted AI harm events. For each observed test event, we construct corrupted counterparts by replacing event participants while avoiding information from the validation and test years. We use AUC to measure how well PHTKG distinguishes observed AI harm events from corrupted counterparts and compare it with seven baseline models. Detailed training and evaluation settings are provided in Appendix D.3. Table 1 summarizes the benchmark results. PHTKG achieves $\mathbf{92.03 \pm 3.30}$ AUC, compared with $87.35 \pm 0.40$ for the strongest baseline, HYPER (Huang et al., 2026). This is **+4.68 AUC points over HYPER** and 29.87– 42.69 points over the other six baselines. The remaining baselines score between 49.34 and 62.16. These results indicate that jointly modeling the seven roles within a temporal hyperedge provides a stronger representation for distinguishing observed AI harm events from corrupted counterparts. Component ablations further show that all five mechanisms contribute to performance: removing state carryover produces the largest observed degradation, while removing explicit role parameters, temporal encoding and decay, recurrence supervision, or provenance weighting also reduces performance. The triple-decomposition comparison likewise performs substantially below the complete model. Together, these results support the combined PHTKG design rather than a single component in isolation. Detailed representation-level analyses are provided in Appendix F.

### 4.3 AI HARM MAP RESULTS

The map (Figure 2) answers 4 questions that a list of incidents cannot: where AI harm is reported, which harms occur where, when new kinds of harm appear, and which actors and systems link cases across borders.

**Where is AI harm reported?** The 2,845 placed cases span 81 countries but are highly concentrated: the United States holds 57%, the United Kingdom 11%, and the Philippines 6%, while half of all countries hold four cases or fewer. This concentration is not an artifact of our placement rules: using only locations named in the reports shifts each category's US share by just 2-6 percentage points and leaves the ranking unchanged (Appendix G.2). It reflects where AI harm is *reported* in English-language news, not necessarily where it occurs. The remaining 193 cases (6.4%) have no single location, as they are mostly online or global harms such as malicious use and misinformation (Appendix G.4).

**Each category has its own geography** Energy / Land is the most U.S.-centered category (65%), following contested data center siting, whereas Deployment Context / Weapons is the least U.S.-centered (35%) and the narrowest (17 countries), clustering in active conflicts: Palestine (27), Ukraine (26), Iran (11), and Russia (10). Compute / Model Behavior and Data are the most widespread (57 and 58 countries) because they follow deployed products rather than territory. Country profiles differ accordingly: weapons harm accounts for 43–65% of category tags in Ukraine, Palestine, Iran, and Russia against a 6% baseline; data harm, led by creative/IP appropriation and data rights, dominates in Germany (52%) and the Netherlands (41%); and 89% of Philippine cases are compute/model-behavior harms, mostly misinformation (Appendix G.3).

**AI is used in one place and supplied from another** Within weapons harm, AI *used* in conflict (5.1–5.3) is reported where fighting occurs, with Palestine and Ukraine holding 50 of 147 cases, whereas AI *supply* and financing (5.4) is reported mainly in the United States and Israel (11 of 18 cases) and names investors, cloud providers, and defense contractors rather than combatants (Appendix G.5). None of the four reviewed frameworks could label these 18 supply cases.

**New kinds of harm appear late and abruptly** Cases rise from 129 reported in 2021 to 889 in 2025, and the first seven months of 2026 (902) already exceed all of 2025 (Figure 5). Because recent news is archived more completely, we compare each category's share of a year rather than raw counts. Compute / Model Behavior is stable at 52–57% throughout. Energy / Land, nearly absent before 2023, grows from 2% of cases in 2023 to 13% in partial 2026, when its count (113) already exceeds all of 2025 (63) as data center build-out enters AI-harm reporting. Weapons harm jumps from 11 cases in 2023 to 41 in 2024, coinciding with the Gaza war and the ongoing war in Ukraine, and reaches 58 in partial 2026. Archive growth lifts all categories alike, so it cannot explain a single category's share nearly doubling within a year (Energy / Land 7% to 13%; weapons 3% to 7%). The map's cumulative growth in matched reports and country coverage over this period is shown in Appendix G.6.

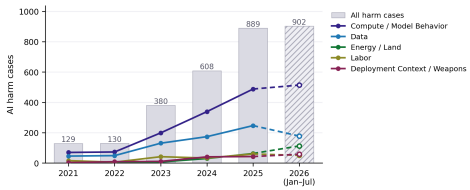



Figure 5: AI harm cases reported per year by our taxonomy category (a case in two categories counts in both); 2026 is partial. Because reporting coverage grows over time, shares within a year are more informative than raw counts.

**The map links cases that no single report connects** Following one actor, system, or affected group across the map reveals cross-border patterns. OpenAI appears in 19 countries and 18 of the 22 subcategories, whereas Clearview AI appears in 7 countries but only three data-harm subcategories. Facial recognition is the most widespread system (33 countries) and is labeled mainly as 4.2 Surveillance, which is also the leading harm for protesters, migrants, and Palestinians alike (Appendices G.7, G.8, and G.9).

### 4.4 LIMITATIONS AND FUTURE WORK

**Limitations** Our work has some limitations. **First,** the fixed taxonomy contains 5 categories and 22 subcategories, with $K_1/K_2$ keywords defined before collection; emerging or out-of-taxonomy

harms may therefore be missed. **Second,** the dataset covers January 2021–July 2026 and relies on text from 27 English-language news domains. It captures only reported AI harms, so observed prevalence reflects reporting and documentation patterns rather than actual harm levels. This limits coverage of non-English sources, social media, non-textual evidence, underrepresented regions, and community-reported harms that do not reach the selected news domains. The United States accounts for 57% of geographically placed cases, which may reflect stronger reporting and documentation capacity rather than higher harm levels. **Third,** automated retrieval may introduce selection bias through the predefined taxonomy, source set, and GDELT-based discovery, favoring more visible or frequently reported harms. **Fourth,** requiring sufficient evidence to link AI to a reported harm may favor harms that are easier to verify and better documented. **Fifth,** PHTKG inherits limitations from automated event extraction and entity resolution, and its provenance signal depends on extractor confidence. Its yearly temporal resolution also limits modeling of short-term dynamics. **Finally,** next-year recurrence and embedding similarity capture structural patterns rather than causal relationships. Evaluation uses 174 held-out predicted AI harm events; broader evaluation across corpora and corruption settings is needed to assess generalization.

**Future work** Future work will expand the dataset's coverage and verification pipeline while improving PHTKG's temporal modeling and evaluation, centering on civil-society partnerships for community-verified evidence and on scaling the interactive dynamic AI Harm Map (Appendix H).

### 4.5 Implications for Civil Society and Governance

Our taxonomy and dataset prototype were developed in a civil-society setting and shared with AI governance and human-rights practitioners. Their continued interest, including requests to document harms in their own contexts and build on the dataset, motivated its design as a shared evidence base that practitioners can contribute to, verify, and extend.

**Shared evidence for civil society** Civil society organizations document AI harm in isolation, with no shared evidence base to check or build on. Our dataset provides that shared, verifiable layer. Its taxonomy also closes two gaps in prior frameworks: an infrastructural definition of AI (Crawford, 2021) brings land and labor harm into scope, and a Deployment Context / Weapons category links AI-harm research to international humanitarian law and arms control. Appendix I details both corrections and the literature on community testimony as evidence.

**Governance** The map reveals an accountability gap that jurisdiction-bound national governance may not fully address. AI use and supply in armed conflict are geographically separated: reported harms are concentrated in Palestine, Ukraine, Iran, and Russia, while the investors, cloud providers, and defense contractors enabling these systems are named mainly in the United States and Israel. The same organizations recur across countries and harm categories; for example, OpenAI appears in 19 countries and 18 of 22 subcategories, and Meta in 18 countries and 18 subcategories. Moreover, 193 cases (6.4%) are not assignable to a specific jurisdiction because they are global or platform-based. These patterns support the need for coordinated cross-border monitoring that considers financiers and suppliers alongside system operators. We document this accountability gap without prescribing specific forms of redress. Appendix I provides further analysis and clarifies the scope of these findings.

## 5 Conclusion

We presented the first automated AI harm map, a novel framework that combines large-scale data collection, an AI harm taxonomy, and PHTKG, a data provenance-aware temporal hypergraph model for representing AI harms. Our PHTKG outperforms the strongest baseline (HYPER) (Huang et al., 2026) by **+4.68 AUC points**, and the remaining six baselines by **29.87 to 42.69 points**. The resulting dynamic map geolocates 2,845 cases across 81 countries shows that weapons harm is the least US-centered category, concentrating instead in conflict-affected countries. By bringing fragmented reports together in a public, interactive map, our framework provides a transparent way to study AI harms across regions and over time, helping researchers, policymakers, and AI developers identify emerging risks and improve AI safety and accountability.

### AI USE STATEMENT

AI assistants were used solely for editorial and coding support during paper preparation, including grammar correction, LaTeX formatting, table and figure layout, code debugging, and minor stylistic polishing. AI assistants were not used to generate experimental results or scientific claims. All scientific content, including the research questions, taxonomy design, dataset curation, experimental design, analyses, and conclusions, was conceived, reviewed, and verified by the human authors, who take full responsibility for the final manuscript.

### REPRODUCIBILITY STATEMENT

The map, code, and dataset used in this paper can be found here. The taxonomy and framework comparison are provided in Appendix B.1; the data collection procedure, including retrieval channels, source list, and keyword lexicon, is described in Appendix C. For the PHTKG model, complete architectural, optimization, and training configurations are provided in Appendix D.3.

### ETHICS STATEMENT

This work uses only information published by professional news outlets and public services. It collects no private information and involves no human-subjects research or interaction with affected individuals. Although the keyword lexicon includes terms for severe harms, the dataset contains no harmful content, only structured metadata about publicly reported incidents.

For politically contested and safety-critical cases, including armed conflict and AI-enabled weapons, inclusion requires an explicit evidentiary threshold (*occurred* or well-sourced *probable*). The dataset aggregates and structures information already publicly reported, without introducing new private information.

The AI harm locations are derived from public reporting, and the dataset is intended for accountability and harm prevention. To reduce classification errors and bias, the LLM-based pipeline was validated against 56 human-verified reference cases and uses conservative inclusion criteria.

## REFERENCES

AIAAIC. Aiaaic repository: User guide. `https://www.aiaaic.org/aiaaic-repository/user-guide`. Accessed: 2026-09-18.

Jiaxin Bai, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, et al. Autoschemakg: Autonomous knowledge graph construction through dynamic schema induction from web-scale corpora. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20557–20584, 2026.

Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations*, pp. 122–131, 2021.

Vincent Boulanin, Neil Davison, Netta Goussac, and Moa Peldán Carlsson. *Limits on autonomy in weapon systems: Identifying practical elements of human control*. Stockholm International Peace Research Institute, 2020.

Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, pp. 535–543, 2018.

J Scott Brennen, Philip N Howard, and Rasmus Kleis Nielsen. An industry-led debate: How uk media cover artificial intelligence. 2018.

Kate Crawford. *Atlas of AI*. Yale University Press, New Haven, 2021.

Boaventura de Sousa Santos. *Epistemologies of the South: Justice against epistemicide*. Routledge, 2015.

Joel Feinberg. *Harm to others*, volume 1. Oxford University Press, 1987.

Guardian News & Media. Guardian Open Platform: Content API Documentation, 2026. URL https://open-platform.theguardian.com/documentation/. Accessed: 2026-08-22.

Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

Mia Hoffmann and Heather Frase. Adding structure to AI harm: An introduction to CSET's AI harm framework. Technical report, Center for Security and Emerging Technology, Georgetown University, July 2023. URL https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

Xingyue Huang, Mikhail Galkin, Michael Bronstein, and Ismail I Ceylan. Hyper: A foundation model for inductive link prediction with knowledge hypergraphs. In *International Conference on Learning Representations (ICLR)*, 2026.

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6669–6683, 2020.

Jinwoo Kim, Saeyoon Oh, and Seunghoon Hong. Transformers generalize deepsets and can be extended to graphs & hypergraphs. *Advances in Neural Information Processing Systems*, 34: 28016–28028, 2021.

Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Jurewicz. Darkbench: Benchmarking dark patterns in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Kalev Leetaru and Philip A. Schrodt. GDELT: Global data on events, location and tone, 1979–2012. In *International Studies Association (ISA) Annual Convention*, 2013.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 85–99, 2024.

Xinyan Ma, Xianhao Ou, Weihao Zhang, Shixin Jiang, Runxuan Liu, Dandan Tu, Lei Chen, Ming Liu, and Bing Qin. Schema-aware planning and hybrid knowledge toolset for reliable knowledge graph triple verification. *arXiv preprint arXiv:2604.04190*, 2026.

Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15458–15463, 2021.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684, 2020.

James Muldoon, Callum Cant, Boxi Wu, and Mark Graham. A typology of artificial intelligence data work. *Big data & society*, 11(1):20539517241232632, 2024.

James Muldoon, Ana Valdivia, and Adam Badger. The politics of artificial intelligence supply chains. *AI & SOCIETY*, 41(2):1175–1187, 2026.

OECD.AI. OECD AI Incidents and Hazards Monitor (AIM), 2024. URL https://oecd.ai/en/incidents. OECD.AI Policy Observatory.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

Aníbal Quijano. Coloniality of power and eurocentrism in latin america. *International sociology*, 15(2):215–232, 2000.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741, 2023.

Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *SuperIntelligence-Robotics-Safety & Alignment*, 1(1), 2024.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *The Seventh International Conference on Learning Representations (ICLR)*, 2019.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9, 2021.

Xing Tang, Ling Chen, Hongyu Shi, and Dandan Lyu. Dhyper: A recurrent dual hypergraph neural network for event prediction in temporal knowledge graphs. *ACM Transactions on Information Systems*, 42(5):1–23, 2024.

The GDELT Project. GDELT 2.0 Event Database Codebook, 2015. URL https://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf.

Mike Thelwall and David Stuart. Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, 57(13):1771–1779, 2006.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *international conference on machine learning*, pp. 3462–3471. PMLR, 2017.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 2016.

Unicode Consortium. Unicode Standard Annex #15: Unicode Normalization Forms, 2023. URL https://unicode.org/reports/tr15/.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229, 2022.

WordPress. WordPress REST API Handbook, 2026. URL https://developer.wordpress.org/rest-api/. Accessed: 2026-08-22.

Ruxin Xue, Jinggui Huang, and Zaitang Huang. Hg-former: Utilizing multi-scale transformer links prediction based on hypergraph structure. *Neurocomputing*, 660:131891, 2026.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *The Third International Conference on Learning Representations (ICLR)*, 2015.

# A Extended Related Work and Baseline Details

The appendix expands on the AI harm resources and baseline descriptions summarized in Section 2. It states each baseline's working principle, its limitations, and why it's a fair comparison.

