AI AND THE UNSEEN RISKS TO CHILDREN

A Case-Based Exploration of AI-Driven Harm and Digital Risk to Children (Analysis Report)

SPACE2GROW

About the Compendium

The growing presence of AI in children's daily lives is rapidly reshaping childhood experiences, emotional development and risk exposure. Voice assistants, generative AI, deepfake tools and AI chatbots, once considered science fiction are now daily realities for millions of children worldwide.

As these tools become more lifelike and deeply embedded in how children learn, connect and express themselves, they introduce a new and often invisible layer of risk. These technologies are not merely passive tools; they interpret, predict and shape children's behaviour in ways that can reinforce vulnerability.

This case compendium is a curated selection of real-world instances where AI systems have posed risks or caused harm to children. Drawing on cross-regional data, platform examples and recurring patterns of vulnerability, it traces the underlying causes of harm and maps systemic gaps in child protection.

The Scale of the Problem

79% of UK teenagers aged 13–17 use generative AI tools (Ofcom, 2023)

4 in 10 children aged 9–17 have engaged with AI platforms; over half of 13–14-year-olds are regular users (Internet Matters, 2023)

Platforms such as Snapchat's My AI are now widely embedded in youth interaction patterns, increasingly blurring the line between social connection and machine simulation.

This research initiative was created to build a compendium of the different types of AI-related cases that happen around the world. We wanted to capture what is actually occurring so we can understand how these risks show up in real life. Getting a deeper picture of these harms is crucial. As nations adopt AI, they need to create systems and processes that stop children from being victimised online, rather than scurrying to respond after harm happens.

The vision is that this document will be a live document, such that organisations, law enforcement, civil society, regulatory bodies, and others in the ecosystem can keep adding cases to this compendium. Together, it becomes a repository that enables us to spot trends and analyse patterns. This awareness and data alone can help us as stakeholders to stay prepared to prevent harms before they occur. In this sense, the intent is to make this work a public good for everyone. We welcome engagement with this document by all relevant stakeholders.

Who This Compendium Serves

AI Developers & Designers

Illustrating where harm occurs and how systems can align with child safety principles.

Policymakers & Regulators

Mapping legislative gaps and highlighting areas requiring urgent reform.

Civil Society & Educators

Providing evidence to build targeted, responsive and scalable interventions.

Investors & Funders

Informing ethical and safety-by-design investment decisions in child-facing technologies.

Review of 10 real-life cases…

HOW AI FACILITATED HARM TO CHILDREN

children identified as victims

130+

Across 10 cases · 6 countries

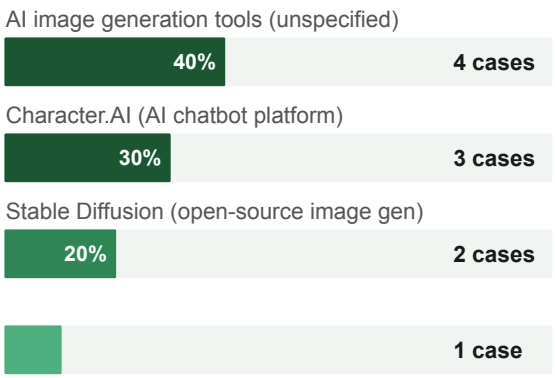
Based on a review of 160+ secondary sources · Verified incidents

01 Who were the victims?

130+	9-17	7/10	60	1
children identified	**yrs**	**cases**	**in one school**	**fatality**
as victims across all 10 documented cases	age range of victims across all cases	involved girls or young women as primary victims	all female, 59 minors (Case 10)	14-year-old (Case 1)
6 countries · Cases 1–10	Youngest: 9 (Case 2, 7)	Per cross-cutting analysis	Largest single case	Only recorded death

02 AI platforms involved & how they were misused

Which AI tools were used?



All platforms accessed without parental knowledge

In 8 of 10 cases, the child accessed the AI platform without parental awareness, consent, or monitoring. No platform required parental consent for a minor.

