AID

English

Incident 807: ChatGPT Introduces Errors in Critical Child Protection Court Report

Description: A child protection worker in Victoria used ChatGPT to draft a report submitted to the Children's Court. The AI-generated report contained inaccuracies and downplayed risks to the child, resulting in a privacy breach when sensitive information was shared with OpenAI.

Editor Notes: Reconstructing the timeline of events: Between July and December 2023, according to reporting, nearly 900 employees of Victoria's Department of Families, Fairness, and Housing (DFFH), representing 13% of the workforce, accessed ChatGPT. In early 2024, a case worker used ChatGPT to draft a child protection report submitted to the Children's Court. This report contained significant inaccuracies, including the misrepresentation of personal details and a downplaying of risks to the child, whose parents had been charged with sexual offenses. Following this incident, an internal review of the case worker's unit revealed that over 100 other cases showed signs of potential AI involvement in drafting child protection documents. On September 24, 2024, the department was instructed to ban the use of public generative AI tools and to notify staff accordingly, but the Office of the Victorian Information Commissioner (OVIC) found this directive had not been fully implemented. The next day, on September 25, 2024, OVIC released its investigation findings, confirming the inaccuracies in the ChatGPT-generated report and outlined the risks associated with AI use in child protection cases. OVIC issued a compliance notice requiring DFFH to block access to generative AI tools by November 5, 2024.

Tools

+ New Report + New Response Discover Citation Info View History

Entities

View all entities

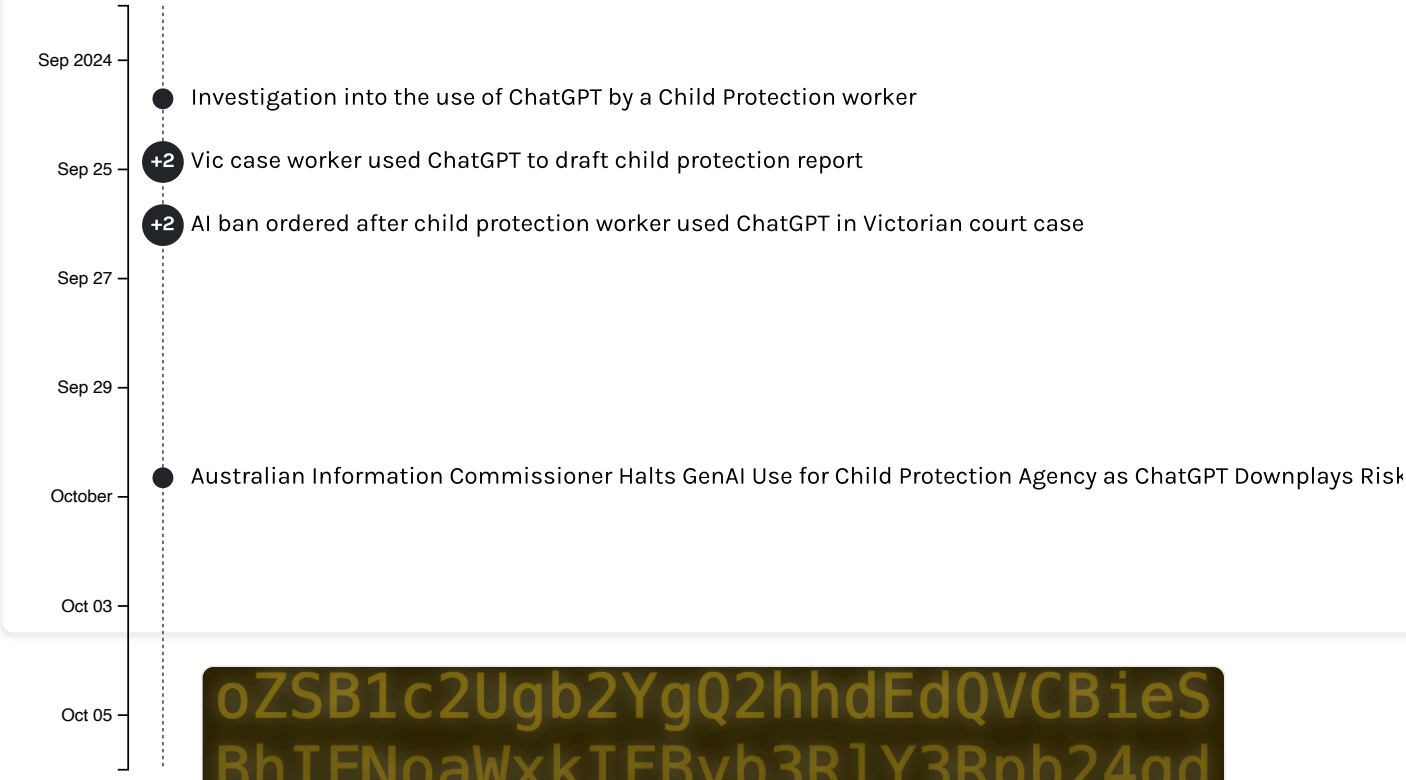
Alleged: OpenAI developed an AI system deployed by Department of Families Fairness and Housing , Government of Victoria and Employee of Department of Families Fairness and Housing , which harmed Unnamed child and Unnamed family of child .

