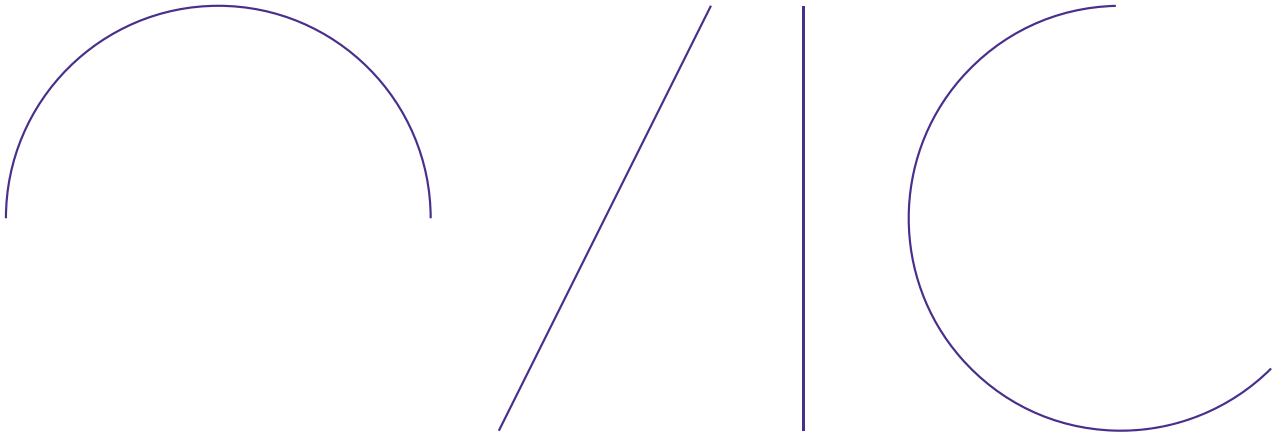
Office of the Victorian Information Commissioner

Investigation into the use of ChatGPT by a Child Protection worker

Under s8C(2)(e) of the *Privacy and Data Protection Act 2014*



Disclaimer

The information in this document is general in nature and does not constitute legal advice.

Copyright

Table of Contents

Foreword

It has not taken long for generative artificial intelligence (GenAI) to grow from a novelty issue to becoming a topic of everyday conversation. Most discussions we now hear about GenAI are split between grand pronouncements about its benefits and dire warnings about its risks.

Until relatively recently, these risks and benefits seemed hypothetical to public sector organisations and their employees who were not using GenAI. We know now that this is no longer the case, as these organisations and their employees are experimenting with using GenAI tools such as ChatGPT.

While some will have noted tangible benefits from this experimentation, what follows in this report is a very real example of the privacy risks associated with GenAI being realised – and the serious harm that may arise when personal information is inappropriately used with these tools. Unfortunately, the case involves a young child at risk of harm.

The investigation undertaken by the Deputy Commissioner found that a Child Protection worker entered a significant amount of personal and delicate information[1] into ChatGPT, including names and information about risk assessments relating to the child. The worker asked ChatGPT to assist in drafting a Protection Application Report – a report that is submitted to the Children's Court to inform decisions about whether a child requires protection.

As a result, the Large Language Model (LLM) on which ChatGPT is based played a role in describing the risks posed to a young child if they continued living at home with their parents, who had been charged with sexual offences.

The inappropriateness of this should be clear when we think of how LLMs work. LLMs do not use reasoning or understand context – they provide statistical predictions on the most likely words to respond to a user prompt. As described in the investigation report:

> *AI systems are not tasked with telling the truth. Sometimes people may mistakenly think that AI systems only get things wrong occasionally while otherwise telling the truth. We need to understand that AI systems make mistakes, so it is important to verify the accuracy of the output before relying on the model. This is especially important when people rely on AI systems to make decisions that affect themselves or others.*

The result in this case was a Protection Application Report that contained inaccurate personal information, downplaying the risks to the child. Fortunately, it did not change the outcome of the child's case, but it is easy to see the potential harm that could have arisen.

[1] The term "delicate information" is used in place of what could, in common usage, be described as "sensitive information". This is because "sensitive information" has a specific definition under the PDP Act – it is any personal information that falls within one of the nine categories listed in Schedule 1 of the Act (such as racial or ethnic origin; religious beliefs; or political opinions). What individuals may think of as information that is sensitive to them, for example, information they regard as embarrassing or secret, may not fall within one of the nine categories. The term 'delicate information' is used to refer to such information.