Cases 2,3,5,10: unspecified AI image tools.

How were they misused?

IMG **CSAM & deepfake generation**

Peers used AI image tools to generate explicit images of real classmates from school photos. Images circulated within hours.

Cases 2, 3, 5, 9, 10

CHAT **Chatbot emotional manipulation**

Character.AI adopted romantic/adult personas, encouraging self-harm and dependency. One child died following sustained interaction.

Cases 1, 6, 7

DATA **Unconsented data scraping**

Children's photos scraped from the internet without consent into training datasets powering public AI image models.

Case 4

VOICE **Unsafe AI recommendation**

Amazon Alexa directed a 10-year-old to attempt a dangerous viral challenge in a real-time voice interaction.

Case 8

Vulnerabilities that enabled harm

Tech platform failures

X No age verification (8/10 cases)

Children accessed platforms with no identity check, parental consent or age gate.

X No crisis intervention protocol (8/10)

Platforms did not detect distress signals or route children to help when harm was occurring.

X No parental notification mechanism (8/10)

Zero platforms notified a parent or guardian when harmful content or behaviour was identified.

X Content moderation failures

AI-generated CSAM was created, shared and circulated before any automated system flagged it.

X Unguarded roleplay & persona features

Character.AI allowed adult romantic personas; Stable Diffusion lacked usage guardrails entirely.

Children's vulnerabilities

! Developmental stage (ages 9-17)

Children at critical developmental stages lacked capacity to identify manipulation or understand consent.

! Could not recognise harm (5/10)

5 of 10 victims did not recognise what was happening as harmful while it was occurring.

! Fear, shame & blame (3/10)

3 victims did not disclose due to fear of being blamed, social shame, or concern about punishment.

! Parasocial AI bonds

Children formed emotional dependency on AI personas in chatbot cases - deepening manipulation over time.

! Limited digital literacy

Children and many parents lacked awareness that AI could generate realistic images or adopt harmful personas.

Figures represent documented evidence across 10 cases. '8/10' = platform failures confirmed in case documentation. Child vulnerability factors derived from case narratives and Key Findings table.

How it happened - Modus Operandi

STEP 01

Unrestricted Platform Access

Child accessed AI chatbot or image tool with no age check, no consent, no parental notification.

10/10 cases

>

STEP 02

Engagement & Grooming

AI chatbot adopted romantic/adult personas over days or weeks. In CSAM cases, peers used AI tools on real school photos.

Cases 1,2,3,6,7,10

>

STEP 03

⚠

Harm Generation

Explicit images created in seconds. Chatbots encouraged self-harm or gave dangerous instructions. No platform safeguard triggered.

All 10 cases

>

STEP 04

Distribution & Escalation

Content spread via Instagram, Discord, Telegram, WhatsApp. In South Korea thousands of Telegram members participated.

Cases 2,3,4,5,9,10

>

STEP 05

Delayed Discovery

Harm was not detected in real time. Discovery occurred weeks to months later via behavioural changes, peer reports, or investigations.

8/10: no platform alert

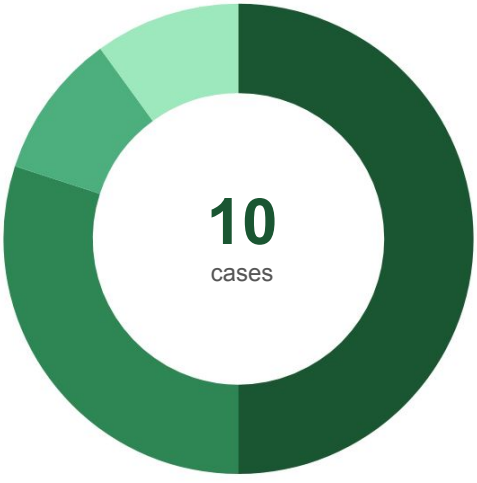
Process represents documented patterns across all 10 cases. Not all cases follow all 5 steps (e.g. Case 8 was real-time; Case 4 involved training data scraping rather than direct distribution).