**AI harm resources** Research has also examined specific dimensions of AI-related harm over time. Boulanin et al. (2020) examine autonomy in weapon systems and the requirements for meaningful human control. Luccioni et al. (2024) quantify the energy costs of AI deployment, and Muldoon et al. (2024) examine AI data-work roles and labor conditions. More recently, Muldoon et al. (2026) connect AI labor and data work to the broader politics of AI supply chains. Together, these works provide evidence and methodological foundations for several harm domains covered by our taxonomy, including labor, military, and environmental harms.

**Knowledge graph embedding baselines** Our benchmark includes three standard knowledge graph embedding models. DistMult (Yang et al., 2015) scores a triple with a diagonal bilinear form. ComplEx (Trouillon et al., 2016) extends this to complex space, which lets it represent asymmetric relations.RotatE (Sun et al., 2019) treats a relation as a rotation in complex space, which lets it capture composition patterns that bilinear models cannot. All three learn a single embedding per entity and per relation, and thus cannot generalize to unseen entities or relations. They also break down a seven-role event into independent triples, throwing away the joint configuration of its participants. We include them to measure the loss of a pairwise decomposition with respect to a single role-aware hyperedge.

**Hypergraph and set-based baselines** We also evaluate four models that function within a multi-participant framework. Deepsets-HG (Kim et al., 2021) encodes a hyperedge by summing participant embeddings, thereby maintaining joint structure while treating the hyperedge as an unordered set. SetTransformer-HG substitutes this summation with self-attention among participants, facilitating interaction; however, it remains permutation-invariant, time-agnostic, and confidence-agnostic. Temporal-SetTransformer (Xue et al., 2026) incorporates temporal aggregation, enabling it to represent the timing of events, yet it does not utilize extraction confidence nor does it incorporate a recurrence objective. HYPER (Huang et al., 2026) serves as the most robust baseline. It is a foundational model for knowledge hypergraphs, capable of generalizing to unseen entities and relations of any arity by learning the interaction of relations through their argument positions. HYPER (Huang et al., 2026) supports continuous time, but it does not model extraction confidence nor expose a recurrence objective. It is also pre-trained on knowledge hypergraphs with a different structure from our seven-role harm events. All four are included to test if joint hyperedge modeling is enough on its own, or if role, time and data provenance conditioning each contribute independently.

**Temporal knowledge graphs** Temporal knowledge graph models are designed to capture the dynamic nature of facts. Know-Evolve (Trivedi et al., 2017) employs a temporal point process to model the timing of events, while RE-Net (Jin et al., 2020) utilizes a recurrent architecture to aggregate information over time.DHyper (Tang et al., 2024) integrates hypergraph representations with recurrent temporal aggregation; however, it does not incorporate extraction confidence and lacks gating of message passing based on reliability.

**Fair comparison** All baselines are trained and evaluated on an identical predicted-event test set, utilizing the same chronological split as outlined in Section 4.2. The triple-decomposition baseline employs the same corruption procedure and binary classification objective as PHTKG, thereby isolating the impact of joint hyperedge modeling from any variations in training signals. Although HYPER is pretrained on a different corpus, we employ its published checkpoint and fine-tune it on the same test set following the same evaluation protocol. This approach ensures that the comparison highlights architectural differences rather than tuning variations.

# B AI Harm Taxonomy, Collection and Classification

This appendix provides the complete taxonomy, detailed category verdicts, and two end-to-end examples of the inclusion protocol.

### B.1 Full Taxonomy and Comparison Against Existing Frameworks

The taxonomy comprises 22 subcategories across 5 categories (Table A1). Each subcategory is assigned an *Inherited*, *Adapted*, or *Original* verdict through literature-based comparison with four reviewed AI harm taxonomies (Table A2). This comparison is disclosed at the subcategory level rather than used to claim taxonomy-level novelty: 11 subcategories are *Inherited*, 5 *Adapted*, and 6 *Original*.

Table A1: The AI Harm Map taxonomy: 5 Categories, 22 Subcategories.

| # | Category | Subcategory | Short definition | Protected interest | Event type |
|---|---|---|---|---|---|
| 1.1 | Compute / Model Behavior | Representational harm | System reinforces stereotypes or subordination of a social group through its outputs. | Dignity, non-discrimination | Acute |
| 1.2 | Compute / Model Behavior | Allocative harm | Opportunity, resources, or information withheld from a person or group by an algorithmic decision. | Livelihood, non-discrimination | Acute |
| 1.3 | Compute / Model Behavior | Misinformation / information hazard harm | System generates or disseminates false or misleading information causing material harm. | Autonomy, safety | Acute |
| 1.4 | Compute / Model Behavior | Malicious use harm | System is used to facilitate fraud, cyberattack, targeted manipulation, or illegitimate surveillance. | Safety, autonomy, privacy | Acute |
| 1.5 | Compute / Model Behavior | Human computer interaction harm | Overreliance, manipulation, or unsafe use enabled by anthropomorphized or persuasive system design. | Autonomy, psychological wellbeing | Chronic |
| 1.6 | Compute / Model Behavior | Quality of service harm | System fails disproportionately for some groups relative to others. | Non-discrimination, livelihood | Chronic |
| 1.7 | Compute / Model Behavior | Physical safety harm | System malfunction, erroneous output, or control failure causes physical injury or death, outside a military deployment context. | Life, bodily integrity | Acute |
| 1.8 | Compute / Model Behavior | Interpersonal / relational harm | AI generated or AI enabled tooling is used to target a specific individual or relationship. | Dignity, autonomy, personal safety | Acute |
| 2.1 | Energy / Land | Environmental resource harm | Energy, water, or emissions impact from operating AI compute infrastructure. | Environment, health | Chronic |
| 2.2 | Energy / Land | Land use and siting harm | Community displacement, land use conflict, or governance failure from AI infrastructure siting. | Livelihood, environment, self determination | Acute and chronic |
| 2.3 | Energy / Land | Resource extraction harm | Mining or material extraction for AI hardware causing environmental or community harm. | Environment, livelihood, health | Chronic |
| 3.1 | Labor | Data labor exploitation | Exploitative conditions in data annotation, content moderation, or other AI supply chain labor. | Livelihood, dignity, psychological wellbeing | Chronic |
| 3.2 | Labor | Algorithmic management harm | Wage suppression, surveillance, or unsafe conditions from AI driven scheduling or performance management. | Livelihood, privacy, dignity | Chronic |
| 3.3 | Labor | Employment displacement / quality harm | Job loss or decline in job quality attributable to AI adoption. | Livelihood | Chronic |
| 4.1 | Data | Privacy violation | System leaks or improperly infers private information about a person. | Privacy | Acute |
| 4.2 | Data | Surveillance harm | AI enabled monitoring used for illegitimate surveillance or censorship of individuals or groups. | Privacy, autonomy, civic space | Chronic |
| 4.3 | Data | Data rights / consent violation | Data collected, used, or sold without valid consent or legal basis. | Privacy, autonomy | Acute |
| 4.4 | Data | Creative / IP appropriation harm | Creative work used as AI training data or reproduced by AI output without consent or compensation. | Livelihood, intellectual property rights | Chronic |
| 5.1 | Deployment Context / Weapons | Autonomous weapons / lethal targeting harm | Death, injury, or destruction from AI assisted target selection or engagement. | Life, bodily integrity | Acute |
| 5.2 | Deployment Context / Weapons | Military decision support harm | AI generated intelligence or targeting recommendations adopted with minimal human review. | Life, due process | Acute |
| 5.3 | Deployment Context / Weapons | Accountability gap harm | Harm for which no actor can be identified as responsible due to AI enabled decision making. | Right to remedy | Acute and chronic |
| 5.4 | Deployment Context / Weapons | Supply chain / investor complicity harm | Producing, exporting, or financing AI systems that inflict harm elsewhere, including arms transfer and investment in military AI suppliers. | Life, right to remedy | Chronic |

Table A2: Verdict and basis for each subcategory.

| # | Subcategory | Verdict | Basis |
|---|---|---|---|
| 1.1 | Representational harm | Inherited | Well covered; Weidinger and Shelby both name it directly. |
| 1.2 | Allocative harm | Inherited | Shelby et al. is the clearest source. |
| 1.3 | Misinformation / information hazard harm | Inherited | MIT and Weidinger both name it. |
| 1.4 | Malicious use harm | Inherited | MIT and Weidinger both name it. |
| 1.5 | Human computer interaction harm | Inherited | MIT and Weidinger both name it. |
| 1.6 | Quality of service harm | Inherited | Shelby et al. names it directly. |
| 1.7 | Physical safety harm | Inherited | CSET and MIT both name it; gap was in an earlier draft, not the literature. |
| 1.8 | Interpersonal / relational harm | Inherited | Shelby et al. names it directly; earlier draft had folded it incorrectly into 1.1. |
| 2.1 | Environmental resource harm | Adapted | Named in the literature but not geolocated or facility verified. |
| 2.2 | Land use and siting harm | Original | Not covered by any of the four reviewed taxonomies. |
| 2.3 | Resource extraction harm | Adapted | Partial coverage in Weidinger and MIT. |
| 3.1 | Data labor exploitation | Original | Not covered by any of the four reviewed taxonomies. |
| 3.2 | Algorithmic management harm | Adapted | Partial coverage in MIT under devaluation of human effort. |
| 3.3 | Employment displacement / quality harm | Inherited | MIT covers this reasonably well. |
| 4.1 | Privacy violation | Inherited | MIT and Weidinger both name it. |
| 4.2 | Surveillance harm | Adapted | State and military scale surveillance not fully covered by any framework. |
| 4.3 | Data rights / consent violation | Adapted | Partial coverage across CSET, MIT, Weidinger. |
| 4.4 | Creative / IP appropriation harm | Inherited | Weidinger names it directly. |
| 5.1 | Autonomous weapons / lethal targeting harm | Original | Developed in IHL/arms control literature, not previously integrated into an AI harm taxonomy. |
| 5.2 | Military decision support harm | Original | Same as above. |
| 5.3 | Accountability gap harm | Original | This is the accountability gap concept from IHL scholarship. |
| 5.4 | Supply chain / investor complicity harm | Original | No reviewed framework reaches the actors who produce, export, or finance AI systems used to cause harm elsewhere. Added in response to UNMAPPED labels. |

### B.2 Worked Examples: Case Inclusion Under the Manual Protocol

These two examples illustrate the inclusion protocol end to end, including the occurrence-status gate for cases where the AI system's causal role is contested.

**Compute / model behavior: COMPAS recidivism risk scores (1.1, 1.2)** A risk assessment algorithm used in Broward County, Florida courts was found to flag Black defendants as high risk at nearly twice the rate of white defendants relative to actual reoffense rates. Occurrence status: occurred. Geolocation: Broward County Circuit Court, the deploying institution. Source: ProPublica, *Machine Bias*, 2016.

**Deployment context / weapons: Lavender and The Gospel targeting systems, Gaza (5.1, 5.3)** Reporting indicates the Israeli military used AI systems to generate a list of tens of thousands of potential targets in Gaza with minimal human review before airstrikes were authorized. Occurrence status: base facts treated as occurred; the AI system's causal role is contested by the Israeli military and is treated as an allegation under active corroboration, not settled fact — an example of the occurrence-status gate admitting a case on well-evidenced base facts while explicitly flagging the contested element rather than either excluding the case outright or treating the contested claim as resolved. Geolocation: specific strike sites within Gaza. Sources: +972 Magazine, 2024; also logged in the AI incident Database.

## C Automatic retrieval and classification

This appendix details the automatic retrieval and classification pipeline, from multi-channel article retrieval and standardized record construction to keyword-based filtering and the resulting corpus.

### C.1 Retrieval Channels

Articles are retrieved through three channels (Table A3): WordPress REST APIs, the Guardian Content API, and GDELT-discovered URLs. All records are mapped to a common schema and filtered locally.

**WordPress REST API** Seven domains expose public WordPress APIs (WordPress, 2026). The pipeline verifies `/wp-json/wp/v2/posts` or `/?rest_route=/wp/v2/posts`, then retrieves posts within the date window using `after`/`before`, newest first, 100 per page (maximum 50 pages). Requests are parallelized across domains with per-domain delays (Thelwall & Stuart, 2006). Records with missing, invalid, or out-of-window `date_gmt` are discarded.

**Guardian content API** The Guardian is queried through `/search` (Guardian News & Media, 2026) using the same date window and ordering. `webUrl`, `fields.headline`, `fields.standfirst`/`trailText`, `fields.bodyText`, and `webPublicationDate` are mapped to the common schema. Filtering remains local because API matching does not support wildcard truncation such as `algorithm*`.

**GDELT-discovered URLs** For 19 domains without usable archive APIs, URLs are obtained from the GDELT 2.0 Event Database (Leetaru & Schrodt, 2013; The GDELT Project, 2015), restricted to its English stream. Yearly files are merged, deduplicated by URL using the earliest collection timestamp, and partitioned by source. Article text is extracted using a cascade of (1) Schema.org JSON-LD (Guha et al., 2016), (2) Trafilatura (Barbaresi, 2021), and (3) an HTML heuristic fallback.

The extractor yielding the longest body is retained and recorded. Bodies shorter than 600 characters are labeled `thin` and excluded.

Table A3: Retrieval channels and source counts.

| Channel | Access method | Domains | Languages |
|---|---|---|---|
| WordPress REST API | `/wp/v2/posts`, public, no key | 7 | en |
| Guardian Content API | `/search`, developer-tier key | 1 | en |
| GDELT-discovered URLs | direct HTTP fetch of article URLs | 19 | en |
| Distinct domains | | 27 | |

### C.2 Record Construction and Output Schema

All three channels emit the same 11-column schema (Table A4), with lexicon-match fields storing the matched entries rather than counts, making each filtering decision directly inspectable. Each source is assigned a stable four-character code, language, and retrieval channel (Table A3), and each article receives a `CODE-YYYY-NNNNNN` identifier. Articles are ordered oldest-to-newest before ID assignment, allowing new data to be appended without changing existing IDs; duplicate URLs are removed.