For example, the Protection Application Report mistakenly described a child's doll, that was used by the child's father for sexual purposes, as a mitigating factor, in that the parents had provided the child with "age appropriate toys". This description was clearly not the result of expert human analysis and reasoning of the facts of the case.

Further, through entering personal and sensitive information into ChatGPT, the information in this case was disclosed to OpenAI, an overseas company, and released outside the control of DFFH. OpenAI now holds that information and can determine how it is further used and disclosed.

The investigation found that the use of ChatGPT in this instance, was a serious breach of the Information Privacy Principles (**IPPs**). Given the seriousness, the Deputy Commissioner decided to issue a Compliance Notice on DFFH, which includes six specified actions. Most importantly, this includes the banning of the use of ChatGPT and other similar tools by Child Protection workers.

While some uses of GenAI may be beneficial, this report illustrates that there are currently circumstances where the privacy risks involved are simply too great – such as where highly delicate information is involved.

Ethics frameworks around the globe have indicated that AI should not be used in high-risk use cases. It is difficult to imagine a higher-risk use case than child protection, where an incorrect opinion could result in lasting serious harm to a child, parents, or both.

I therefore encourage all organisations to assess the risks involved in their employees' use of GenAI across their different functions and activities. In line with their obligations under the IPPs, organisations must put in place appropriate controls to mitigate these risks.

Sean Morrison
Information Commissioner
September 2024

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 ChatGPT[2] 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

[2] ChatGPT stands for Chat Generative Pre-Trained Transformer. It is an example of a GenAI tool that responds to a user's prompt by generating human-like text content.

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 2024[3].
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.

[3] The compliance notice that was issued to DFFH specified a deadline of 17 September 2024. However, DFFH subsequently sought an extension to comply with this specified action. OVIC agreed to extend the deadline to 24 September 2024.

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.

1. Background

1. In December 2023, the Department of Families, Fairness and Housing (**DFFH**) reported a privacy incident to the Office of the Information Commissioner (**OVIC**), indicating that a Child Protection worker (**CPW1**) had used ChatGPT[4] when drafting a Protection Application Report (**the PA Report**) incident.

2. The report had been submitted to the Children's Court, to provide information about risks of harm to a young child and whether they should be placed in Out of Home Care. However, at a further Court hearing around a week later, a legal officer who reviewed the report prior to the Court hearing identified deficiencies in the report that indicated that it had been drafted using a Large Language Model (**LLM**) tool, later identified to be ChatGPT.

3. DFFH investigated these concerns and confirmed that CPW1 had used ChatGPT in drafting the report. Its investigation also indicated potential other instances of the use ChatGPT as part of CPW1's child protection related duties.

4. OVIC conducted preliminary inquiries with DFFH and determined that the issues raised warranted more formal regulatory action. Therefore OVIC[5] commenced an investigation under section 8C(2)(e) of the *Privacy and Data Protection Act 2014* (**PDP Act**) in March 2024.

Child protection and its functions

5. The Victorian Child Protection Service (**Child Protection**) is part of DFFH, and is specifically directed to support children and young people at risk of harm or where families are unable to protect them.

Functions of Child Protection

6. The main functions of Child Protection are to investigate and respond to matters where it is alleged that a child is at risk of significant harm and needs protection. This includes making applications to the Children's Court if the child's safety cannot be ensured within the family home.

7. As such, understanding and assessing risk of harm to children are "at the heart"[6] of Child Protection's functions.

8. Among other things, assessing risk of harm to children requires Child Protection staff to properly analyse information, exercise their professional judgment, and make informed decisions. All such activities rely upon accurate and detailed information, properly recorded in various types of

[4] ChatGPT stands for Chat Generative Pre-Trained Transformer and is explained in greater detail below.

[5] References in the report to the carrying out of regulatory powers by OVIC under part 3 of the PDP Act means the exercise of powers by the Privacy and Data Protection Deputy Commissioner.

[6] DFFH, SAFER children framework guide The five practice activities of risk assessment in child protection, October 2021, p. ii. Available at: https://www.cpmanual.vic.gov.au/sites/default/files/2021-10/SAFER%20children%20framework%20guide%20October%202021.pdf

documentation – such as court reports, case notes, case plans, investigation plans, and risk assessments.

Protection Application Report

9. If Child Protection has concerns about a child's safety and welfare, it may make a Protection Application to the Children's Court. In doing so, Child Protection workers complete a Protection Application Report (**PA Report**) which is submitted to the court.

10. The PA Report is an essential element in protecting vulnerable children who require court ordered protective intervention to ensure their safety, needs and rights. The report needs to contain enough information and analysis to enable the Children's Court to make decisions about whether a child needs protection and, if so, the nature of the order[7] required to address the child's safety, development, and wellbeing needs.