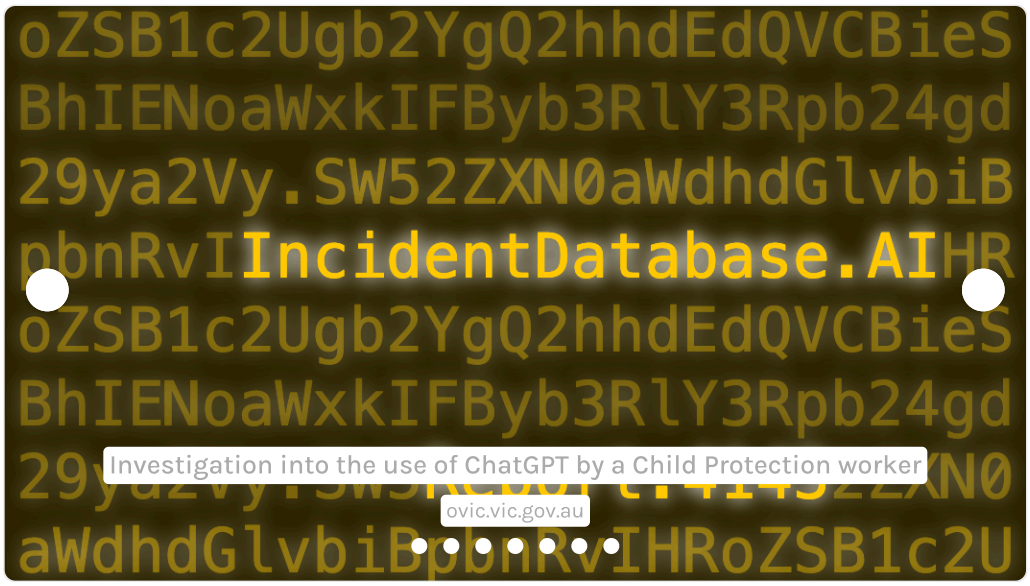
Incident Stats

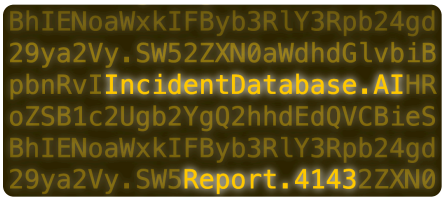
Incident ID	807
Report Count	7
Incident Date	2024-09-25
Editors	Daniel Atherton