Table A4: Output schema, identical across the three channels.

| Column | Contents |
|---|---|
| `id` | `CODE-YYYY-NNNNNN` identifier |
| `source_language` | lexicon language used for matching |
| `source_name` | source domain |
| `title` | |
| `standfirst` | subtitle/standfirst, or the first 400 characters of the body if absent |
| `all_new_text` | the concatenated search text $T(d)$ actually passed to the matcher |
| `source_url` | canonical article URL, used as the deduplication key |
| `publication_date` | ISO 8601, UTC |
| `country_of_news` | source country, where recorded in the configuration |
| `keyword_1st_group_founded` | the matched $K_1$ entries, pipe-separated |
| `keyword_2nd_group_founded` | the matched $K_2$ entries, pipe-separated |

### C.3 The keyword lists

**AI list ($K_1$)**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AI, | AGI, | intelligen*, | learning, | model*, | model*, | model*, | model*, |
| LLM*, | ML, | NLP, | machine | neural | foundation | language | generative | transformer | chatbot*, |
| RLHF, | GPT*, | learning, | net*, large | model*, | model*, | ai, | model*, | conversational |
| artificial | deep | language | frontier | generative | multimodal | diffusion | agent*, ai |

assistant*, ai agent*, agentic*, ai system*, ai model*, ai tool*, ai power*, ai driven, ai generat*, ai enabl*, ai assist*, algorithm*, automated decision*, automated system*, decision support*, predictive analytic*, predictive model*, recommender system*, recommendation algorithm*, ranking algorithm*, computer vision, speech recognition, natural language process*, model output*, training data, risk score*, risk assessment tool*, scoring algorithm*, automated moderation, self driving, autonomous vehicle*, driverless, robotaxi*, autopilot, driver assist*, surgical robot*, diagnostic algorithm*, triage algorithm*, data cent*, server farm*, compute cluster*, ai infrastructure, compute infrastructure, training run*, model training, inference workload*, data annotat*, data label*, annotator*, data worker*, content moderat*, automated scheduling, productivity monitor*, worker surveillance, bossware, training corpus, training corpora, facial recognition, face recognition, biometric*, emotion recognition, gait recognition, predictive polic*, behavioural inference, behavioral inference, profiling algorithm*, voice clon*, deepfake*, synthetic media, synthetic imag*, autonomous weapon*, lethal autonomous*, killer robot*, loitering munition*, target recognition, targeting system*, target selection, ai targeting, military ai, defense ai, defence ai, intelligence analysis system*, drone swarm*, armed drone*, unmanned aerial*, openai, chatgpt, anthropic, deepmind, google gemini, meta llama, microsoft copilot, deepseek, midjourney, character ai, chip*, AI chip*, chipmaker*, semiconductor*, GPU*, TPU*, AI accelerator*, wafer*, foundry, Nvidia, TSMC, ASML, AI hardware, hardware supply chain, hyperscale*, colocation, AI data cent*, training cluster*, supercomputer*, AI factory, power purchase agreement*, megacampus, Sama, Scale AI, Appen, Remotasks, Mechanical Turk, Toloka, Clickworker, Surge AI, red team*, human feedback, human in the loop, data enrichment, ghost work, algorithmic management, gig platform*, dispatch algorithm*, driver deactivation, rating algorithm*, keystroke logging, productivity score*, workforce analytics, time tracking software, warehouse algorithm*, route optimi*, credit scoring, tenant screening, hiring algorithm*, resume screening, applicant tracking system*, automated eligibility, benefits algorithm*, fraud detection algorithm*, welfare algorithm*, claims algorithm*, utilization management algorithm*, algorithm*, sentencing algorithm*, parole algorithm*, child welfare algorithm*, safety filter*, alignment, fine tun*, system prompt*, content classifier*, moderation model*, reinforcement learning, synthetic data, license plate reader*, automatic number plate recognition, ANPR, iris recognition, voice recognition, lie detection, emotion AI, aggression detection, gunshot detection, ShotSpotter, social media monitoring, Lavender, The Gospel, Project Maven, Maven Smart System, Palantir, Anduril, battle management system*, command and control, ISR, fire control, automatic target recognition, decision advantage, Clearview AI, Stability AI, xAI, Grok, Mistral, Perplexity, Hugging Face, Cohere, ElevenLabs, Suno, Runway, Sora, Claude, Gemini, Llama, LAION, Common Crawl, Books3, The Pile, ImageNet, web crawl*.

## Harm ($K_2$) list

harm, harms, harmed, harmful, damag*, injur*, victim*, abus*, misus*, violat*, breach*, misconduct, negligen*, wrongdoing, scandal*, controvers*, backlash, outcry, alleged, allegedly, allegation*, accus*, lawsuit*, litigat*, class action*, complaint*, fined, fines, penalt*, sanction*, regulator*, enforcement action*, banned, liab*, discriminat*, bias*, racis*, sexis*, stereotyp*, prejudic*, disparate impact, disproportionate, unequal treatment, wrongful*, wrongly*, misidentif*, false match*, denied benefit*, marginalis*, marginaliz*, dehumanis*, dehumaniz*, degrading, misinformation, disinformation, fake news, false claim*, fabricated, misleading, propaganda, manipulated media, election interference, astroturf*, information operation*, fraud*, scam*, phishing, impersonat*, identity theft, cyberattack*, hacked, hacking, hacker*, ransomware, malware, extort*, money launder*, social engineering, psychological harm, trauma*, PTSD, distress*, anxiet*, depression, addict*, compulsive*, manipulat*, overrelian*, self harm*, suicid*, vulnerable, death*, died, kill*, fatality, fatalities, casualt*, crash*, collision*, malfunction*, defect*, hospitalis*, hospitaliz*, wounded, safety incident*, misdiagnos*, harass*, stalk*, cyberstalk*, doxx*, non consensual*, nonconsensual*, intimate imag*, nudif*, revenge porn*, sexual exploitation, defam*, bully*, blackmail*, sextortion, water use, water consumption, water scarcit*, drought*, aquifer deplet*, emission*, carbon footprint*, greenhouse gas*, pollut*, contaminat*, toxic waste, e waste, noise pollution, energy consumption, grid strain, electricity price*, blackout*, deforest*, biodiversity loss, land grab*, displac*, evict*, expropriat*, community opposition, protest, protests, protesters, protestors, environmental justice, exploit*, low pay, underpaid, poverty wage*, wage theft, wage suppression, unpaid overtime, precarious*, unsafe condition*, working condition*, union bust*, unionis*, unioniz*, strike*, retaliat*, layoff*, job loss*, job cut*, redundanc*, deskill*, burnout, quota pressure, privacy violation*, privacy breach*, data breach*, data leak*, without consent, lack of consent, unlawful processing, GDPR, data protection, unlawful surveillance, mass surveillance, spying, censor*, chilling effect*, repress*, re identif*, copyright*, infring*, without permission, without compensation, unlicensed, plagiaris*, plagiariz*, lost income, airstrike*, bombing*, shelling, civilian death*, civilian casualt*, civilian harm, war crime*, atrocit*, unlawful killing*, extrajudicial, international humanitarian law, indiscriminate, collateral damage, accountability, gap*, impunity, lack of oversight, no human review, no recourse, denied remedy, opaque decision*, unexplained decision*, unable to appeal, responsibility gap*, denied due process, unpaid royalt*, targeting of activist*, targeted for surveillance, hate speech, toxicity, toxic output*, slur*, demean*, denigrat*, misgender*, caste discriminat*, islamophob*, antisemit*, xenophob*, homophob*, transphob*, ableis*, racial stereotyp*, harmful stereotype*, subordinat*, underrepresent*, offensive content, denied loan*, loan denial*, credit denial*, denied credit, mortgage denial*, denied housing, welfare fraud, benefits cut off, eligibility determinat*, wrongly flagged, wrongly denied, clawback*, overpayment notice*, debt notice*, rejected applicant*, screened out, denied coverage, coverage denial*, claims denial*, prior authorization denial*, denied

```
care, denied parole, denied
bail, denied admission, false positive*, false negative*, hallucinat*, confabulat*, fabricated citation*, fabricated source*, made up, false information, inaccurate information, conspiracy theor*, voter suppression, synthetic text, false attribution, CSAM, child sexual abuse material, child abuse imagery, grooming, voice scam*, vishing, romance scam*, jailbreak*, prompt injection, bioweapon*, biosecurity, chemical weapon*, botnet*, spam campaign*, business email compromise, credential theft, anthropomorph*, parasocial, emotional depend*, companion app*, dark pattern*, sycophan*, engagement optimi*, unhealthy attachment, delusion*,
psychosis, encouraged self harm, isolation, error rate*, accuracy gap*, accuracy disparit*, underperform*, fails for, failure rate*, darker skin, skin tone, accent*, dialect*, transcription error*, mistranslat*, low resource language*, safety recall*, product recall*, NHTSA, struck a pedestrian, pedestrian, run over, electrocut*, amputat*, burn injur*, workplace accident*, industrial accident*, robot injur*, crushed, near miss, phantom braking, unintended acceleration, adverse event*, patient harm, sensor failure*, swatting, coercive control, intimate partner, domestic abuse, gender based violence, image based abuse,
synthetic nude*, undress app*, tracked without consent, ratepayer*, utility bill*, electricity bill*, energy bill*, diesel generator*, backup generator*, gas turbine*, peaker plant*, air quality, particulate*, PM2.5, smog, asthma, respiratory, wastewater, water withdrawal*, potable water, cooling water, megawatt*, gigawatt*, power demand, transmission line*, fossil fuel*, coal plant*, climate target*, grid capacity, tax abatement*, tax break*, tax incentive*, subsid*, zoning, rezoning, moratorium, public hearing*, planning permission, permit denial*, eminent domain, farmland, agricultural land, sacred
site*, indigenous land, land rights, resettle*, property value*, nondisclosure agreement*, shell compan*, community consultation, water rights, host community, mining, mine*, miner*, cobalt, lithium, rare earth*, nickel, coltan, tantalum, tungsten, bauxite, silica, tailings, artisanal mining, child labour, child labor, forced labour, forced labor, smelter*, mine collapse, land degradation, acid drainage, conflict mineral*, resource extraction, outsourc*, subcontract*, gig work*, crowdwork*, click work*, microwork*, piece rate, no sick leave, mental health support, counsel*, graphic content,
disturbing content, sub minimum wage, mass dismissal*, wrongful dismissal, visa status, refugee worker*, algorithmic management, automated management, deactivat*, automated firing, fired by algorithm, terminated by algorithm, time theft, unpredictable schedul*, just in time schedul*, clopening, pace of work, injury rate*, bathroom break*, productivity quota*, algorithmic wage discriminat*, personalized pay, opaque ranking, replaced by AI, replaced by automation, automated away, hiring freeze*, headcount reduction, restructur*, contract cancel*, contract terminat*, fewer commissions, rate cut*, wage stagnation, entry level job*, PII, personally identifiable, memoriz*,
training data extraction, model inversion, membership inference, sensitive data, health data, medical record*, location data, inference about, private conversation*, chat log*, leaked prompt*, exposed database, wrongful arrest*, false arrest*, arrest*, detained, detention, spyware, Pegasus, wiretap*, dissident*, human rights defender*, journalist*, watchlist*, no fly list, social credit, Uyghur, ethnic profiling, racial profiling, immigration enforcement, border control, checkpoint*, deportation, student monitoring, school surveillance, protest monitoring, scrap*, web scraping, scraped image*, scraped data, data broker*, sold data,
opt out, terms of service change, privacy policy change, secondary use, repurpos*, BIPA, biometric privacy, CCPA, COPPA, children s data, Garante, CNIL, Information Commissioner, consent decree, legitimate interest, unlawful basis, fair use, likeness, right of publicity, voice likeness, in the style of, style mimic*, derivative work*, licensing deal*, residual*, opt out of training, trained on my, without attribution, uncredited, market dilution, lost commission*, kill chain, kill list, target list*, target bank, strike approved, proportionality principle of distinction, precautions in attack, drone strike*, munition*, friendly
fire, non combatant*, protected person*, Geneva Convention*, Article 36 review, ICRC, SIPRI, meaningful human control, automation bias, rubber stamp*, minimal human review, cursory review, seconds per target, human in the loop, human on the loop, target recommendation*, unverified intelligence, intelligence failure, wrong target, mistaken identity, over trust, black box, no explanation, unexplainable, trade secret*, proprietary system, unable to contest, no redress, no appeal, liability shield, immunity, unaccountable, who is responsible, no audit trail, refused to disclose, no transparency.
```

### C.4 The Filtering Rule

The filtering rule standardizes article text and applies boundary-aware keyword matching and list-specific thresholds.

**Text preparation** For each article $a$, the search text is the concatenation of its non-empty title, standfirst, and body: $T(a) = \text{title}(a) + \text{standfirst}(a) + \text{body}(a)$. HTML is converted to plain text, whitespace is collapsed, and Unicode NFKC normalization followed by NFD decomposition removes combining marks (Unicode Consortium, 2023). Both case-folded and case-preserving forms are retained. Each keyword is prepared once as its original form, first word, regular expression, and case-sensitivity flag, with the first word used as a fast pre-filter before applying the full regular expression.

**Word boundaries** Keywords match only when not adjacent to letters; digits, hyphens, apostrophes, and punctuation act as boundaries. Thus, `GPT*` matches *GPT-4* and `AI` matches *AI-powered*, while `harm*` excludes *charm*.

**Truncation** A trailing `*` enables right-side prefix matching: `algorithm*` matches *algorithm*, *algorithmic*, and *algorithmically*, but not *logarithm*. Truncation is used in 47% of English $K_1$ and 50% of English $K_2$ entries to capture inflections and derivations without stemming.

**Internal separators** Whitespace in multi-word entries matches whitespace, hyphens, or underscores; thus `machine learning` matches *machine learning*, *machine-learning*, and *machine_learning*, but not *machinelearning*.

**Acronyms** Uppercase alphabetic entries of up to six characters are matched case-sensitively, including 14 English $K_1$ entries such as `AI`, `ML`, `NLP`, `AGI`, and `LLM*`. Table A5 summarizes the resulting lexicon sizes and structural properties.

**Threshold selection** We count distinct matched entries, not occurrences; thus, fifty occurrences of "AI" still yield $|K_1(a)| = 1$. We set $|K_1(a)| \geq 3$ and $|K_2(a)| \geq 2$ based on their vocabulary structures and inspection of retrieved articles across threshold combinations, prioritizing precision.

Table A5: Lexicon sizes. "Trunc." counts entries ending in the truncation `*`; "Acr." counts entries treated as case-sensitive acronyms; "Multi." counts multi-word entries.

| | $K_1$ (AI) | | | | $K_2$ (harm) | | | |
|---|---|---|---|---|---|---|---|---|
| Language | Entries | Trunc. | Acr. | Multi. | Entries | Trunc. | Acr. | Multi. |
| English (en) | 231 | 109 | 14 | 167 | 588 | 294 | 11 | 342 |

### C.5 COUNT AFTER FILTERING

Table A6 summarizes the number of news articles retrieved from each source before and after applying the proposed filtering procedure.

Table A6: Number of retrieved news extracted before/after filtering

| Source Name | Group | News before filtering | News after filtering |
|---|---|---|---|
| 404media.co | GDELT URLs | 57 | 9 |
| aljazeera.com | GDELT URLs | 52038 | 143 |
| bbc.com | GDELT URLs | 50492 | 348 |
| dailymaverick.co.za | GDELT URLs | 29381 | 78 |
| english.elpais.com | GDELT URLs | 10593 | 109 |
| equaltimes.org | GDELT URLs | 513 | 24 |
| euobserver.com | GDELT URLs | 5820 | 56 |
| independent.co.uk | GDELT URLs | 84442 | 217 |
| time.com | GDELT URLs | 6954 | 11 |
| wired.com | GDELT URLs | 4404 | 1068 |
| abc.net.au | GDELT URLs | 65670 | 233 |
| biometricupdate.com | GDELT URLs | 5050 | 3946 |
| mg.co.za | GDELT URLs | 10008 | 21 |
| militarytimes.com | GDELT URLs | 2332 | 8 |
| morningstaronline.co.uk | GDELT URLs | 15689 | 10 |
| rferl.org | GDELT URLs | 19794 | 6 |
| scroll.in | GDELT URLs | 21069 | 47 |
| thenation.com | GDELT URLs | 7018 | 35 |
| theverge.com | GDELT URLs | 7577 | 819 |
| theguardian.com | The Guardian | 406415 | 5337 |
| 972mag.com | Wordpress | 1445 | 99 |
| digitaltrends.com | Wordpress | 58903 | 1663 |
| euromaidanpress.com | Wordpress | 30096 | 268 |
| motherjones.com | Wordpress | 10912 | 197 |
| rappler.com | Wordpress | 110317 | 1344 |
| theintercept.com | Wordpress | 4658 | 173 |
| vice.com | Wordpress | 48218 | 696 |

### C.6 AI HARM CLASSIFICATION: FULL PROCEDURE

**Protocol** Classification is performed per source by an LLM following one source-agnostic written protocol. The input for each source is the set of scraped articles, each with its title, standfirst, full body, URL, publication date, and the AI and harm keywords matched by the filter. For every source the agent produces two files — a single-label and a multi-label version of the same records — and a coverage report stating how many articles were scraped, kept, and discarded, the reason for each discard, the subcategory distribution, the countries covered, and any judgment calls it was unsure about.