11. The PA Report contains an assessment outlining the Child Protection worker's assessment and judgement of the risks and needs of the child, and the capacity of the parent/s to provide for the child's safety and development.[8]

Personal and sensitive information

12. Central to this investigation is the large volume of very personal and sensitive information captured and compiled throughout child protection cases.

13. For example, a Protection Application Report includes child and family details, summaries of events leading to a protection application, concerns, services accessed, family strengths and protective factors, a child's current circumstances, a risk assessment, and recommendation.

14. Any mismanagement of this delicate[9] information can have serious consequences for children and their families.

---

[7] If the Children's Court finds that the child needs protection, it may make one of the following protection orders – an order requiring a person to give an undertaking to the court, a family preservation order, a family reunification order, a care by Secretary order (giving parental responsibility to DFFH) or a long term care order.

[8] Department of Health and Human Services, Court report writing guide: For Victorian Child Protection Practitioners, v.1, May 2020.

[9] The term "delicate information" is used in place of what could, in common usage, be described as "sensitive information". This is because "sensitive information" has a specific definition under the PDP Act – it is any personal information that falls within one of the nine categories listed in Schedule 1 of the Act (such as racial or ethnic origin; religious beliefs; or political opinions). What individuals may think of as information that is sensitive to them, for example, information they regard as embarrassing or secret, may not fall within one of the nine categories. The term 'delicate information' is used to refer to such information.

Artificial intelligence and ChatGPT

AI and GenAI

15. An Artificial Intelligence (**AI**) system is a "machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment"[10].

16. Generative AI (**GenAI**) is a specific type of AI which responds to user prompts by generating new content such as text, images, audio, video, and code. GenAI tools do not reason, but perform their tasks through models based on statistical analysis of large data sets.

ChatGPT

17. ChatGPT is an example of a GenAI tool, developed by OpenAI. It is an online chatbot[11] which responds to a user's prompt by generating human-like text content. Since being released in November 2022 as a free web-based tool that any internet user can access, its use has become widespread.

18. ChatGPT relies on its LLM[12] to respond to prompts. The LLM is trained using vast amounts of publicly available information and information that has been scraped from the Internet and other sources, as well as data entered by users[13].

19. When a user enters a prompt, ChatGPT tries to detect patterns, context and meaning based on the LLM's training and any adjustments made by the AI's developer. It then makes a word-by-word prediction of the statistically most appropriate response. OpenAI describes that this is "similar to auto-complete capabilities on search engines, smartphones, and email programs."[14]

20. In this way, ChatGPT (like other GenAI tools) does not understand prompts and context in the same way as humans do, and does not use reasoning to provide a response.

---

[10] Organisation for Economic Co-operation and Development (OECD), Explanatory memorandum on the updated OECD definition of an AI system, OECD Artificial Intelligence papers, No.8 (March 2024), available at: https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en

[11] A chatbot can be defined as 'a computer program that interacts with humans through natural language conversations. Some chatbots use LLMs to generate content according to user inputs'. See Fan Yang, Jake Goldenfein, and Kathy Nickels, 'GenAI concepts', ADM+S and OVIC (Web Page, 2024), https://www.admscentre.org.au/genai-concepts/#chatbot (Gen AI concepts).

[12] See definitions of 'LLM' and 'Machine learning' in: 'Gen AI concepts', above n. 11.

[13] For details on how OpenAI uses user content to train its LLM, see: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

[14] https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed.

Privacy risks and ChatGPT

21. Despite its popularity and utility in some cases, there are a range of privacy risks[15] associated with the use of GenAI tools such as ChatGPT. Most relevant in the present circumstances are risks related to:

    - collection, use, and disclosure of inaccurate personal information
    - unauthorised disclosure of personal information.

22. Based on the fact GenAI tools make statistical predictions, generating content using ChatGPT may result in inaccurate personal information being generated, and subsequently used or disclosed, by the user.[16] The fact that inaccurate responses from ChatGPT may appear convincing and authoritative creates a risk that such inaccuracies may be overlooked by users.

23. Concerns about the risk of GenAI tools producing "hallucinations" – a term used to describe the generation of inaccurate information by AI tools – have been described in the following way:

> *AI systems are not tasked with telling the truth. Sometimes people may mistakenly think that AI systems only get things wrong occasionally while otherwise telling the truth. That is not true. We need to understand that AI systems make mistakes so it is important to verify the accuracy of the output before relying on the model. This is especially important when people rely on AI systems to make decisions that affect themselves or others.*[17]

24. The use of ChatGPT also creates the risk of unauthorised disclosures of personal information. Where a user inputs personal information into the "free" version of ChatGPT, that information is disclosed to OpenAI[18] and released outside the control of the relevant public sector organisation.