Incident Reports

Reports Timeline


Sep 2024
Investigation into the use of ChatGPT by a Child Protection worker
Sep 25
+2
Vic case worker used ChatGPT to draft child protection report
+2
AI ban ordered after child protection worker used ChatGPT in Victorian court case
Sep 27
Sep 29
October
Australian Information Commissioner Halts GenAI Use for Child Protection Agency as ChatGPT Downplays Risk
Oct 03
Oct 05


oZSB1c2Ugb2YgQ2hhdEdQVCBieS
BhIENoaWxkIFByb3RlY3Rpb24gd
29ya2Vy.SW52ZXN0aWdhdGlvbiB
pbnRvIIncidentDatabase.AIHR
oZSB1c2Ugb2YgQ2hhdEdQVCBieS
BhIENoaWxkIFByb3RlY3Rpb24gd
Investigation into the use of ChatGPT by a Child Protection worker
ovic.vic.gov.au
aWdhdGlvbiBpbnRvIHRoZSB1c2U



Investigation into the use of ChatGPT by a Child Protection worker

ovic.vic.gov.au · 2024

The following is a copy of the executive summary of the report. To view the report in full, please download the PDF provided by OVIC.

Executive summary

Background

In December 2023, the Department of Families, Fairness and Housing (DFFH) reported a privacy incident to the Office of the Information Commissioner (OVIC), explaining that a Child Protection worker (CPW1) had used ChatGPT2 when drafting a Protection Application Report (PA Report). The report had been submitted to the Children's Court for a case concerning a young child whose parents had been charged in relation to sexual offences.

PA reports are essential in protecting vulnerable children who require court ordered protective intervention to ensure their safety, needs and rights. These reports contain Child Protection workers' assessment of the risks and needs of the child, and of the parents' capacity to provide for the child's safety and development.

Despite its popularity, there are a range of privacy risks associated with the use of generative artificial intelligence (GenAI) tools such as ChatGPT. Most relevant in the present circumstances are risks related to inaccurate personal information and unauthorised disclosure of personal information.

After conducting preliminary inquiries with DFFH, the Privacy and Data Protection Deputy Commissioner commenced an investigation under section 8C(2)(e) of the Privacy and Data Protection (PDP) Act with a view to deciding whether to issue a compliance notice to DFFH under section 78 of that Act. OVIC may issue a compliance notice where it determines that:

a. an organisation has contravened one or more of the Information Privacy Principles (IPPs);

b. the contravention is serious, repeated or flagrant; and

c. the organisation should be required to take specified actions within a specified timeframe to ensure compliance with the IPPs.

Findings

OVIC's investigation confirmed DFFH's initial findings – that CPW1 used ChatGPT in drafting the PA Report and input personal information in doing so.

There were a range of indicators of ChatGPT usage throughout the report, relating to both the analysis and the language used in the report. These included use of language not commensurate with employee training and Child Protection guidelines, as well as inappropriate sentence structure.

More significantly, parts of the report included personal information that was not accurate. Of particular concern, the report described a child's doll – which was reported to Child Protection as having been used by the child's father for sexual purposes – as a notable strength of the parents' efforts to support the child's development needs with "age-appropriate toys".

The use of ChatGPT therefore had the effect of downplaying the severity of the actual or potential harm to the child, with the potential to impact decisions about the child's care. Fortunately, the deficiencies in the report did not ultimately change the decision making of either Child Protection or the Court in relation to the child.

By entering personal and sensitive information about the mother, father, carer, and child into ChatGPT, CPW1 also disclosed this information to OpenAI (the company which operates ChatGPT). This unauthorised disclosure released the information from the control of DFFH with OpenAI being able to determine any further uses or disclosures of it.

While the focus of the investigation was on the PA Report incident, OVIC also considered other potential uses of ChatGPT by CPW1 and their broader team, as well as examining the general usage of ChatGPT across DFFH. This revealed that:

• A DFFH internal review into all child protection cases handled by CPW1's broader work unit over a one year period, identified 100 cases with indicators that ChatGPT may have been used to draft child protection related documents.

• Within the period of July to December 2023, nearly 900 employees across DFFH had accessed the ChatGPT website, representing almost 13 per cent of its workforce.

Contravention of the IPPs

While the PA Report incident may have involved the contravention of multiple IPPs, OVIC's investigation specifically considered DFFH's management of the risks associated with the use of ChatGPT through the lens of two IPPs:

• IPP 3.1 – which requires organisations to take reasonable steps to make sure that the personal information it collects, uses or discloses is accurate, complete and up to date.