**Pre-processing** Publishers repeat navigation and related-link blocks on every page, and a single AI-bearing block makes every article from that site pass the keyword filter. The agent detects such blocks by finding AI-matching sentences that recur across a large fraction of a source's pages; on +972 Magazine, a related-link headline about the *Lavender* targeting system appeared on 91 of 99 pages. The specific repeated string is removed before keywords are counted, rather than the words in it, so the article it links to is still admitted. The agent also repairs character-encoding errors (e.g. in 77 of 78 Daily Maverick records), drops 69 duplicate articles published under different URLs, and excludes non-article pages such as section indexes, author profiles, and live blogs.

**Inclusion rule** An article is included only if it documents a case in which an identifiable AI system caused an adverse effect on a protected interest — life, bodily integrity, livelihood, environment, autonomy, dignity, or non-discrimination — regardless of intent. Each candidate is assigned an occurrence status: *occurred* (documented as having happened), *probable* (strongly evidenced but contested or partly sourced), *implied*, or *speculative*. Only occurred cases, and probable cases that meet the same sourcing bar, are admitted. The agent discards four recurring false-positive types: a keyword used in another sense (e.g. *LLM* as a Master of Laws, or *The Gospel* as scripture rather than an Israeli targeting system), a passing mention of AI with no incident, a harm in which the AI system played no causal role (a human-piloted drone strike is not subcategory 5.1), and governance news with no implicit event.

**Source-specific rule** biometricupdate.com is a biometrics and digital-identity trade publication: every one of its 3,946 articles mentions biometrics, so the keyword filter does no work there. For this source, acceptance additionally requires a named deployment or incident, an identifiable affected population, and a documented adverse outcome, such as a wrongful arrest or misidentification, a breach of real records, or a regulator finding a deployment unlawful. A harm pattern applied to titles and standfirsts produced 681 candidates, which were reviewed individually against this rule; 207 were admitted. National digital-ID rollouts, vendor launches, contract awards, detection products, and legislative news are excluded by rule.

**Labeling** Each admitted article is assigned every applicable subcategory; a single primary label is then chosen by the severity of the protected interest, in the fixed order $5.1 > 1.7 > 5.2 > 5.3 > 5.4 > 1.8 > 4.2 > 1.4 > 4.1 > 4.3 > 1.2 > 1.3 > 1.1 > 1.6 > 3.1 > 3.2 > 3.3 > 2.1 > 2.2 > 2.3 > 1.5 > 4.4$. The acute/chronic event type is copied from the taxonomy rather than inferred per article, and the country is where the harm occurred, not the location of the publisher or the vendor. An article that fits no subcategory is marked UNMAPPED, so that gaps in the taxonomy remain countable. The 21 UNMAPPED labels revealed two gaps — the weapons supply chain, including investor complicity, and labor in the compute supply chain — which we resolved by adding subcategory 5.4 and assigning the latter to 3.3. Subcategory 5.4 is placed directly after 5.3 in the severity ordering, since both protect the right to remedy and 5.4 additionally protects life; all 18 cases labeled 5.4 carry it as their only subcategory, so its position does not change any primary label.

**Output checks** Every output file is checked for a fixed field order, sequential identifiers, event types that match the taxonomy, a null subcategory identifier if and only if the label is UNMAPPED, no duplicate URLs, and an occurrence status of only *occurred* or *probable*. Table A7 lists the fields of a record.

**Example** A +972 Magazine report that the Israeli army "marked tens of thousands of Gazans as suspects for assassination, using an AI targeting system with little human oversight" is admitted

as *occurred*, with country Palestine. It is labeled 5.1 (Autonomous weapons / lethal targeting), because an identified AI system selected targets and deaths followed, and 5.3 (Accountability gap), because no actor is identifiable as responsible; its primary label is 5.1. A report on a human-piloted quadcopter firing on civilians is discarded: the harm is severe, but the source never establishes that AI played a role in selecting the target.

**Verification** A second model audits the classification with a separate prompt that instructs it not to assume the first model was correct, to independently re-derive both the exclusions and the labels, and to report every disagreement, concentrating on judgment-call exclusions and on the 5.1/5.2 and 4.1/4.2/4.3 boundaries (the audit prompt is specified for +972 Magazine's 99 articles; a separate audit of the `country` field flagged 31 of 755 US/Global records, 17 of which were misassigned and corrected).

**Result** Table A8 gives the per-source funnel. Acceptance rates track how close a publication's beat sits to AI harm, from 5% on biometricupdate.com to 61% on The Intercept. Of the 2,689 admitted articles, 2,455 are occurred and 234 probable; 2,329 carry one subcategory, 357 two, and 3 three.

Table A7: Fields of a case record produced by the classifier.

| Field | Content |
|---|---|
| `report_id` | Sequential identifier of the admitted case |
| `original_text` | Opening of the article body, with title and boilerplate removed |
| `source_url`, `publication_date` | Provenance of the report |
| `country` | Where the harm occurred, inferred from the article |
| `occurrence_status` | *occurred* or *probable* |
| `taxonomy_hints` | AI and harm keywords matched by the keyword filter |
| `classification(s)` | category, subcategory, and event type: the primary label, plus every applicable label in the multi-label version |

Table A8: Classification funnel per source.

| Source | Scraped | Kept | % | Source | Scraped | Kept | % |
|---|---|---|---|---|---|---|---|
| *WordPress REST API* | | | | *GDELT-discovered URLs (cont.)* | | | |
| 972mag.com | 99 | 45 | 45 | english.elpais.com | 109 | 38 | 35 |
| digitaltrends.com | 1,663 | 239 | 14 | dailymaverick.co.za | 78 | 15 | 19 |
| euromaidanpress.com | 268 | 43 | 16 | euobserver.com | 56 | 14 | 25 |
| motherjones.com | 197 | 99 | 50 | scroll.in | 47 | 25 | 53 |
| rappler.com | 1,344 | 393 | 29 | thenation.com | 35 | 13 | 37 |
| theintercept.com | 173 | 105 | 61 | equaltimes.org | 24 | 7 | 29 |
| vice.com | 696 | 238 | 34 | mg.co.za | 21 | 4 | 19 |
| *GDELT-discovered URLs* | | | | time.com | 11 | 5 | 45 |
| biometricupdate.com | 3,946 | 207 | 5 | morningstaronline.co.uk | 10 | 6 | 60 |
| wired.com | 1,068 | 288 | 27 | 404media.co | 9 | 8 | 89 |
| theverge.com | 819 | 189 | 23 | militarytimes.com | 8 | 1 | 12 |
| bbc.com | 348 | 112 | 32 | rferl.org | 6 | 4 | 67 |
| abc.net.au | 233 | 93 | 40 | *Guardian Content API* | | | |
| independent.co.uk | 217 | 96 | 44 | theguardian.com | 5,337 | 352 | 7 |
| aljazeera.com | 143 | 50 | 35 | **Total** | **16,965** | **2,689** | 16 |

# D PHTKG

This appendix provides additional details of the PHTKG framework and its implementation. It describes the human-annotated reference set used for training, the extraction and verification procedure applied to these reports, the resulting enriched event representation, the formal PHTKG construction and training procedure, the inference process applied to automatically extracted events, and the Neo4j representation. We distinguish the structured event information retained in the enriched JSON output from the learned representations produced by PHTKG. The former preserves the entities, event metadata, evidence, provenance, and geolocation information used for graph construction and map generation, while the latter encodes the relational and temporal structure learned by PHTKG and supports similarity-based retrieval and relational analysis.

### D.1 EXTRACTION PROMPT AND EXTRACTION OUTPUT

The extraction model is prompted with a fixed system instruction specifying the extraction task, the eleven-field output schema, and the entity-resolution rules for each field (e.g., that the AI system field should capture what technology is involved even absent a commercial product name, and that the affected-group field should be filled with the broadest supported group rather than defaulting to "Not specified in report"). An abridged excerpt of this prompt is given below; the full prompt additionally specifies field-by-field extraction rules for the remaining fields and the JSON schema itself.

**Prompt — Extraction prompt**

```
You are the event-extraction model inside a multilingual AI-harm knowledge-graph pipeline.

Your job is to extract ONE central AI-related harm event from the supplied report.

The event does NOT need to explicitly use words such as: "harm", "bias", "discrimination",
"violation", or "victim". Neutral, legal, institutional, corporate, regulatory, and
cautious language can still describe an AI-harm event.

ABSOLUTE OUTPUT RULES
1. NEVER output JSON null for any field.
2. NEVER omit any required field.
4. If a fact is genuinely unavailable, use exactly: "Not specified in report"
   Do NOT use null. Do NOT use an empty string. Do NOT invent a fact.
6. Always return exactly ONE event.
9. Return valid JSON only.

AI SYSTEM: extract the actual technology described by the report.
Do NOT output "Not specified in report" merely because there is no commercial product name -
the field asks WHAT TECHNOLOGY is involved, not the product name.

ORGANIZATION: the actor that deployed, operated, purchased, commissioned, or used the
AI system, not the publisher of the report.

AFFECTED GROUP: the most specific supported group; do not automatically use the
organization as the affected group.

ACTION: describe what the AI-enabled system did, not the harm.

CONSEQUENCE: the adverse effect that resulted -- distinct from the action itself.
```

**Input — Raw input record, report R001**


```
{
  "report_id": "R001",
  "source_language": "en",
  "original_text": "A leaked Defense Ministry tender lays out the
      army's training program for manipulating public opinion in
      Israel and abroad. Israel's defense establishment is training
      soldiers and other defense officials to conduct psychological
      operations designed to 'influence public consciousness' in
      Israel and abroad, an internal Defense Ministry tender
      published last July and obtained by the Israeli investigative
      outlet The Hottest Place in Hell reveals. The courses, taught
      in Hebrew and English by academics who are not affiliated
      with the military, are intended for defense personnel based
      both domestically and overseas, as well as unspecified
      'foreign partners.'",
  "translated_text": null,
  "source_url": "https://www.972mag.com/leaked-idf-propaganda-
                 israel-intelligence/",
  "publication_date": "2026-06-04",
  "country": "Israel",
  "occurrence_status": "occurred",
  "taxonomy_hints": [
    "AI", "deepfake*", "Lavender", "propaganda",
    "information operation*", "manipulat*", "bombing*"
  ],
  "classifications": [
    {
```

```
      "category": "Compute / Model Behavior",
      "subcategory_id": "1.3",
      "subcategory": "Misinformation / information hazard harm",
      "event_type": "Acute"
    }
  ]
}
```


**Extraction verification** After generation, we apply three verification stages before incorporating an extracted AI harm event into the predicted-event graph.

**1. Schema verification** We first verify that the generated output is valid JSON and conforms to the required schema. We check that all required fields are present, that the output represents a single central AI harm event, and that each field follows the expected format. Outputs that fail these structural checks are not treated as valid structured events.

**2. Evidence and entailment verification.** We next verify that the extracted event is supported by the source report. The extracted original-language and translated evidence spans are checked against the corresponding source texts to ensure that they are grounded in the report. We additionally use natural language inference (NLI) to assess whether the extracted event and its associated roles are supported by the source evidence. This step is intended to identify extracted information that is not explicitly supported by the underlying report.

**3. Harm-category verification.** Finally, we verify the assigned harm category against the predefined AI harm taxonomy. The predicted category is compared with its corresponding taxonomy definition and inclusion criterion to check whether the extracted event satisfies the requirements of that category. This verification helps prevent structurally valid extractions from being assigned to an incompatible harm category.

**Human-annotated reference set.** We use 1,162 human-annotated AI harm reports as the reference set for event extraction and PHTKG training. Each report provides a human-verified event structure, including the event type, seven PHTKG roles: organization, AI system, affected group, action, harm category, consequence, and location, as well as date and evidence information.

These annotations serve both as reference labels for evaluating automatic extraction and as gold event graphs for evaluating PHTKG. For each report, we apply the same frozen extraction model used for the larger corpus to obtain a predicted structured event from the original report, its canonical English translation, source metadata, and taxonomy information. We then compare the predicted and human-annotated events field by field, separating extraction errors from errors introduced by graph representation and modeling.

**Learned representation** After training, R001 has a 128-dimensional event embedding and a temporal-encoding vector, both shown truncated below, together with a soft assignment to one of the model's learned recurring event patterns, a novelty score relative to the training distribution, and its five nearest neighbors under cosine similarity. These learned outputs are appended to the structured event fields to form the final PHTKG-enriched JSON record.

**Example output — Extraction model output, report R001**


```
{
  "event_type": "Psychological Operations Training",
  "ai_system": "Academics not affiliated with the military",
  "organization": "Defense Ministry of Israel",
  "affected_group": "Soldiers and defense officials",
  "action": "Train personnel in psychological operations",
  "consequence": "Manipulate public opinion in Israel and abroad",
  "location": "Israel and abroad",
  "event_date": "2026-06-04",
  "original_evidence_span": "A leaked Defense Ministry tender lays
      out the army's training program for manipulating public
      opinion in Israel and abroad. ... intended for defense
```

```
      personnel based both domestically and overseas, as well as
      unspecified 'foreign partners.'",
  "translated_evidence_span": "A leaked Defense Ministry tender
      lays out the army's training program for manipulating
      public opinion in Israel and abroad. ... intended for
      defense personnel based both domestically and overseas, as
      well as unspecified 'foreign partners.'",
  "confidence": 1.0,
  "harm_category": [
    {
      "category": "Compute / Model Behavior",
      "subcategory_id": "1.3",
      "subcategory": "Misinformation / information hazard harm",
      "event_type": "Acute"
    }
  ],
  "report_id": "R001",
  "source": "https://www.972mag.com/leaked-idf-propaganda-israel-intelligence/",
  "learned_representation": {
    "event_embedding": [0.101688, -0.000132, 0.035200, -0.111806,
        0.003090, 0.099490, -0.087780, 0.068742, ... ], // dim = 128
    "temporal_representation": [-2.228652, 0.985528, -0.747057,
        -0.029241, 1.089348, -0.332850, -0.045613, 0.674024, ... ],
        // dim = 128
    "similar_events": [
      {"report_id": "R931", "similarity": 0.9812},
      {"report_id": "R494", "similarity": 0.9647},
      {"report_id": "R048", "similarity": 0.9503},
      {"report_id": "R462", "similarity": 0.9388},
      {"report_id": "R467", "similarity": 0.9214}
    ],
    "ai_system_recurrence_probability": 0.524037,
    "harm_category_recurrence_probability": 0.586489
  }
}
```


### D.2 PHTKG: FORMAL CONSTRUCTION

This appendix provides the equations summarized in Section 3.3; all symbols are defined in Table A9.