25. OpenAI then has ownership of that information, and can determine how it is used and disclosed – for example, using it to train its LLM, or sharing information with other third parties. This means there is limited ability for an organisation to take reasonable steps to protect its information after it has been entered into the LLM.

---

[15] For more information, see: OVIC, 'Public Statement: Use of personal information with ChatGPT', February 2024, available at: https://ovic.vic.gov.au/privacy/resources-for-organisations/public-statement-use-of-personal-information-with-chatgpt/

[16] The risks associated with inaccurate information being generated by GenAI tools have been noted in other important non-privacy contexts. See, for example: https://www.supremecourt.vic.gov.au/forms-fees-and-services/forms-templates-and-guidelines/guideline-responsible-use-of-ai-in-litigation

[17] See 'Gen AI concepts', above n.11. This is also recognised by OpenAI as a footer on the home screen of ChatGPT states 'ChatGPT can make mistakes. Check important info'.

[18] https://openai.com/policies/privacy-policy/.

OVIC's investigation

26. The Deputy Commissioner decided to commence an investigation based on the potential seriousness of the issue, noting the highly delicate nature of the personal information involved, and the significance of the rights and interest at stake. The Deputy Commissioner also considered the likely educative impacts of an investigation, noting the emerging prevalence of the use of GenAI tools across public sector organisations.

27. OVIC conducted the investigation with a view to deciding whether to issue a compliance notice to DFFH. Under section 78 of the PDP Act, OVIC may issue a compliance notice where it determines that:

    a. an organisation has contravened one or more of the IPPs;

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

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

28. To determine whether DFFH committed a serious, flagrant or repeated contravention of the IPPs, OVIC's primary focus was on the PA Report incident. However, OVIC also analysed other potential uses of ChatGPT by CPW1 as well as examining the general usage of ChatGPT across DFFH.

29. OVIC was primarily focussed on considering the controls that DFFH has in place to regulate the use of ChatGPT or any other GenAI platform by its staff. The investigation therefore considered the following IPPs:

    IPP 3.1 which requires DFFH to take reasonable steps to ensure that the personal information it collects, uses or discloses is accurate, complete and up to date.

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

30. To inform our findings under section 78 of the PDP Act, OVIC considered the following issues:

    - The circumstances and impacts of the PA Report incident
    - Other potential uses of ChatGPT by CPW1 and other members of their team
    - General usage of ChatGPT across DFFH
    - Controls that DFFH had in place to regulate the use of ChatGPT – at the time of the PA Report incident
    - Controls that DFFH had in place to regulate the use of ChatGPT – in the present day.

31. The investigation involved the following steps:

    - Meetings with DFFH senior employees
    - Analysis of documentary evidence, including policies, court report extracts, audit logs, management briefs, correspondence, and DFFH communications with employees

- Analysis of written commentary from DFFH addressing questions posed by OVIC.

DFFH response to the investigation

32. At the conclusion of the evidence gathering phase of the investigation, OVIC prepared a draft investigation report outlining the Deputy Commissioner's proposed findings, and preliminary view about whether to issue a compliance notice on DFFH.

33. DFFH was provided a reasonable opportunity to respond to the proposed findings and conclusions in the draft report. This report was finalised taking into account DFFH's response to the draft report. A copy of DFFH's formal response to the final report is included at **Annexure B.**

2. Discussion of findings

The PA Report incident

34. The relevant child protection case involved a young child. Child Protection received information that the child's parents had been charged in relation to sexual offences. These charges did not relate to sexual offences against the child.

35. Within days of the concerns being reported, Child Protection investigated[19], assessed that the child was at risk of unacceptable harm, and issued a Protection Application by Emergency Care to the Children's Court. This resulted in the Court making interim orders removing the child from the care of their parents.

36. The matter continued to come before the Children's Court as Child Protection sought to manage the protection of the child through the interim orders, and to obtain a final protection order setting out long term protection arrangements for the child.

37. As part of this, CPW1 prepared a PA Report supporting Child Protection's concerns about risks to the child. The team manager reviewed and signed off the report which was submitted to the Children's Court.

38. A DFFH legal representative reviewed the PA Report on the day of a further Court hearing approximately a week later and identified concerns relating to unusual language used in the report and the adequacy of the risk assessment. Suspecting that the PA Report had been drafted using ChatGPT, the representative reported their concerns to the relevant Child Protection Area.