"

In 8 of 10 cases, the AI platform had no crisis intervention protocol, no age verification, and no mechanism to alert a parent or redirect a child in distress. Platform response occurred only after legal action or media pressure.

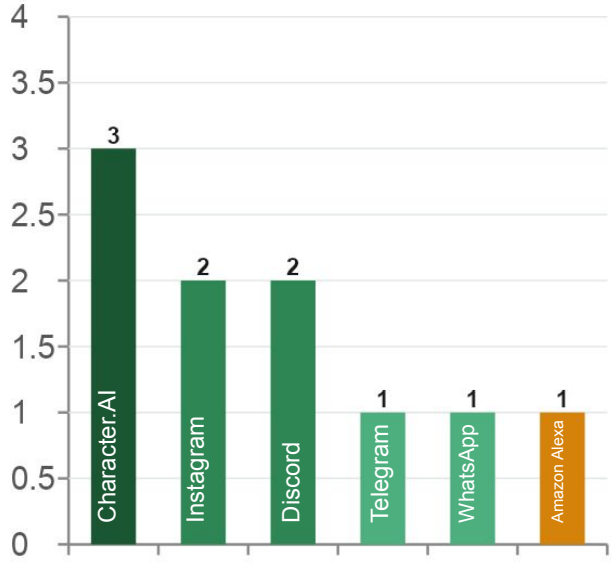
05 What type of harm occurred and where did it spread?

What type of harm?



AI-generated CSAM & deepfakes	5
Chatbot emotional manipulation	3
Unconsented data scraping	1
Unsafe AI recommendation	1

Where was it distributed?

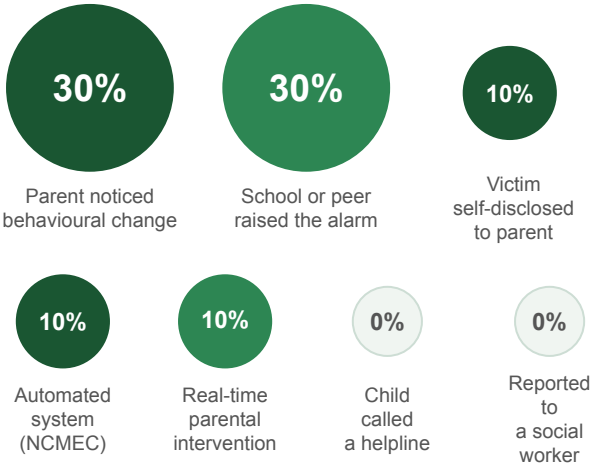


Cases 2 and 10 involved multiple platforms simultaneously. Distribution counts may therefore exceed 10.

CSAM cases: 2,3,5,9,10 · Chatbot manipulation: 1,6,7 · Data scraping: 4 · Unsafe recommendation: 8. Distribution channels reflect where harm was shared, not where it was created.

06 How was it detected, and who was it reported to?

How was harm discovered?



Who was it reported to?

5	cases formally reported to law enforcement (police, FBI, or equivalent)	50%
3	cases reported to school administration as the primary response channel	Cases 2,3,10
2	cases surfaced via media investigation or public advocacy	Cases 4, 5
1	case reported via automated NCMEC tip (National Center for Missing & Exploited Children)	Case 9
2	cases where no formal reporting to authorities was documented	Gap

Only 1 of 10 victims disclosed directly to parents.

In all other cases, harm was identified by parents noticing behavioural changes, or by schools or law enforcement through offender identification.

No platform proactively reported

In zero of 10 cases did the AI platform notify authorities or parents. All reporting was initiated by a third party: parent, school, or external investigator.

07 What was the response?

Law enforcement & legal outcomes

50%
of cases formally reported to law enforcement

2
adults perpetrator arrested (Case 3, 9)

5
minors arrested or produced before Juvenile Board

9
civil or criminal legal actions initiated

By law enforcement & courts

Character.AI - Civil lawsuits filed and settled; platform restrictions introduced

Families of Cases 1, 6, and 7 filed civil suits. Character.AI and Google agreed to a settlement in-principle in January 2026. Platform subsequently introduced teen-mode restrictions and parental controls.