• IPP 4.1 – which requires organisations to take reasonable steps to protect the personal information it holds from misuse and loss and from unauthorised access, modification or disclosure.

DFFH submitted to OVIC's investigation that it had a range of controls in place at the time of the PA Report incident in the form of existing policies, procedures, and training materials (such as its Acceptable Use of Technology Policy and eLearning modules on privacy, security and human rights).

However, OVIC found that these controls were far from sufficient to mitigate the privacy risks associated with the use of ChatGPT in child protection matters. It could not be expected that staff would gain an understanding of how to appropriately use novel GenAI tools like ChatGPT from these general guidance materials.

There was no evidence that, by the time of the PA Report incident, DFFH had made any other attempts to educate or train staff about how GenAI tools work, and the privacy risks associated with them. Additionally, there were no departmental rules in place about when and how these tools should or should not be used. Nor were there any technical controls to restrict access to tools like ChatGPT.

Essentially, DFFH had no controls targeted at addressing specific privacy risks associated with ChatGPT and GenAI tools more generally. The Deputy Commissioner therefore found that DFFH contravened both IPP 3.1 and IPP 4.1 and determined that the contraventions were "serious" for the purposes of section 78(1)(b)(i) of the PDP Act.

Issuing of a compliance notice

The decision on whether to issue a compliance notice required OVIC to look at the present circumstances and consider whether DFFH currently has reasonable controls in place to prevent similar breaches of IPP 3.1 and IPP 4.1.

Since the PA Report incident, DFFH has released specific Generative Artificial Intelligence Guidance to "help employees understand the risks, limitations and opportunities of using GenAI tools such as ChatGPT". It has also promoted this guidance through awareness raising activities.

While the content of this guidance is broadly fit for purpose, it must be noted that DFFH has almost no visibility on how GenAI tools are being used by staff. Despite the extent of use of GenAI tools across DFFH, it has no way of ascertaining whether personal information is being entered into these tools and how GenAI-generated content is being applied.

In these circumstances, the controls that DFFH has in place are insufficient to mitigate the risks that using GenAI tools will result in inaccurate personal information or in the unauthorised disclosure of personal information. This is particularly the case in child protection matters, where the risks of harm from using GenAI tools are too great to be managed by policy and guidance alone.

Given this, OVIC considers that a major gap in DFFH's controls is the use of technical solutions to manage employee access to GenAI tools. Specifically, the Deputy Commissioner considers that ChatGPT and similar GenAI tools should be prohibited from being used by Child

Protection employees. OVIC therefore issued a compliance notice requiring that DFFH must take the following specified actions:

1. Issue a direction to Child Protection staff setting out that they are not to use any web-based or external Application Programming Interface (API)-based GenAI text tools (such as ChatGPT) as part of their official duties. This direction must be issued by 24 September 20243.

2. Implement and maintain Internet Protocol blocking and/or Domain Name Server blocking to prevent Child Protection staff from using the following web-based or external API-based GenAI text tools: ChatGPT; ChatSonic; Claude; Copy.AI; Meta AI; Grammarly; HuggingChat; Jasper; NeuroFlash; Poe; ScribeHow; QuillBot; Wordtune; Gemini; and Copilot. The list does not incorporate GenAI tools that are included as features within commonly used search engines. This action must be implemented by 5 November 2024 and maintained until 5 November 2026.

3. Implement and maintain a program to regularly scan for web-based or external API-based GenAI text tools which emerge that are similar to those specified in Action 2 – to enable the effective blocking of access for Child Protection staff. This action must be implemented by 5 November 2024 and maintained until 5 November 2026.

4. Implement and maintain controls to prevent Child Protection staff from using Microsoft365 Copilot. This action must be implemented by 5 November 2024 and maintained until 5 November 2026.

5. Provide notification to OVIC upon the implementation of each of Specified Actions 1 – 4 explaining the steps taken to implement the respective Specified Actions.