**Role-aware participant representation** The same entity can appear under different semantic roles across events. PHTKG therefore combines a global entity embedding of dimension $d$ together with learned role embeddings, representing each role-indexed entity as $h(v_{e,r}) = h(v_{e,r}) + \rho_r$. Attention over the seven role-slots weights role information according to $\alpha_{e,r} = \mathrm{softmax}_r\left(\frac{q_e^\top k_{e,r}}{\sqrt{d}} + b_r\right)$, where $q_e$ is projected from the current event state, $k_{e,r}$ and the corresponding value are projections of $h(v_{e,r}, r)$, and $b_r$ is a learned role-dependent bias. The attention-weighted sum of participant values, after dropout, gives the aggregated entity message $m_e^{\text{entity}}$.

**Temporal and provenance encoding** Temporal information is encoded as $T(t) = tW_{\text{linear}} + b_{\text{linear}} + \sin(tW_{\text{periodic}} + b_{\text{periodic}})$, where $t$ is the normalized event year and $T(t)$ the resulting temporal encoding, combining a linear drift term $(W_{\text{linear}}, b_{\text{linear}})$ with a periodic term $(W_{\text{periodic}}, b_{\text{periodic}})$. Events with no known date use a learned missing-time embedding rather than an imputed date. Provenance contributes an event-level message $m_e^{\text{provenance}} = \tanh(W_p p_e + b_p)$, with $p_e \in [0,1]$ the event's provenance score and $W_p, b_p$ its projection parameters, and an outgoing gate $g(e) = \mathrm{sigmoid}(p_e)$ controlling how strongly the event propagates information to neighboring entity states. The event state $h_e^{(\ell)}$ is updated as $h_e^{(\ell+1)} = \mathrm{GRU}_{\text{event}}\left(m_e^{\text{entity}} + m_e^{\text{time}} + m_e^{\text{provenance}}, h_e^{(\ell)}\right)$, where $\mathrm{GRU}_{\text{event}}$ is the event-level GRU and $m_e^{\text{entity}}, m_e^{\text{time}}$ are the aggregated entity- and time-derived messages.

**Event-to-entity propagation** Updated events propagate to their participants through role-specific transforms, $m_{e\to v,r} = w_{e,v,r} W_r h_e^{(\ell+1)}$, $w_{e,v,r} = \kappa_r(\Delta t_{e,v})\,\mathrm{sigmoid}(p_e)$, where $m_{e\to v,r}$ is the message from event $e$ to participant $v$ in role $r$, $W_r$ a role-specific projection, and $\Delta t_{e,v}$ the time

gap between event and entity update, with $\kappa_r(\Delta t) = \exp[-\mathrm{softplus}(\gamma_r)|\Delta t|]$ a learned role-specific temporal decay ($\gamma_r$ the decay rate for role $r$). When temporal information is unavailable for either the event or the entity, the temporal factor is set to 1, so propagation is modulated by data provenance alone. Incoming messages for each entity are combined through a weight-normalized aggregation $\bar{m}_v$, after which $h_v^{(\ell+1)} = \mathrm{GRU}_{\mathrm{entity}}\left(\bar{m}_v,\ h_v^{(\ell)}\right)$, with $\mathrm{GRU}_{\mathrm{entity}}$ the entity-level GRU. Only entities participating in the current event batch are updated; non-participating entities retain their previous states, and entity states are carried forward across successive year buckets during chronological training.

Table A9: Summary of notation used throughout the paper.

| Symbol | Meaning |
|---|---|
| $a$ | A single retrieved article/document |
| $K_1$ | AI-related keyword lexicon (231 entries) |
| $K_2$ | Harm-related keyword lexicon (588 entries) |
| $\lvert K_1(a)\rvert, \lvert K_2(a)\rvert$ | Number of distinct $K_1/K_2$ entries matched in article $a$ |
| $e = (V, \tau, p)$ | One AI harm event, represented as a role-aware hyperedge |
| $V$ | Set of role-tagged participants of event $e$ |
| $\tau$ | Temporal information (year) of event $e$ |
| $p \in [0, 1]$ | Provenance / confidence score of event $e$ |
| $d$ | Shared entity/role embedding dimension ($d = 64$ default) |
| $r$ | A role slot (organization, AI system, affected group, action, harm category, consequence, location) |
| $v_{e,r}$ | Entity/participant value filling role $r$ in event $e$ |
| $h(v_{e,r})$ | Entity embedding of $v_{e,r}$ |
| $\rho_r$ | Learned role embedding for role $r$ |
| $h(v_{e,r}, r)$ | Role-value embedding, $h(v_{e,r}) + \rho_r$ |
| $\alpha(v, r)$ | Attention weight of participant $v$ in role $r$ |
| $q$ | Attention query, projected from current event state |
| $k(v, r)$ | Attention key, projection of $h(v_{e,r}, r)$ |
| $b(r)$ | Learned role-dependent attention bias |
| $T(t)$ | Temporal encoding of year $t$ |
| $W_{\text{linear}}, b_{\text{linear}}$ | Linear component of temporal encoder |
| $W_{\text{periodic}}, b_{\text{periodic}}$ | Periodic component of temporal encoder |
| $W_p, b_p$ | Parameters of the provenance-message projection |
| $m_e^{\text{provenance}}$ | Provenance-derived event message, $\tanh(W_p p_e + b_p)$ |
| $g(e)$ | Provenance outgoing gate, $\mathrm{sigmoid}(p_e)$ |
| $h_e^{(l)}$ | Event hidden state at layer $l$ |
| $m_{\text{entity}}, m_{\text{time}}, m_{\text{provenance}}$ | Aggregated messages fed into the event GRU |
| $m(e \to v, r)$ | Message from event $e$ to participant $v$ in role $r$ |
| $W_r$ | Role-specific projection matrix |
| $w(e \to v, r)$ | Propagation weight from event to participant $v$ |
| $\kappa_r(\Delta t)$ | Role-specific temporal decay function |
| $\gamma_r$ | Learned decay-rate parameter for role $r$ |
| $\Delta t$ | Time gap between an event and an update |
| $h_v^{(t)}$ | Entity $v$'s hidden state at year $t$ |
| $s(e)$ | Compatibility score of event $e$ |
| $\mathcal{L}_{\text{pred}}$ | Compatibility loss (BCE + margin-ranking) |
| $e_{\text{pos}}, e_{\text{neg}}$ | Positive (real) and negative (corrupted) hyperedges |
| $\lambda_{\text{rank}}$ | Margin-ranking loss weight (0.65) |
| margin | Margin-ranking threshold (0.25) |
| $z_{v,t}$ | Concatenated entity + temporal feature, $[h_v^t; T(t+1)]$ |
| $\mathrm{sim}(e_i, e_j)$ | Cosine similarity between event embeddings |
| $R$ | Number of roles in the triple-decomposition baseline ($R = 7$) |
| $\rho_{ij}$ | Learned relation embedding for ordered role pair $(i, j)$ (triple-decomposition baseline) |
| $s_{\text{triple}}(e)$ | Triple-decomposition baseline compatibility score |

### D.3 Training Details and Hyperparameters

Negative hyperedges are generated per positive event using three corruption strategies drawn with fixed weights: same-role substitution (55%, replacing one role value with a different value from that role's training vocabulary), a directional role swap between the organization and affected-group slots (25%), and cross-event borrowing of a role value from another training event (20%). Each attempted corruption is rejected and retried, for up to eight attempts, if it reproduces the original event or collides with a known positive hyperedge in the current split. Events whose vocabulary is too small to support a valid corruption fall back to a deterministic role-cycling procedure that respects the same collision check, keeping the negative pool leakage-safe across training, validation, and test splits.

The compatibility loss $\mathcal{L}_{\text{pred}}$ (Section 3.3) scores each positive hyperedge $e_{\text{pos}}$ against its corrupted counterpart $e_{\text{neg}}$ using the compatibility score $s(e)$, combining a binary cross-entropy (BCE) term with a margin-ranking term of weight $\lambda_{\text{rank}} = 0.65$ and margin $= 0.25$:

$$\mathcal{L}_{\text{pred}} = \underbrace{\tfrac{1}{2}\Big(\text{BCE}(s(e_{\text{pos}}), 1) + \text{BCE}(s(e_{\text{neg}}), 0)\Big)}_{\text{compatibility}} + \lambda_{\text{rank}} \cdot \max\big(0,\ \text{margin} - s(e_{\text{pos}}) + s(e_{\text{neg}})\big).$$

The auxiliary next-year recurrence objective is weighted by $\lambda_{\text{extra}} = 0.10$ relative to $\mathcal{L}_{\text{pred}}$ and linearly warmed up from 0 over the first 30 epochs, limiting its influence during the initial stage of compatibility training.

The scorer is trained with AdamW (learning rate $10^{-3}$, weight decay $3 \times 10^{-4}$), gradient norms clipped to 1.5 (global norm), and a `ReduceLROnPlateau` schedule (factor 0.5, patience 6 epochs) driven by the paired validation AUC. Training checkpoints the model whenever the paired validation AUC improves by more than $10^{-4}$ and stops after 25 epochs without improvement (maximum 120 epochs). The default configuration uses embedding dimension $d = 64$, 2 message-passing layers, dropout 0.10, 6 negatives per positive, and a chronological 15%/15% validation/test split of dated AI harm events; the same configuration is used for both the gold- and predicted-event graphs to avoid introducing a separate hyperparameter-tuning difference between the two settings. An optional self-parametric search (*Auto-PHTKG*) is used to select the final PHTKG configuration. It evaluates up to 30 candidate configurations over embedding dimension, number of layers, dropout, learning rate, weight decay, and negative-sampling mixture, training each for at most 100 epochs with a patience of 12. The configuration with the best paired validation AUC, penalized by its standard deviation across repeated corruption draws for stability, is selected. The search uses only the training and validation splits; the held-out test split is evaluated only after the final configuration has been fixed, preserving the chronological evaluation protocol.

The cross-validation protocol requires at least 20 dated AI harm events in a calendar year before that year can serve as a training, validation, or test interval, yielding three valid folds per graph. Fold AUCs were 0.980 / 0.675 / 1.000 ($n_{\text{test}} = 201/336/333$) on gold-verified events and 0.902 / 0.978 / 0.974 ($n_{\text{test}} = 195/338/304$) on predicted events. The middle gold-event fold trains on fewer years of history ($n = 292$) than its neighbors, which likely explains its lower score; the larger predicted-event graph does not show comparable fold-to-fold variance.

### D.4 Neo4j Graph Database Representation

To preserve the structured relationships identified in the extracted AI-harm events, we additionally store the resulting event corpus in a Neo4j graph database. Each harm event is represented as a central Event node linked to the entities that characterize it, including the associated Organization, AISystem, AffectedGroup, Action, HarmCategory, Consequence, and Location, together with source and event-similarity relationships when available. Event nodes retain the corresponding temporal information, evidence, confidence scores, provenance metadata, and learned representations produced by the inference pipeline. During ingestion, named entities are canonicalized to reduce duplication caused by variations in capitalization, spelling, abbreviations, or equivalent names, while the original extracted values are preserved to maintain traceability to the source data. Neo4j is therefore used as a persistent graph representation of the corpus, facilitating structured storage and relational querying, while the Dynamic AI Harm Map itself is generated directly from the enriched JSON output of the pipeline

## E Taxonomy Distribution

### E.1 Distribution of Classified AI-Harm Cases

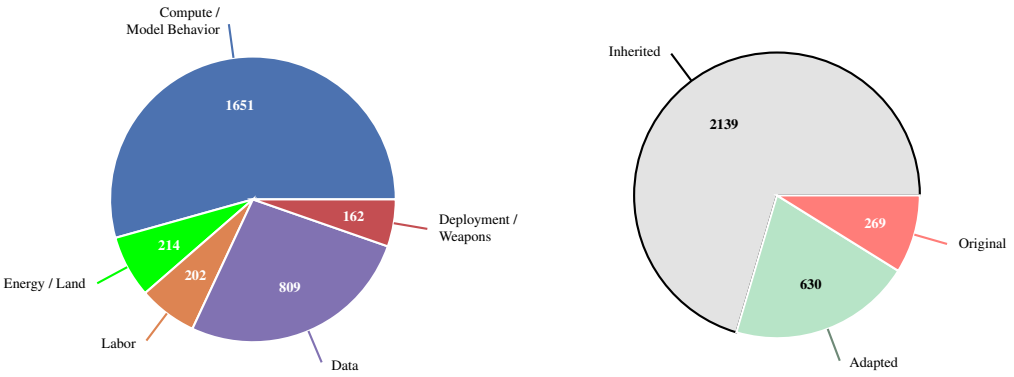
The 3,038 classified cases show how harms are distributed across the five categories and across the Inherited, Adapted, and Original categories (Figure 7a and Figure 7b).

**Counting rules** Category counts in this paper follow one of three rules, stated wherever they are used. (i) *Primary label*: each case is counted once, under the category of its primary subcategory; these counts sum to 3,038 (Compute / Model Behavior 1,651, Data 809, Energy / Land 214, Labor




Figure 6: Neo4j representation of an AI harm event. Each Event node is connected to its seven semantic role entities, source information, and, when available, similar events.

202, Deployment Context / Weapons 162) and are used in Section 4.1, Figure 4, and Figure 7a. (ii) *Category tags*: a case is counted under every category it falls in, so the 56 cases spanning two or more categories are counted more than once (1,683, 825, 214, 208, 165); this rule is used wherever a case must appear in each of its categories, namely the per-year trends (Figure 5), the geographic and country analyses (Appendix G), and the weapons analyses. (iii) *Subcategory labels*: every label is counted (3,202 in total); this rule is used only for the acute/chronic breakdown (Appendix E.2).The one exception is the overview in Appendix G.1 (Figure 11), which counts each case once and shows the 56 cases spanning two or more categories as a separate group.




(a) Category-level distribution.

(b) Inherited/Adapted/Original.

Figure 7: Distribution of classified AI-harm cases across the five taxonomy categories (left) and the Inherited, Adapted, and Original categories (right)

### E.2 Acute and chronic harm

The taxonomy assigns an event type to each subcategory rather than to each case: ten subcategories are acute, ten chronic, and two — 2.2 Land use and siting and 5.3 Accountability gap — are both. Figure 8 therefore reports how the corpus distributes across that fixed assignment. It is a property of which harms are reported, mediated by a design decision, and not an independent empirical finding: no case was judged acute or chronic on its own facts.