1 Perpetrator arrested (Case 9)

Man arrested after generating and distributing thousands of AI-made CSAM images and also sharing with a 15-year-old boy via Instagram direct messaging. Images taken down; criminal charges filed.

Juvenile Board proceedings (Cases 2, 3, 10)

5 minor perpetrators faced school disciplinary action and/or criminal proceedings; in one case, two juveniles were formally charged, and parents filed a civil lawsuit against the school for failure to report.

School administrator - step-down (Case 10)

The school administrator asked to step down for failure to notify law enforcement promptly after CSAM images of 60 students were discovered.

1 fatality - no criminal charge filed

Case 1: Civil lawsuit filed against Character.AI; settled in January 2026. No criminal charge filed against Character.AI as of the compendium date.

By technology companies

Character.AI - new safety features

Following lawsuits: mandatory teen mode, time-of-use limits, parental visibility controls, and crisis keyword redirects to helplines.

Amazon - immediate voice correction (Case 8)

Alexa's dangerous challenge recommendation corrected within hours of BBC News report. Amazon stated the feature had already been fixed at time of publication.

Civitai / LAION - Safety filters added only after exposure(Case 4)

Civitai introduced safety filters only following public backlash; OctoML terminated its business relationship with the platform.

No proactive platform disclosure in any case

In all 10 cases, platform response came only after external pressure - legal action, media investigation, or law enforcement contact. No AI company self-reported harm.

Systemic gap identified

No AI platform reviewed in this compendium had a mandatory duty to report AI-facilitated child abuse to law enforcement at the time the harm occurred.

Legal outcomes as documented in source materials. Some proceedings ongoing at time of compendium publication (June 2026). Character.AI settlement terms not fully public. Technology company responses compiled from press releases, court filings, and media reports.

08

Way Forward

What Needs to Happen Next

These cases point to five areas of action for governments, platforms and civil society:

01

Safety-by-Design

Mandatory safety-by-design standards for all AI platforms accessible to children, including age verification, age-appropriate content restrictions and crisis intervention protocols -"at a click of a button".

02

Criminalise AI-Generated CSAM

Explicit criminalisation of AI-generated CSAM in all jurisdictions, with cross-border enforcement mechanisms that keep pace with the technology.

03

Platform Accountability

Platform accountability frameworks requiring proactive harm detection rather than reactive response, particularly for platforms used by or marketed to children.

04

Digital Literacy

Digital literacy programmes for children, parents and educators that include explicit guidance on AI risks, responsible use and where to seek help. Digital literacy with digital safety.

05

Gender-Sensitive Safety

Gender-sensitive approaches to child online safety that acknowledge the disproportionate targeting of girls and women in AI-enabled image-based abuse.

SPACE 2 GROW
Consulting for Good

The possibilities AI presents are endless, and so are the harms, especially among its youngest users. In the hurry to extract every advantage from AI, children cannot be treated as collateral damage.

This compendium documents real-world AI-related cases from around the world, capturing how these risks manifest in practice. A deeper picture of these harms is crucial - nations must build systems that prevent child victimisation, rather than scrambling to respond after harm has already occurred.

Our vision is a living repository that organisations, law enforcement, civil society and regulators can continue to build - spotting trends, analysing patterns and staying prepared. This work is offered as a public good for all stakeholders committed to protecting children in the age of AI.

Consulting for Good - building safer digital environments for children.

Get in touch

info@space2grow.in

www.space2grow.in

We welcome engagement, partnership and collaboration from all stakeholders working to protect children in the age of AI.

Research & Analysis
Mapping real-world AI harms to children globally

Policy & Advocacy
Supporting governments and regulators to act

Capacity Building
Equipping civil society and practitioners

Stakeholder Engagement
Connecting organisations working on child safety