Designing *Against* Harm: Translating Real-World AI Failures into Actionable Design Guidelines

HUMAN CENTERED DESIGN & ENGINEERING
UNIVERSITY of WASHINGTON

Pitch Sinlapanuntakul | **Auli** Badoni* | **Avleen** Kaur* | **Chang** Li* | **Hsiang**-Yi Lung* | **Otto** Roesler* | Te-Hsin **Angela** Shen* | **Tirata** Vittayaareekul* | **Mark** Zachry

* Authors contributed equally to this research

A. CONTEXT

AI-infused products can generate outputs that mislead, misinform, violate rights, or lead to unsafe real-world consequences. These harms are not only caused by technical failures, but also arise from design decisions that fail to account for the risks of human–AI interaction.

Existing AI design/ethics resources often ask designers to design *for* positive abstract values ("be fair," "be transparent"), which leave designers without a concrete, failure-grounded vocabulary for preventing harm.

Our work derives **6 design principles and 19 patterns from 155 documented real-world AI harm incidents**. By translating past failures into a structured, practice-oriented resource, we support designers in proactively designing *against* harm during early-stage AI concept development.

B. METHODOLOGY & ANALYSIS

We draw on **155 cases from the AIAAIC (AI, Algorithmic, and Automation Incidents and Controversies) archive**, an independent, non-partisan public interest initiative that documents real-world AI incidents through news reporting, investigative journalism, and public disclosures.

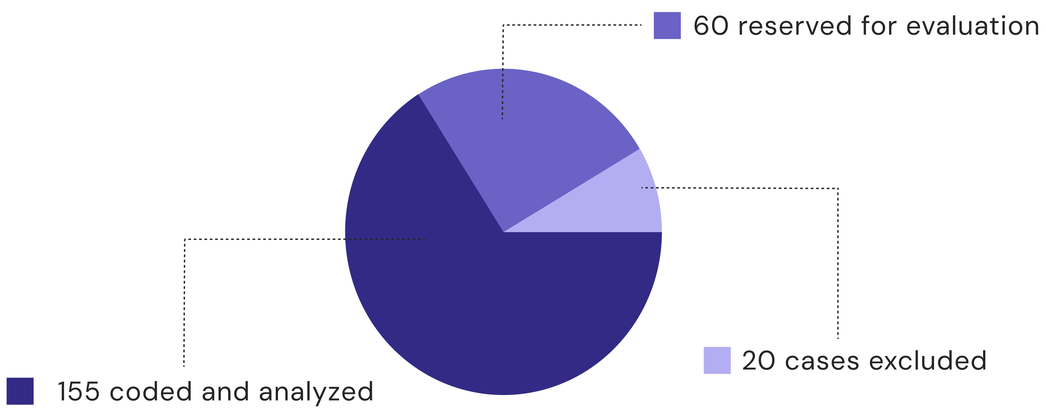

Fig 1. 235 AI harm cases from AIAAIC (2025).

Unlike research reports, AIAAIC captures incident-level reporting without pre-applied interpretive framing, allowing our analysis to remain grounded in the raw characteristics of real failures across diverse AI applications, domains, harm types, and socio-technical contexts.

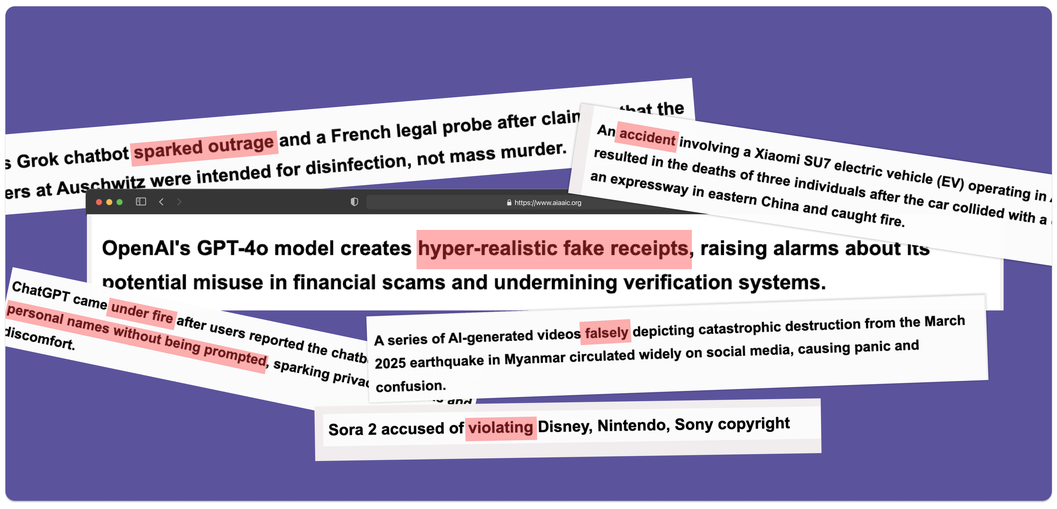

Fig 2. Examples of AI harm cases covered.

1 **Calibration**
15 shared cases

2 **Coding**
140 cases coded

3 **Affinity Clustering**
24 clusters → 20 patterns

4 **Restructuring**
26 patterns → 15 principles

5 **Probe + Refine**
11 design ideas → 3 major revisions

6 **Guidelines (v.1)**
6 design principles → 19 patterns

C. DESIGN GUIDELINES (version 1)

MOST EXISTING RESOURCES	OUR GUIDELINES
Starting point: **literature & abstract values**	Starting point: **documented failures**
*"How do we **design for good**?"*	*"How do we **design against harm**?"*

VS

PRINCIPLE 1

Design Against Unbounded AI Authority

Establish what the AI is permitted to do, and build in human control for when those permissions are exceeded.

37 cases

PRINCIPLE 2

Design Against Blind Trust

Make the AI's capability boundaries and output reliability visible so users can judge when and how much to rely on what they receive.

34 cases

PRINCIPLE 3

Design Against Misattributed Content

Ensure AI-produced outputs remain clearly identifiable as machine-generated, so it cannot be mistaken for human work.

41 cases

PRINCIPLE 4

Design Against Impulsive Harm

Introduce deliberate pauses at moments where fast decisions or fast automated action could produce harm that is difficult to reverse.

14 cases

PRINCIPLE 5

Design Against Hidden Influence

Make visible both what data the product collects or draws on and how it shapes the output, so users can see and adjust the forces acting on what they receive.

21 cases

PRINCIPLE 6

Design Against Safety Drift

Keep safety signals and correction mechanisms active and accessible across the full duration of an interaction, not only at its start.

14 cases

D. FUTURE WORK