**Categories differ sharply in temporal character** Of 3,202 labels, 2,012 are acute, 1,097 chronic and 93 both. The split is far from uniform across categories. Compute / model behavior is 89% acute

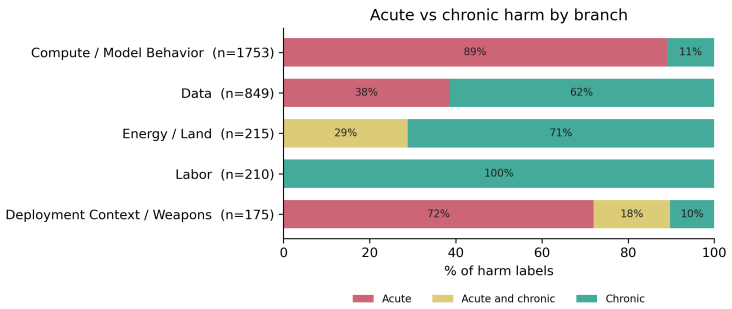



Figure 8: Share of harm labels by event type within each category; $n$ denotes labels, not cases.

(1,560 of 1,753 labels), because the subcategories it populates most heavily - 1.3 Misinformation, 1.4 Malicious use, 1.7 Physical safety - are discrete incidents. Labor is 100% chronic by construction: all three of its subcategories describe conditions rather than events. Energy / land is 71% chronic with a further 29% in the dual-typed 2.2, and data is 62% chronic, carried by 4.2 Surveillance and 4.4 Creative / IP appropriation. Deployment context and weapons is the most acute category after compute at 72%, with 18% dual-typed through 5.3 Accountability gap and 10% chronic through the newly added 5.4.

**Why this matters for the unit of analysis** An incident-based resource records what the first column of this figure describes: a dated event with a victim and a proximate cause. Roughly a third of our labels, and the overwhelming majority in three of the five categories, are of the other kind - water drawn by a facility over years, a labor market reshaped, a surveillance system left running. These have no incident date and would either be excluded from an incident register or forced into one arbitrary entry. Taking the harm rather than the incident as the unit is what allows both kinds to be recorded in the same structure, and the imbalance visible here is the practical consequence of that choice.

## F PHTKG Analysis

This appendix complements the main benchmark and component ablations with representation-level analyses of PHTKG. We examine role-aware attention, temporal encoding and decay, provenance-conditioned propagation, recurrence, role conditioning, and similarity-based retrieval, while Table A10 summarizes the structural capabilities of PHTKG and the evaluated baselines and Table A11 provides a qualitative retrieval example.

Table A10: Structural characteristics of the models evaluated in the benchmark.

| Method | Joint hyperedge | Cont. time | Prov.-gated | Recurrence |
|---|---|---|---|---|
| DeepSets-HG | yes | no | no | no |
| SetTransformer-HG | yes | no | no | no |
| Temporal-SetTransformer | yes | yes | no | yes |
| HYPER | yes | yes | no | no |
| RotatE | no | no | no | no |
| DistMult | no | no | no | no |
| ComplEx | no | no | no | no |
| **PHTKG (ours)** | **yes** | **yes** | **yes** | **yes** |

**Role-aware attention** Figure 9 shows that the final-layer attention distribution is strongly concentrated on organization, action, and consequence, which receive mean weights of 0.323, 0.311, and 0.259, respectively. Together, these three roles encode the actor involved, the AI-enabled behavior being performed, and the resulting adverse effect, and therefore describe much of the relational structure that distinguishes one AI harm event from another. Their higher attention weights suggest that the model relies strongly on this actor-action-outcome configuration when constructing the joint event representation.

Table A11: Analogical retrieval example. Top-5 nearest neighbors of query event R1148 in the learned gold-graph embedding space, by cosine similarity.

| Report ID | AI system | Year | Similarity |
|---|---|---|---|
| R916 | Universal Forensic Extraction Device | 2021 | 0.9946 |
| R816 | Facial recognition ad-targeting system | 2021 | 0.9945 |
| R1154 | (not specified) | 2021 | 0.9940 |
| R1156 | ShotSpotter | 2021 | 0.9931 |
| R914 | ShadowDragon | 2021 | 0.9930 |

The substantially lower average weights assigned to location, AI system, harm category, and affected group do not imply that these roles are irrelevant. Rather, some of their information may be redundant with the higher-weighted roles or may serve primarily as contextual information. For example, harm category summarizes a type of harm that may already be expressed through the action and consequence fields, while location may be discriminative only for a subset of geographically dependent incidents. Similarly, AI-system and affected-group descriptions vary considerably in specificity across reports. The attention distribution should therefore be interpreted as the model's allocation of representational emphasis within the learned event embedding, rather than as a direct measure of causal or stand-alone feature importance.

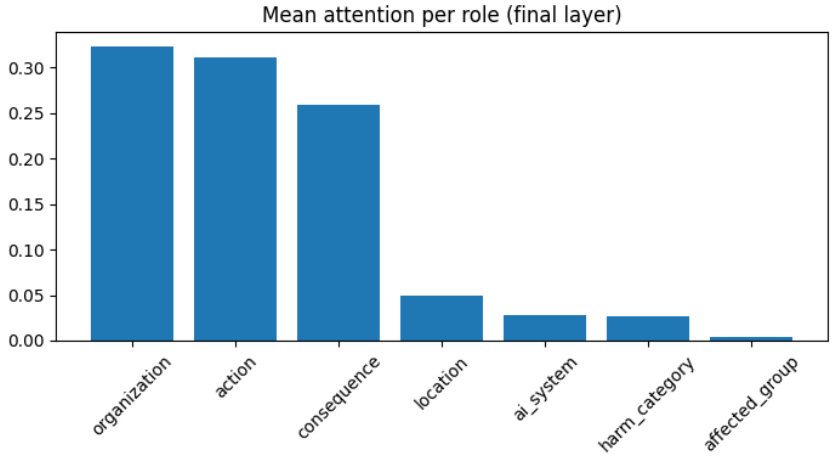



Figure 9: Mean final-layer attention assigned to the seven semantic roles in PHTKG. Organization, action, and consequence receive the largest average attention weights.

**Temporal modeling** The temporal component operates at two levels. Temporal encoding represents events according to their occurrence time, while the role-specific decay kernel controls how strongly earlier events influence later entity states.

The controlled temporal-encoding analysis shows that changing only the event year can substantially alter its representation. For the same event content, cosine similarity is 0.599 between 2018 and 2020, 0.659 between 2020 and 2024, and $-0.129$ between 2018 and 2024. Similarity therefore does not decrease monotonically with time difference. This is expected from the linear and periodic components of the temporal encoding, which distinguish temporal positions rather than enforce simple chronological distance. The low similarity between 2018 and 2024 nevertheless shows that temporal context can substantially change the event representation. Overall, the decay kernel controls how strongly older events influence later states, whereas temporal encoding allows event representations to vary non-linearly with time.

Table A12 reports next-year entity recurrence on the gold-event graph. AUC remains above chance across all four transitions, ranging from 0.666 (2022–2023) to 0.945 (2024–2025).

**Controlled ablations** Two additional analyses examine aspects of the representation not captured directly by the component visualizations. First, the complete PHTKG model is compared with a DistMult-style triple-decomposition baseline, in which the seven-role event is replaced by independently scored role pairs. The triple-decomposition representation reaches an AUC of 0.5877, com-

Table A12: Next-year entity recurrence AUC on the gold-event graph, by year transition.

| Transition | AUC |
|---|---|
| 2021–2022 | 0.762 |
| 2022–2023 | 0.666 |
| 2023–2024 | 0.768 |
| 2024–2025 | **0.945** |

pared with 0.9203 for PHTKG (an absolute difference of 0.333), consistent with the broader benchmark results in which standard pairwise knowledge-graph models also perform substantially below PHTKG. Because the controlled baseline removes the joint event state together with the temporal and provenance-aware mechanisms, this difference should be attributed to the complete structured PHTKG formulation rather than to the hyperedge representation in isolation.

Second, a controlled role-conditioning probe evaluates whether semantic role identity affects the representation independently of the underlying surface value. Representing the same surface value under two different roles produces a cosine similarity of $-0.121$, indicating that role conditioning substantially alters the resulting representation and prevents identical textual values from being treated as semantically equivalent when they occupy different positions within an event. This is complementary to the attention analysis: attention determines how strongly each participant contributes to the event, whereas role conditioning determines how that participant is represented before its contribution is aggregated.

# G THE AI HARM MAP

Section 3.4 describes how a case becomes a dot and reports what the resulting map covers. This appendix supplies what the page limit excludes: the distribution of the case base the map renders, the analyses the map makes possible, and the conventions a reader needs in order to interpret any figure drawn from it.

The map is a view of the classified case base rather than a separate artifact. Every dot is one case, carrying the labels assigned during classification, the evidence span admitted by the extraction protocol, and a link to the article it came from; the interactive map and the static figures in this appendix are drawn from the same records under the same seeded layout, so a pattern visible in one can be checked against the other. Two conventions hold throughout. First, a case may carry subcategory labels from more than one category: 56 of the 3,038 cases do, and each is counted under every category it falls in, so per-category counts sum to slightly more than 3,038, and a percentage computed over category tags is a share of tags, not of cases.

The figures below describe reporting, not incidence. A country appears often because journalism about AI harm there reaches our sources, not because more harm occurs there; the same caution applies to every organization, system and affected group counted in this appendix.

The interactive AI Harm Map is shown in the main text (Figure 2). Figure 10 gives its static per-category view: the same 2,845 placed cases under the same seeded layout, split by taxonomy category.

## G.1 DISTRIBUTION OF THE CASE BASE

Figure 11 summarizes the case base the map renders, by category and by country.

**By category** Counting each case once, Compute / model behavior accounts for 1,639 cases and Data for 772, together 79% of all cases. The three infrastructural and deployment categories are far smaller — Energy / land 213, Labor 201, and Deployment context / weapons 157, with a further 56 cases spanning two or more categories — but their size is the point rather than a defect of the corpus: these are precisely the harms that model-behavior-centered taxonomies do not name, and a case base built on such a taxonomy would record them as zero rather than as a few hundred. The

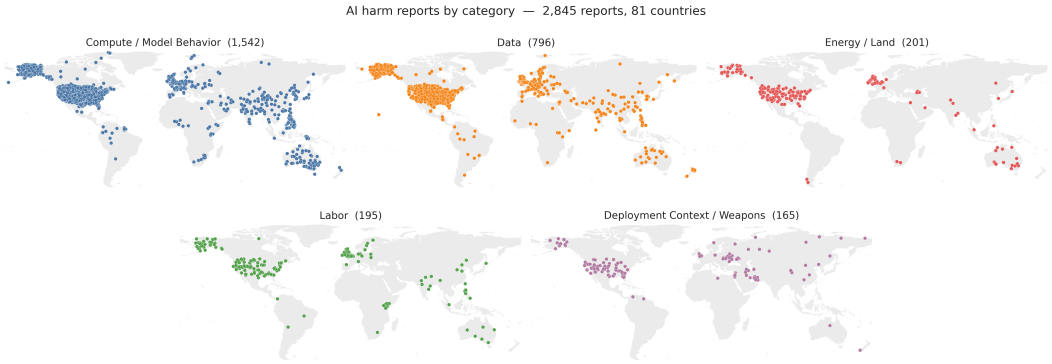



Figure 10: Static per-category view of the AI Harm Map (Figure 2): 2,845 AI harm cases by category across 81 countries; dots randomized within countries.

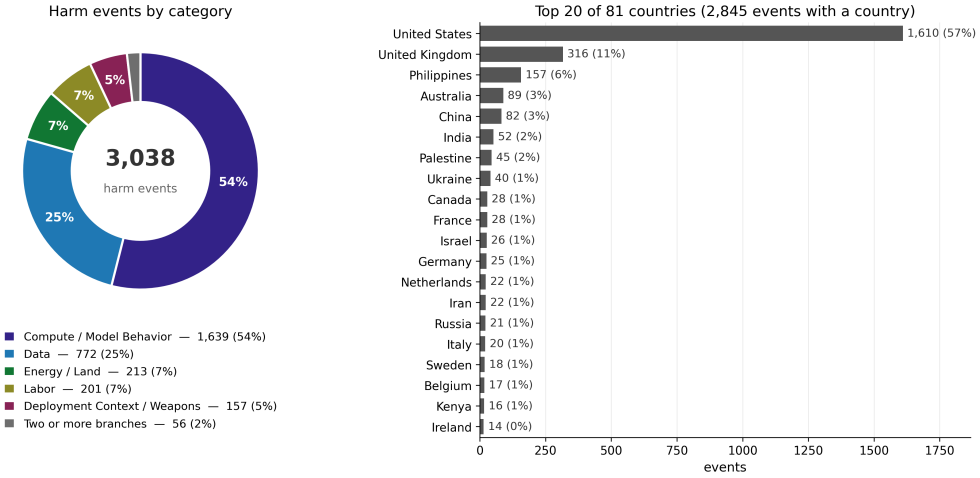



Figure 11: The mapped case base. Left: the 3,038 cases by taxonomy category, each case counted once; the 56 cases that fall under more than one category form a separate slice. Right: the twenty countries with the most cases, as counts and as a share of the 2,845 placed cases.

asymmetry also sets the scale for the rest of this appendix: a pattern involving weapons harm rests on hundreds of cases, not thousands, and is reported with that in mind.

**By country** Placement is strongly concentrated. The United States alone holds 1,610 of the 2,845 placed cases (57%), the United Kingdom 316 (11%), and the Philippines 157 (6%); the twenty countries shown account for 94% of all placed cases. The tail is correspondingly thin: the median country holds 4 cases, and 48 of the 81 countries hold five or fewer. Two of the leading positions are traceable to specific sources rather than to any property of AI deployment — the Philippines' share rests largely on Rappler's fact-checking beat, and the United Kingdom's on the volume of British national coverage in the corpus — which is why per-country shares are reported here only as a description of the corpus, and why Section 4 tests the United States concentration against locations named in the reports themselves rather than inferred.

### G.2 Geographic concentration by category

The United States holds 57% of all placed cases, and a resource built on English-language reporting invites the objection that this concentration is an artifact of how cases were located rather than a property of what was reported. Two of our four placement rules could produce it: the headquarters rule assigns a case to the country of the organization held responsible, and most of the organizations in the corpus are American. This subsection separates the two possibilities by recomputing every share over the subset of cases whose country the report itself names - *reported* and *city* placements, 2,303 of the 2,845 - and comparing it against the share over all placed cases.

Table A13 reports the geographic spread of each category, and Figure 12 shows the US share of cases within each category.