39. Child Protection staff investigated the matter, which included interviewing CPW1 and their colleagues. CPW1 admitted that they had used ChatGPT in drafting the report, but denied inputting personal information as part of this.

40. Based on its investigation and review, DFFH determined that CPW1 had used ChatGPT in drafting the PA Report, and had input personal information in doing so. DFFH arranged with the Court for the PA Report to be withdrawn and a replacement one submitted. It also notified affected parties about the incident, as well as notifying OVIC.

Indicators of ChatGPT usage

41. In reviewing the PA Report incident, DFFH identified a range of indicators of possible use of ChatGPT in the report, relating to both the analysis and the language used in the report. These included:

    - Inappropriate language not commensurate with training and Child Protection guidelines

[19] For information about Child Protection investigations generally, see Child Protection Manual, https://www.cpmanual.vic.gov.au/policies-and-procedures/phases/investigation/investigation-policy.

- Inappropriate sentence structure
- Inaccurate information.

42. Of particular concern, the PA Report referenced a child's toy in conflicting ways – such that personal information was inaccurate[20] and the risk assessment was inappropriate.

> **Conflicting analysis about a child's toy**
>
> First, the PA Report referenced how the child's father used a particular toy, a doll, for sexual purposes.
>
> The report later referenced the same toy as a notable strength, in that the parents had provided the child with "age-appropriate toys", which was used to underscore their efforts to nurture the child's development needs.

43. DFFH advised OVIC's investigation that:

> *This paragraph does not correlate with the seriousness of the sexual harm and uses sophisticated language that describes the adults involved in an overly positive light. This overly positive wording minimises the level of risk posed to the child. "The presence of age-appropriate toys" is inappropriate in the context of sexual activity occurring with the child's doll; sexually deviant and violent behaviours are minimised ("alleged misconduct"); and stating the adults demonstrated care and support in their parenting alongside sexual violence and deviance highlights a double bind.*

44. This aspect suggests that information about the child protection case was entered into ChatGPT but, consistent with the how LLMs operate and the associated risks mentioned above[21], ChatGPT failed to properly understand the relevant context, and generated inappropriate and inaccurate content as a result. The generated content presents what should clearly be an indicator of risk to the child as an indication of positive caregiving capacity of the parents.

45. Additionally, a theme throughout the PA Report was the use of language and terminology that were not standard for reports of this nature, and which were not in keeping with the clear and concise style of writing required for reports under Child Protection's Court report writing guide.[22]

46. DFFH explained that, at times, the language was overly sophisticated, complex and descriptive with the use of unusual sentence structure and linking words. In other instances, the language and corresponding analysis was overly simplistic which, for example, served to minimise the significance of

[20] The content inaccurately described the actions of the father and mother, as well as inaccurately describing the care of the child and risks relating to them.

[21] See paragraphs 18 – 25 above.

[22] See above, n.8.

potential long-term impacts to the child. In other parts of the PA Report there was use of American spelling, as well as sentences that were "non-sensical".

47. OVIC found that CPW1 had used ChatGPT in drafting the PA Report on the basis that:

    - there were multiple indicators of ChatGPT usage throughout the report (as described above) – particularly the analysis of the child's toy, which was so inappropriate that it cannot reasonably be expected to have been provided by a human child protection practitioner; and
    - CPW1 admitted to DFFH that they had used ChatGPT in preparing the report.

Did CPW1 input personal information into ChatGPT?

48. Whether CPW1 input personal information into ChatGPT goes only to the question of whether there was an unauthorised disclosure of personal information to OpenAI. Regardless of whether CPW1 input personal information, it is clear that they collected[23] personal information through ChatGPT, used this in the PA Report, then disclosed it to the Court.

49. It is important to note that – for both DFFH and OVIC – it was impossible to objectively verify whether personal information was entered into ChatGPT because of organisations' lack of visibility of staff inputs into ChatGPT.

> **Lack of visibility of staff inputs to ChatGPT**
>
> Individual users of ChatGPT must create an account to use the platform. When logged into their account, users can see the history of prompts they entered and responses provided by ChatGPT.
>
> However, organisations cannot view this information – even when staff log in to ChatGPT on the organisation's systems. If an organisation checks a staff member's internet browsing history, logs will show that they accessed the ChatGPT website – but not what information the staff member put in, or how ChatGPT responded.

50. DFFH interviewed CPW1[24], who admitted using ChatGPT to generate the PA Report as well as using it on other occasions[25] – in order to save time and to present work more professionally (such as formulating dot points into full paragraphs).