6. Provide a report to OVIC on its monitoring of the efficacy of Specified Actions listed 1 – 4 on 3 March 2025; 3 September 2025; 3 March 2026; and 3 September 2026.

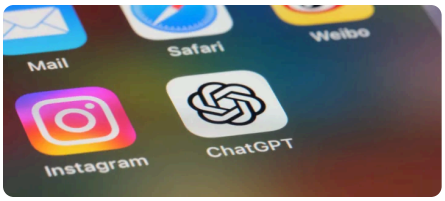
DFFH response to the investigation

OVIC welcomes DFFH's response to this report's findings and conclusions, as shown at Annexure B.

In summary, DFFH accepts the finding that there was a breach of IPPs 3.1 and 4.1 and commits to addressing the actions specified in the Compliance Notice within the required timeframes.

However, in its response DFFH contends that the report "did not find that any staff had used GenAI to generate content for sensitive work matters". In fact, the report presents the opposite – the Deputy Commissioner found on the balance of probabilities that CPW1 used ChatGPT to generate content which was used in a very sensitive work matter – the drafting of the PA Report which was submitted to the Children's Court for a child protection case.

Collapse



Vic case worker used ChatGPT to draft child protection report

itnews.com.au · 2024

Victoria's Department of Families, Fairness and Housing (DFFH) has been directed to ban and block access to a range of generative AI tools after a child protection worker used ChatGPT to draft a report submitted to the Children's Court.

The...

Read More

Victorian welfare agency banned from GenAI after child protection debacle

themandarin.com.au · 2024

Victoria's Department of Families, Fairness and Housing (DFFH) child protection service has been banned from using generative artificial intelligence in the workplace for at least a year.

The ban comes after an investigation into a case wor...

Read More

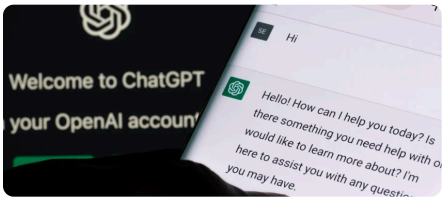
AI ban ordered after child protection worker used ChatGPT in Victorian court case

theguardian.com · 2024

Victoria's child protection agency has been ordered to ban staff from using generative AI services after a worker was found to have entered significant amounts of personal information, including the name of an at-risk child, into ChatGPT.

T...

Read More



Victorian child protection worker uses ChatGPT for protection report

cyberdaily.au · 2024

Victoria's child protection agency has been ordered to ban the use of AI tools after a case worker used ChatGPT to write a child's protection report, resulting in sensitive data being submitted and a number of inaccuracies being generated.

...

Read More

Victoria's child protection agency bans AI use after report debacle

newsbytesapp.com · 2024

What's the story

Victoria's child protection agency has imposed a ban on its staff using generative artificial intelligence (AI) services. This decision comes after an employee was discovered to have inputted substantial personal informatio...

Read More

Australian Information Commissioner Halts GenAI Use for Child Protection Agency as ChatGPT Downplays Risk

medianama.com · 2024

A state-level Australian Information Commissioner has ordered Victoria state's child protection agency to stop using generative AI services. According to the Information Commissioner, the agency staff entered a significant amount of persona...

Read More

Variants

A "variant" is an incident that shares the same causative factors, produces similar harms, and involves the same intelligent systems as a known AI incident. Rather than index variants as entirely separate incidents, we list variations of incidents under the first similar incident submitted to the database. Unlike other submission types to the incident database, variants are not required to have reporting in evidence external to the Incident Database. Learn more from the research paper.

Add Variant

Similar Incidents

By textual similarity

Did **our** AI mess up? Flag the unrelated incidents

Australian Automated Debt Assessment System Issued False Notices to Thousands

Jul 2015 · 39 reports

Australian Retailers Reportedly Captured Face Prints of Their Customers without Consent

May 2022 · 2 reports

Airbnb's Trustworthiness Algorithm Allegedly Banned Users without Explanation, and Discriminated against Sex Workers

Jul 2017 · 6 reports

← Previous Incident

Next Incident →