Table A13: Geographic spread of each category. *Reported* counts only cases whose country the article itself names; *all placed* adds cases located by an organization's headquarters or by a platform. A case labeled in two categories appears in both rows.

| Category | Cases | Countries | US share | | Top countries |
|---|---|---|---|---|---|
| | | | reported | all placed | |
| Energy / land | 201 | 20 | 65% | 67% | US 134, UK 23, Australia 14 |
| Data | 796 | 58 | 55% | 57% | US 452, UK 84, China 22 |
| Compute / model | 1542 | 57 | 52% | 58% | US 894, UK 168, Philippines 140 |
| Labor | 195 | 22 | 49% | 52% | US 102, UK 38, Kenya 10 |
| Weapons | 165 | 17 | 35% | 37% | US 61, Palestine 27, Ukraine 26 |
| All categories | 2845 | 81 | 53% | 57% | — |

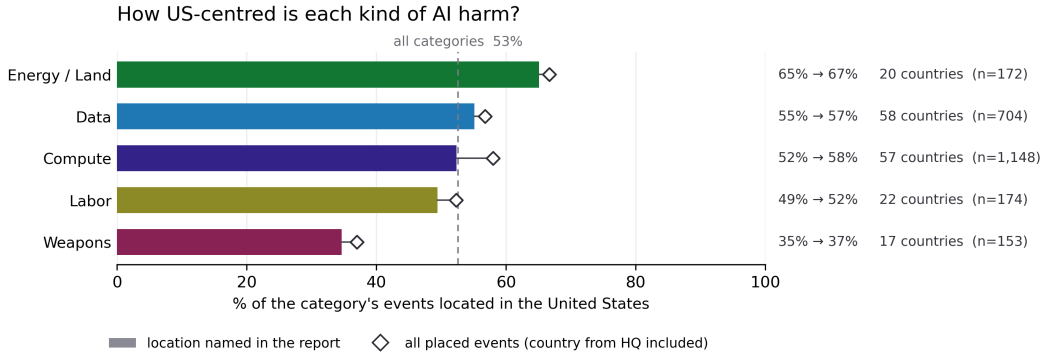



Figure 12: US share of cases by category. Bars count only cases whose location the report names; diamonds add cases located by an organization's headquarters or by a platform, and the connector between them is the shift the fallback introduces. The dashed line is the reported-only share across all categories

.

**The placement rules do not drive the concentration** The headquarters and platform fallbacks move each category by between two and six percentage points, and never change its rank: Energy / land is the most U.S.-centered category on either measure and weapons the least. The concentration survives restriction to locations the reporting itself names, so it is a property of the corpus — of which harms get written about, and where — rather than of our resolver. The largest single shift, six points in compute / model behavior, has a straightforward cause: this category carries the most cases about model behavior in general, which reporting attributes to a named developer rather than to a place, so it is the category most often located by a fallback rule (394 of its 1,542 cases).

**Categories differ in kind, not only in degree** Weapons harm is both the least American (35%) and the most narrowly distributed (17 countries), and its leading non-US positions — Palestine 27, Ukraine 26, Iran 11, Russia 10 — are active conflicts rather than large AI markets. Energy / land is its mirror image: the most American (65%) and also narrow (20 countries), reflecting where data center siting and water use are being contested and covered. The two large categories are the geographically widest, at 57 and 58 countries, because model behavior and data harms follow deployed products rather than territory. A taxonomy that collapsed these five categories into a single notion of AI harm would average these distributions together and report only the 57% figure, which describes none of them.

### G.3 Harm profiles of individual countries

Counting cases per country says where reporting concentrates; it does not say whether countries differ in the *kind* of harm reported there. Figure 13 asks that question directly, by reducing each of the fifteen most-reported countries to the distribution of its cases across the five categories and setting those distributions against the distribution over all 81 countries.

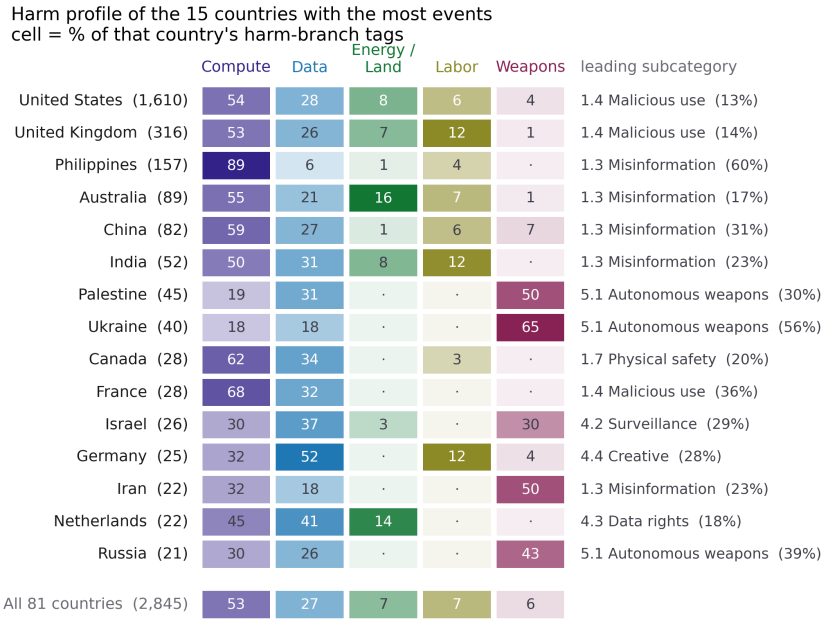



Figure 13: Harm profiles of the 15 countries with the most cases. Cells show the percentage of category tags per country (rows sum to 100%), with column-wise shading indicating concentration. The final row gives the profile across all 81 countries, and the rightmost column identifies each country's largest subcategory.

**Three profiles, not one** Most countries reproduce the global mix: roughly half compute / model behavior, a quarter data, and single digits elsewhere - the United States (54/28/8/6/4) sits almost exactly on the all-country baseline (53/27/7/7/6), which is unsurprising given that it supplies 57% of the cases that define that baseline. Two groups depart from it sharply. In Ukraine, Palestine, Iran and Russia, weapons harm accounts for 43–65% of category tags against a 6% baseline, and their leading subcategory is 5.1 Autonomous weapons in three of the four cases. In Germany and the Netherlands the departure runs the other way: data harm reaches 52% and 41%, with 4.4 Creative / IP appropriation and 4.3 Data rights leading - the profile of jurisdictions where AI harm is contested through copyright and data protection rather than through deployment. The Philippines is a third pattern again: 89% compute / model behavior, and 60% of its labels in a single subcategory, 1.3 Misinformation, the most concentrated profile in the table.

**What the profiles cannot establish** A country's profile is a property of the reporting available for it, and the concentration visible in the Philippines is partly a property of one outlet's fact-checking beat rather than of Philippine AI deployment. Sample sizes also fall quickly outside the first few rows: the fifteenth country holds 21 cases, so a shift of two or three cases moves its profile by several points, and no significance should be attached to differences of that size. The profiles are offered as a description of what the corpus contains and as a demonstration that a category-level taxonomy separates cases that a single AI-harm count would merge - not as a measurement of national differences in AI harm.

### G.4 The unplaced cases

193 of the 3,038 cases (6.4%) receive no country. They are not failures of the resolver: in almost every instance the reporting is explicit about location and the location it gives is not a place. Figure 14 shows what these cases offer instead.

**What an unplaceable location looks like** Of the 193, 106 name a platform or the internet itself (*Online platforms*, *Online*), 65 give a global scope (*Global*, *Around the world*), 4 name a multi-country region, and 1 is unspecified; the remaining 17 name a setting rather than a territory, such as remote video interviews or a category of workplace. Assigning these a country would require inventing one. The two available inventions were both rejected: routing them to the platform operator's

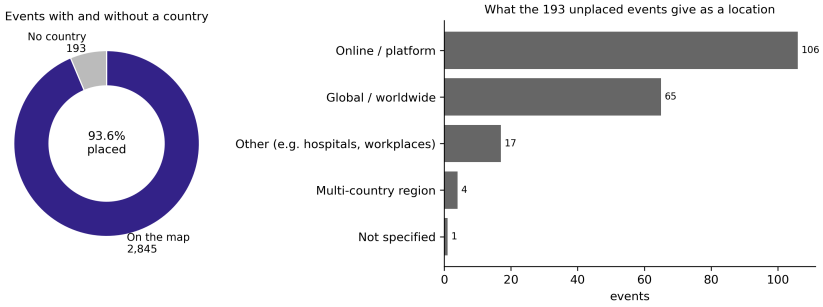



Figure 14: Left: cases with and without a country. Right: what the 193 unplaced cases give as their location, grouped by the kind of answer.

headquarters would have added roughly a hundred more cases to the United States and manufactured the very concentration Section 4 tests for, and collecting them under a *Global* pseudo-country would have created the third-largest "country" in the data and placed it nowhere on the map.

**The unplaced set is not a random sample** Its composition is lopsided: 141 of its category tags are compute / model behavior against 29 data, 13 energy / land and 13 labor, and its two most common labels are 1.4 Malicious use (50) and 1.3 Misinformation (49). This is the expected signature — a deepfake circulating on a platform or a model generating harmful output at scale has users everywhere and victims in no particular jurisdiction. Two consequences follow for every country-level figure in this appendix. First, information harms are systematically under-represented in them, so per-country compute / model-behavior shares are lower bounds. Second, no weapons case is unplaced at all: armed conflict always has a locus, which is why the geographic analysis in Section G.2 is strongest exactly where the map matters most.

### G.5 Use and supply of AI in armed conflict

The weapons category contains two populations that a map distinguishes and a list does not. Subcategories 5.1–5.3 record AI *used* in conflict: targeting systems, decision support, and harms for which no actor can be held responsible. Subcategory 5.4 records the production, export and financing that puts those systems in the field. The corpus holds 147 cases of the first kind and 18 of the second, and Figure 15 places them side by side.

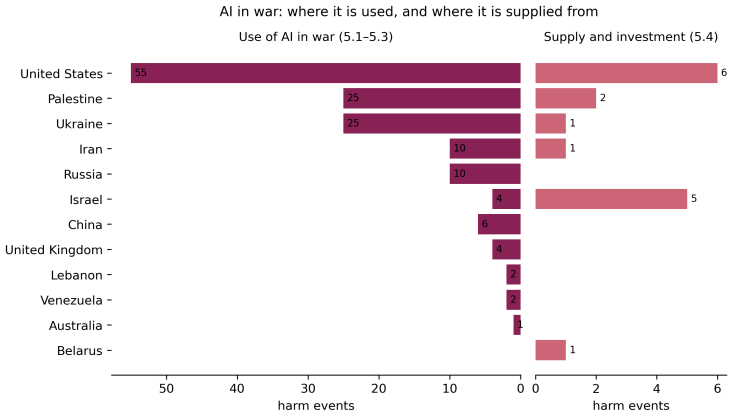



Figure 15: Countries where AI is reported in armed conflict (5.1–5.3, left) versus reported suppliers or financiers (5.4, right); note the different panel scales.

**Use and supply are geographically disjoint** Use concentrates in active conflicts: Palestine and Ukraine hold 25 cases each, Russia and Iran 10 each, across 14 countries in total. Supply concentrates elsewhere, in 8 countries, and its cases name financial and industrial actors rather than combatants — the Norwegian sovereign wealth fund's holdings in firms tied to the occupation, more

than $27 billion of European investment in Palantir, Amazon's and Google's cloud contracts with the Israeli ministry of defense under Project Nimbus, a Pentagon contract placing AI targeting software on Ukrainian drones. The United States is the only country prominent in both panels, and for different reasons in each: it holds the largest number of use cases (55), mostly concerning domestic military AI procurement and testing, and the largest number of supply cases (6). No other country appears substantially in both.

**Why 5.4 exists** These 18 cases were the direct evidence for adding the subcategory. Under the original 21-subcategory taxonomy they could not be labeled: 5.1–5.3 require that an AI system be used against someone, and a pension fund's portfolio, a cloud contract or a venture investment is not a use. Rather than force them into the nearest weapons subcategory — which would have recorded investors as conducting lethal targeting — they were left UNMAPPED so that the gap stayed countable, and 5.4 was added once the pattern was clear. Appendix B.1 records the resulting verdict.

**A lower bound, not a measurement** Eighteen cases is small, and the number is depressed by our own acceptance rule as much as by the world: a supply case is admitted only when reporting connects a named AI system to a documented adverse outcome, and most coverage of arms transfer and defense investment does neither. Ten of the eighteen come from two outlets that cover this beat closely, +972 Magazine and The Intercept, so the count also tracks where investigative capacity happens to be directed. We report it as demonstrating that the category is populated and separable, not as an estimate of how much AI supply-chain harm occurs.

### G.6 Growth of the AI Harm Map over time

Figure 16 shows six snapshots of the interactive map, each restricted to reports published up to the date shown, illustrating how coverage accumulated as the underlying 2021–2026 news corpus grew.

**Reports and countries grow together, but at different rates** Between the first and last snapshot, matched reports increase roughly 22-fold (126 to 2,845) while the number of countries represented increases roughly 5-fold (16 to 81). This gap indicates that early growth mostly deepens coverage within already-represented countries rather than reaching new ones: doubling from 126 to 267 reports (2022–2023) adds only 7 countries, whereas the map is already geographically broad all countries. The report count roughly doubles in each of the first three annual snapshots ($126 \to 267 \to 662 \to 1{,}194$), then grows more moderately from 2025 to 2026 ($1{,}194 \to 2{,}133$, a 79% increase) and again through the partial 2026 window ($2{,}133 \to 2{,}845$, a 33% increase over roughly six additional months). Snapshots cut off at the end of January and include only placed reports, so their yearly increments differ from the calendar-year totals in Figure 5.

### G.7 Cross-case patterns: organizations

Cases in the map are linked not only by taxonomy and place but by the actors they name, and it is this linkage that turns a collection of individual reports into a resource for cross-case analysis. Grouping the case base by the organization each case holds responsible answers a question no single report can: across how many countries, and across how many distinct kinds of harm, does one actor appear?

We group by lexical families over the `organization` field — a family collects an organization together with its products and subsidiaries, so that *ChatGPT*, *Sora* and *DALL-E* resolve to OpenAI, and *Instagram* and *WhatsApp* to Meta. Seventeen families cover 1,112 of the 3,038 cases; the remainder name smaller vendors, public bodies or no actor at all. Country counts here are computed over reported locations only. This is not a stylistic choice: a case placed by the headquarters rule would otherwise locate a company's harm in the company's own country by construction, and every firm would appear to cause harm mainly at home.

Figure 17 shows the organizations most frequently implicated, ranked by the number of countries in which their harms were reported.

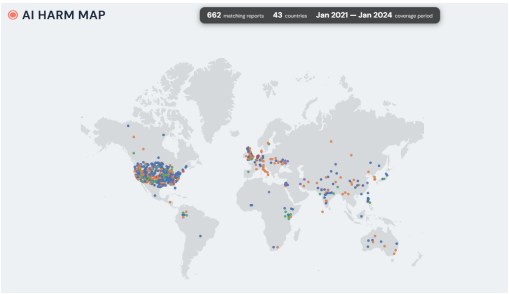
**Breadth and narrowness are both informative** The three largest general-purpose deployers appear across both the widest geography and nearly the whole taxonomy: OpenAI in 19 countries and 18 of the 22 subcategories, Meta in 18 countries and 18 subcategories, Google in 13 countries and




(a) Jan 2021–Jan 2022: 126 reports, 16 countries.