51. However, CPW1 denied inputting personal information into ChatGPT, saying that they would remove names and identifying information. CPW1 also said that when using ChatGPT, they would check the content and make modifications to ensure it was relevant and made sense.

---

[23] When ChatGPT generates new content containing personal information in response to a user prompt, this generated content constitutes a new 'collection' of personal information for the purposes of the IPPs. See OVIC, 'Public Statement: Use of personal information with ChatGPT', above n.15.

[24] DFFH's engagement about the PA Report incident was limited owing to CPW1's resignation.

[25] See below, paras 60 – 61.

52. DFFH also interviewed staff in CPW1's team. They indicated that CPW1 had demonstrated their use of ChatGPT to others, and that this involved inputting client names into the tool to create content.

53. It must be stressed that the question of whether information constitutes personal information does not depend simply on whether it includes names or not. The definition of 'personal information' in the PDP Act[26] sets out that it is information about a person whose identity is apparent or can reasonably be ascertained[27]. Even where a person tries to de-identify information by removing names, it may be re-identified using the rest of the information as well as other information sources.

54. OVIC found, on the balance of probabilities, that CPW1 input personal information, including names[28], into ChatGPT in drafting the PA Report, despite CPW1 denying this. This finding was based on the following:

    - there were indications of ChatGPT throughout the PA Report as explained above
    - in those sections of the PA Report where there were signs that ChatGPT had been used, the names of individuals were frequently used, alongside other forms of personal information
    - given that one of CPW's reasons for using ChatGPT was to save time, it is to be doubted that they would have removed all identifying information before inputting it into ChatGPT and then re-inserted the identifying information into the PA Report
    - this level of care and quality assurance is open to further doubt given the clearly inappropriate nature of the LLM response, and the strangeness of aspects of the content in the report – which CPW1 did not notice or amend; and
    - as noted above, it was indicated that CPW1 had input client names into ChatGPT when demonstrating use of the tool to others.

Impacts of ChatGPT usage

55. Firstly, the use of ChatGPT in the PA Report resulted in inaccurate personal information being collected, used, and disclosed. That is, the report inaccurately described the actions and caregiving capacity of the parents, and inaccurately described risks posed to the child.

[26] PDP Act, section 3.

[27] Whether a person's identity can be reasonably ascertained in a given circumstance depends on a range of factors such as the nature and amount of information involved, and the nature of any individuals or entities who have access to the information (including additional information or other resources they have access to). See 'OVIC, IPP Guidelines, Key Concepts' for more analysis of what constitutes personal information and the factors to be considered as part of this.

[28] Based on finding that names were input into ChatGPT, OVIC determined that the identities of individuals mentioned in the report were apparent. It was therefore not necessary to assess whether their identities would have been reasonably ascertainable without these names being input into ChatGPT.

56. This downplayed the severity of the actual or potential harm to the child, with the potential to impact decisions about the child's care. DFFH described this impact:

> *[The use of ChatGPT] impacts the risk assessment by reducing the severity of the risk of sexual and psychological harm to the child. The actual or potential impact relating to potential ChatGPT use and/or poor practice is to minimise the actual or potential harm severity means [and] there is potential for an insufficient intervention by Child Protection to mitigate or stop ongoing harm to a child.*

57. Fortunately, despite having the potential to cause significant negative consequences, these deficiencies in the report did not ultimately change the decision-making of either Child Protection or the Court in relation to the child.

58. Secondly, by entering highly personal and sensitive information about the mother, father, carer, and child into ChatGPT, CPW1 disclosed this information to OpenAI. This unauthorised disclosure released the information from the control of DFFH. OpenAI was solely in control of any further uses or disclosures of this information.

59. The unauthorised disclosure of this type of information could clearly cause harm– such as emotional distress – to affected individuals. In the present case, the mother, father, and carer variously described feelings of anxiety and being overwhelmed when informed about their information being disclosed.

Potential other instances of ChatGPT usage by CPW1 and colleagues

60. When interviewed following the PA Report incident, CPW1 indicated that they had used ChatGPT on other occasions – in preparing Client Relationship Information System (CRIS) notes, client correspondence, and two court reports.

61. CPW1 indicated that they used ChatGPT regularly for around one month prior to the PA Report incident. CPW1 claimed that they used ChatGPT to save time and to present work more professionally.

62. DFFH also interviewed staff in CPW1's team. It was indicated that CPW1's use of ChatGPT had been well known within the team for a period of around 3 – 4 months and possibly longer, and that they had demonstrated to other team members how it could be used. There were no admissions that other members of the team had used ChatGPT.

63. DFFH decided to review all child protection cases that had been handled by CPW1's work unit within a 12-month period, to identify potential other uses of ChatGPT for child protection-related documents.