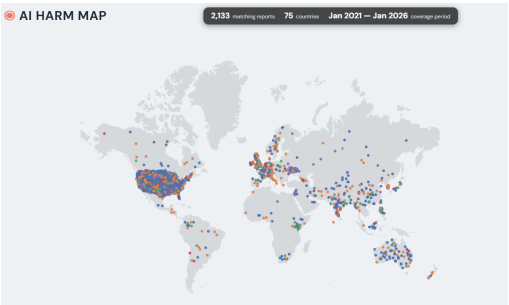

(b) Jan 2021–Jan 2023: 267 reports, 23 countries.



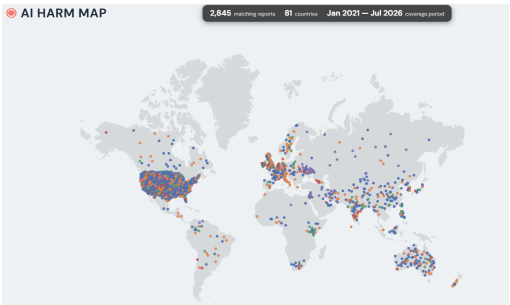

(c) Jan 2021–Jan 2024: 662 reports, 43 countries.




(d) Jan 2021–Jan 2025: 1,194 reports, 63 countries.




(e) Jan 2021–Jan 2026: 2,133 reports, 75 countries.




(f) Jan 2021–Jul 2026 (full corpus): 2,845 reports, 81 countries.

Figure 16: The AI Harm Map at 6 successive coverage cutoffs, each a cumulative snapshot of all reports published up to that date. Report count and country count both grow throughout the window; the map's spatial footprint widens from a handful of dense clusters in the United States and Western Europe (2022) to broad coverage across 81 countries (2026)

19 subcategories. Their leading harms differ despite comparable volume — OpenAI's most frequent label is 4.4 Creative / IP appropriation (46 cases), Meta's is 1.3 Misinformation (58), Google's is 1.3 followed closely by 1.7 Physical safety and 2.1 Environmental resource — which is a difference in product surface, not in scale. Narrow profiles are equally legible: Clearview AI appears in 7 countries but only 3 subcategories, all of them data harms (4.1, 4.2, 4.3); Tesla's 40 cases fall in 2 countries and are dominated by 1.7 Physical safety (32); the Israeli military's 23 cases concentrate in Palestine (19) and lead with 5.1 Autonomous weapons. A single-purpose system produces a single-purpose harm profile, and the taxonomy makes that visible without any manual reading of the cases.

**Limits of this grouping** The families are lexical and therefore conservative: a case that describes a harm without naming the responsible company is not counted, and an ambiguous token is resolved by context rather than assumed. These counts measure *reported involvement*, not culpability or market share — an organization appears here because journalism connected it to a harm, and the largest deployers attract the most scrutiny as well as deploying the most systems. They are reported to show the kind of cross-case question the map supports, and to motivate the relational representation, in

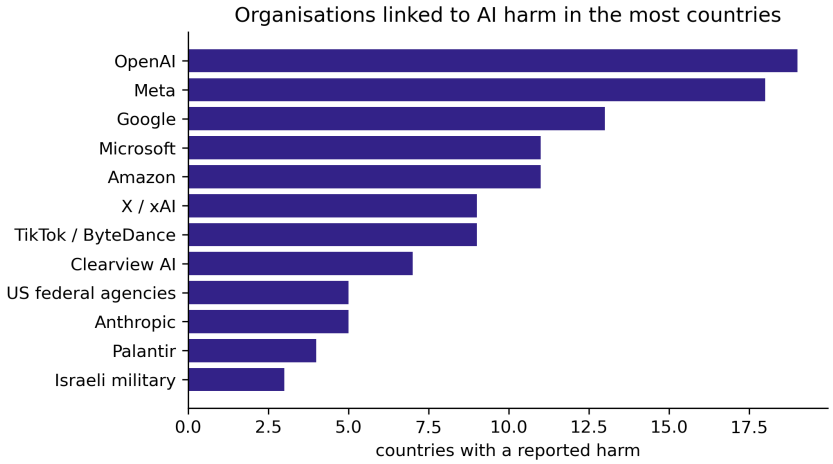



Figure 17: Organizations by the number of countries in which a harm involving them is reported, counting only cases whose country the report itself names.

which organizations are entities shared across hyperedges rather than strings repeated in separate records.

### G.8 Cross-case patterns: AI systems

The complement to grouping by actor is grouping by technology. Each case records the AI system it concerns, and collecting those strings into families — by capability rather than by vendor, so that *Lavender*, *Gospel* and armed drones fall together under military targeting — shows which systems the reporting connects to which harms. Twelve families match 1,393 of the 3,038 cases. Families are not exclusive: a case about a synthetic-voice companion app matches both deepfakes and chatbots, and is counted in each. (Figure 18)

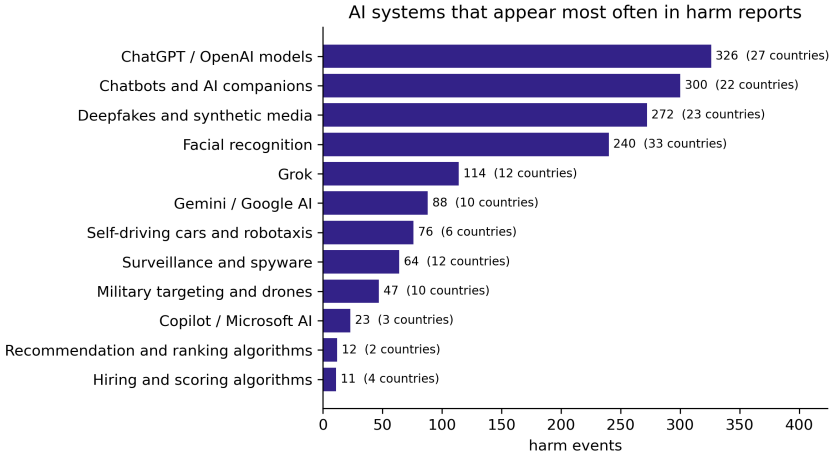



Figure 18: AI system families by number of cases, with the number of countries in which each family appears. A case naming two families is counted in both.

**Each family carries a characteristic harm signature** Generative text and conversational systems supply the largest share of cases — the ChatGPT and OpenAI-model family in 326 cases across 27 countries, chatbots and AI companions in 300 across 22, deepfakes and synthetic media in 272 across 23 — and their labels sit almost entirely inside category 1, led by 1.3 Misinformation, 1.4 Malicious use and 1.5 Human–computer interaction. The other families look nothing like this. Facial recognition is the most widely distributed system in the corpus, appearing in 33 countries, more than any other family and more than either large language model family, and 142 of its labels are 4.2 Surveillance: a municipal technology deployed in many places, each contributing a few cases. Self-driving vehicles are the opposite shape, 76 cases in 6 countries with 62 labels in 1.7 Physical safety — narrow, concentrated and bodily. Military targeting and drones is the only

family whose leading country is not the United States (Ukraine 17, then the United States 9 and Russia 8), and the only one whose labels reach 5.4 Supply chain / investor complicity.

**What the families do and do not show** Matching is lexical over the recorded system string, so a case that describes an automated decision without naming the system behind it joins no family, and the 1,393 matched cases are a lower bound on each technology's presence. Family size reflects reporting volume rather than deployment scale or severity: conversational systems are both widely used and heavily covered, while an algorithm allocating welfare payments may affect more people per case and appear only eleven times. The grouping shows that harm types often vary with system type—surveillance technologies are associated with data harms, vehicles with physical-safety harms, and generative systems with information harms—providing a regularity that the relational representation can learn across cases rather than treating each record in isolation.

### G.9 Cross-case patterns: affected groups

Each case records who the harm falls on, and grouping those strings gives the distribution the taxonomy is ultimately about: not which systems exist or which companies deploy them, but which populations appear in the reporting as harmed. Thirteen groups match 1,560 of the 3,038 cases. As elsewhere in this appendix the families overlap — a case naming schoolgirls matches both children and women — and a case is counted in each group it names. (Figure 19)

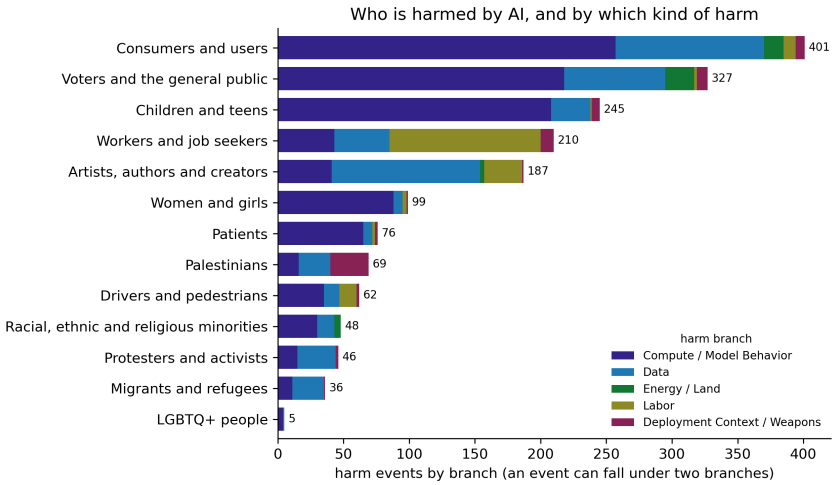



Figure 19: Affected groups by number of cases, split by taxonomy category. A case labeled in two categories contributes to both segments, so segment totals exceed the case count printed at the end of each bar.

**Who is harmed predicts which harm** The category composition of a group is rarely mixed. Workers and job seekers (208 cases) are the only group whose largest category is labor (115), led by 3.3 Employment displacement and 3.1 Data labor exploitation. Artists, authors and creators (184) are overwhelmingly a data-harm group (113), of which 88 labels are 4.4 Creative / IP appropriation. Three groups — protesters and activists (42), migrants and refugees (35) and Palestinians (56) — share a single signature: 4.2 Surveillance is the leading label for each (28, 21 and 24 labels respectively), and it is the only place in the corpus where data harm and deployment harm co-occur at scale, since the Palestinian cases carry both 4.2 Surveillance and 5.1 Autonomous weapons. The diffuse publics behave differently again: consumers and users (390) and voters and the general public (323) are dominated by compute / model-behavior harms, with 1.3 Misinformation alone accounting for 170 of the voters group's labels. Geography follows the same logic — voters and the general public appear in 36 countries, the widest of any group, while patients appear in only 5.

**Reading these counts carefully** The group field is extracted from the article text, so a group is present here when reporting names it. Size differences therefore mix three things that cannot be separated in this corpus: how often a harm occurs, how often it is reported, and how specifically an article identifies its victims. A group's absence is especially weak evidence: LGBTQ+ people

appear in 5 cases, which reflects how rarely the reporting names that population as the affected party rather than a claim about incidence. The counts are informative about the *structure* of AI harm — that labor harm falls on workers, appropriation on creators, and surveillance on politically exposed populations — and uninformative about its prevalence.

## H Future Work

Future work will expand the dataset across sources, languages, regions, time periods, and harm categories while reducing retrieval bias. This will include partnerships with civil-society and grassroots organizations to collect and verify community-submitted accounts, as well as multimodal evidence such as images and video. We will also explore LLM-based debate for ambiguous cases, allowing agents to compare evidence, provenance, and knowledge-graph context before human verification. For PHTKG, we will evaluate calibrated provenance signals, finer-grained temporal modeling, additional AI harm corpora, and incremental graph updates as new events are added to the Dynamic AI Harm Map.

## I Civil Society interest and AI governance: Extended Discussion

**Taxonomy scope corrections** Our proposed taxonomy addresses two distinct gaps through two separate extensions. First, its infrastructural definition of AI, following Crawford (2021), pulls land and labor harm into scope because CSET (Hoffmann & Frase, 2023), the MIT AI Risk Repository(Slattery et al., 2024), Weidinger et al. (2022) and Shelby et al. (2023) structurally exclude them under a narrower, computational definition. Second, the Deployment Context / Weapons category connects the AI ethics literature on harm with international humanitarian law and arms-control research. This domain is outside the scope of all four taxonomies, regardless of how they define AI. Both extensions are deliberate contributions of this work.

**Evidentiary standing of community testimony** A separate and narrower question remains open at the evidentiary layer, distinct from category scope. Once a harm type is in scope, whether a specific reported case is treated as sufficiently evidenced still depends on the project team's inclusion tests, occurrence status gate, and severity ordering, and on a source base of 27 English-language domains plus manual review. A community's own account of harm to itself has no path into this dataset unless it first passes through journalism written and indexed by one of those sources. Scholarship on epistemic exclusion in AI governance describes a version of this pattern: testimony, land claims, and lived experience of harm are treated as complaints rather than as evidence with the same standing as a formal audit, and this dynamic reproduces inside AI systems and research practice specifically, not only inside governance institutions (Mohamed et al., 2020; Quijano, 2000; de Sousa Santos, 2015). Addressing this within the current pipeline is future work, specifically extending the planned civil society sourcing workstream described above to accept and verify community-submitted accounts directly, rather than only through published reporting.

**Governance: three cross-case patterns**

1. **Disjoint use and supply** Use and supply of AI in armed conflict are geographically disjoint: harm concentrates in Palestine, Ukraine, Iran, and Russia, while the financial and industrial actors enabling it, namely investors, cloud providers, and defense contractors, are named almost entirely in the United States and Israel. A national regulator in either group of countries sees only half of this picture.
2. **Recurrent actors across borders** The same organizations recur across dozens of countries and most of the taxonomy: OpenAI appears in 19 countries and 18 of 22 subcategories, and Meta in 18 countries and 18 subcategories. No single national enforcement action, however well resourced, can see the pattern a company's conduct forms across borders; only a resource built to connect cases can.
3. **Harm without jurisdiction** 193 cases (6.4% of the corpus) resist jurisdictional placement entirely, because the harm is genuinely global or platform-based rather than local. These are not edge cases to be resolved by better geolocation; they show that a meaningful share of

> AI harm has no jurisdiction by construction, and national frameworks premised on locating harm within a territory will miss them regardless of enforcement quality.

Together, these findings support coordinated, cross-border monitoring and accountability mechanisms that extend liability upstream to financiers and suppliers, not only operators, and that track actors and patterns across jurisdictions rather than within them.

**Accountability findings (documenting is not resolving)** Subcategories 5.3 and 5.4 record 41 cases in which no actor could be held responsible or in which financing and supply relationships enabled harm elsewhere. Our paper stops at naming that gap; what redress would look like for the people in these cases remains an open question it does not answer. The 5.4 count is a lower bound in the strict statistical sense, but it should not be read only as a taxonomy gap that got filled. Eighteen cases naming specific investors, cloud contracts, and financing relationships identify the kind of actor that divestment campaigns, shareholder resolutions, and procurement challenges are already built to target, and the map makes that actor identifiable at the case level rather than only in aggregate.