64. As well as seeking to identify possible usage of ChatGPT, the review also considered broader 'practice concerns' about CPW1's work unit. Amongst other things, these concerns related to the depth and thoroughness of the risk assessments and child protection interventions, in line with Child Protection's process requirements.

65. The review was conducted by senior Child Protection employees. It involved review of 796 cases in line with the Child Protection Practice Manual that is the primary point of reference for practitioners and managers regarding statutory child protection policy, procedures and supporting advice.

66. Through the review, DFFH sought to identify any closed cases that required further action from Child Protection, on the basis that actions taken on the case may have been inappropriate – either due to ChatGPT usage, practice issues, or both.

67. OVIC met with employees who carried out the review, who noted the difficulty in assessing whether ChatGPT had been used in drafting documents. This was understandable considering the lack of organisational visibility of inputs and outputs from ChatGPT, as noted at paragraph 49.

68. Reviewers noted that it was often difficult to form a view on whether inappropriate content within a document was a consequence of using ChatGPT, or was a result of the general practice deficiencies identified within the work unit. On occasion, this led to different staff reaching different conclusions on the issue.

69. Ultimately, DFFH found there were 100 cases with indicators that ChatGPT may have been used to draft child protection documents. The types of documents involved court reports, case notes, case plans and risk assessments. Some indicators and examples were as follows:

- Sophisticated language
- Overly positive descriptors
- Inaccurate information
- Unusual content
- Unusual terminology
- Unusual reference to legal intervention
- Unusual Child Protection intervention
- Nonsensical references
- American spelling and/or phrasing

Example 1: Overly positive, utopian language

"… underscoring her genuine care and unwavering dedication to the child's wellbeing. These family strengths and protective factors serve as a testament to the support system that has been established for the children, promoting their overall welfare and development."

Example 2: Sophisticated language

The mother highlighted the positive co-parenting relationship between herself and the father. They effectively communicate to provide the necessary support for the children. The mother expressed her commitment to monitor the children's behaviours closely and to address any concerns promptly.

Example 3: Unusual description of family home

The family home was a cluttered mess. Piles of unwashed dishes covered the kitchen countertops, along with empty pizza boxes and scattered food crumbs. The living room was a maze of toys, clothes, and scattered papers, with a thick layer of dust coating the furniture. The carpets were stained and littered with debris, and the air was heavy with a musty odor. Overflowing trash cans emitted an unpleasant smell, and dirty laundry was piled up in every corner. The bathrooms were grimy, with mildew creeping up the walls, and the bedrooms were strewn with clothes, books, and random items. It was clear that cleaning and organizing had taken a backseat in this chaotic household.

70. DFFH reviewers noted that despite the suspected use of ChatGPT, none of the further instances had the same potential impacts on the assessment of risk to children as was the case in the PA Report incident.

71. OVIC sought a sample of documents from the review, to gain an understanding of the types of indicators of ChatGPT usage as well as their potential impact. DFFH provided OVIC with nine examples.

72. On a review of these nine samples, OVIC shared DFFH's view that the negative impacts on the quality of the risk assessments did not reach the level of the PA Report incident.

73. Nevertheless, the samples suggested that ChatGPT had been used in the formulation of more documents than the PA Report, and that personal information had likely been input as part of this. This would have involved the unauthorised disclosure of large amounts of highly sensitive information to OpenAI.

General use of ChatGPT across DFFH

74. In response to the PA Report incident, DFFH sought to identify how many and which employees across the department had used ChatGPT in their official duties. To do this, DFFH analysed audit logs for July to December 2023.

75. Consistent with the inherent lack of visibility mentioned at paragraph 49 above, the logs showed only which staff had accessed the ChatGPT website on a department device, but not what information they had input or what content was generated by ChatGPT.

76. Through this analysis, DFFH identified that nearly 900 employees had accessed the ChatGPT website within this period. This represents almost 13 per cent of DFFH's workforce of around 7,000 employees.

77. In April 2024, DFFH's Chief Information Officer (**CIO**) sent an email to the nearly 900 employees, requesting them to help DFFH to better understand the use and risk exposure of using GenAI applications on government devices. The email asked recipients to provide examples of "the constructive and safe use of GenAI in the department's work". However, only ten employees

responded to the email. The uses identified included language translation, seeking resources, seeking explanation, generating an Outlook macro, study, document writing, and essay writing.

3. Did DFFH contravene the IPPs?

78. Given the above, OVIC found that that CPW1's use of ChatGPT regarding the PA Report incident involved the disclosure of personal information to an unauthorised third party (OpenAI) as well as the collection, use and disclosure of inaccurate personal information.

79. While the PA Report incident may therefore have involved the contravention of a number of IPPs[29], the investigation specifically considered DFFH's general management of the risks associated with the use of ChatGPT through the lens of IPP 3.1 and IPP 4.1.

80. Both of these IPPs require organisations to take "reasonable steps". OVIC therefore considered what controls DFFH had in place to protect personal information from unauthorised disclosure and to ensure the accuracy of the personal information it collected, used and disclosed – and whether these were reasonable in the circumstances.

What controls were in place at the time of the PA Report incident?

81. DFFH submitted that it had a range of controls in place to mitigate the privacy risks associated with the use of ChatGPT by Child Protection staff at the time of the PA Report incident. These controls consisted of:

- Acceptable Use of Technology Policy
- eLearning modules on privacy awareness and security awareness
- the DFFH values
- the VPS code of conduct
- Human Rights legislation and associated eLearning module
- communications to leadership and management by way of three education sessions in May 2023 that referred to data security, privacy and other risks associated with GenAI.

[29] This may have involved, for example, unnecessary and unfair collection of personal information in contravention of IPP 1.1 and 1.2; use and disclosure of personal information for unauthorised secondary purposes in contravention of IPP 2.1; and unauthorised transfer of personal information outside Victoria in contravention of IPP 9.1.

Were the controls reasonable?

82. What is considered 'reasonable' for the purpose of IPP 3.1[30] and IPP 4.1[31] depends on the particular context. In this case, the reasonableness of the controls DFFH had in place must be considered in the child protection context within which the PA Report incident took place.

83. This included consideration of the following factors:

    - the volume, nature, and sensitivity of the personal information
    - the potential consequences for individuals concerned if personal information was inaccurate or subject to unauthorised disclosure
    - the foreseeability of risks relating to inaccurate personal information or the unauthorised disclosure of personal information.

84. Child protection matters involve a significant volume of information about individuals and some of the most delicate personal information held by any government organisation.

85. This information is used to make decisions affecting significant rights and interests of individuals. It is used to assess risks to children and determine protection arrangements – including whether a child should be placed in out of home care, or whether they can remain living with their family.

86. As a result, it is clear that where personal information is either inaccurate or is inappropriately disclosed in this context, it can have serious impacts on the individuals concerned.

87. Weighed against this, at the time of the PA Report incident, DFFH should have been aware of the privacy risks posed by ChatGPT and other AI tools (as described at paragraphs 21 – 25) if they were to be used by Child Protection staff.

88. Taking these factors together, OVIC found that the controls DFFH had in place at the time of the PA Report were insufficient to mitigate these privacy risks.

89. DFFH pointed to existing policies, procedures, and training materials containing general obligations. In particular, it noted that CPW1 breached multiple provisions in the Acceptable Use of Technology Policy. Despite this, it remains that DFFH did not provide clear training and guidance to employees regarding the use of ChatGPT and GenAI more broadly. It could not be expected that staff would gain an understanding of whether they could use novel GenAI tools like ChatGPT or how to appropriately use them based only on existing general guidance materials.

[30] See 'OVIC, IPP Guidelines, IPP 3 – Data Quality', para 3.21 – 3.26 for a discussion of the factors that are relevant to determining reasonable steps in relation to IPP 3.1.

[31] See 'OVIC, IPP Guidelines, IPP 4 – Data Security', para 4.8 – 4.27 for a discussion of the factors that are relevant to determining whether a security measure is reasonable in relation to IPP 4.1.

90. While DFFH referred to education sessions which covered risks associated with GenAI, these were only directed at managers and leaders. DFFH advised that these sessions were designed to allow attendees to have contextual discussions with their team members. However, this was not an effective way of educating the general workforce about how GenAI tools work and the privacy risks associated with them.

91. There was no evidence that at the time of the PA Report incident, DFFH had made any other attempts to educate or train non-managerial or non-leadership staff about how GenAI tools work and the privacy risks associated with them.

92. Similarly, there were no specific 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.

93. Essentially, DFFH had no controls targeted at addressing specific privacy risks associated with ChatGPT and GenAI tools more generally. OVIC therefore found that DFFH contravened:

    - IPP 3.1 – by failing to take reasonable steps to mitigate risks that ChatGPT use would result in the collection, use and disclosure of inaccurate personal information

    - IPP 4.1 – by failing to take reasonable steps to mitigate risks that ChatGPT use would result in the unauthorised disclosure of personal information.

94. Taking into account the nature of the personal information involved, and the impacts of the PA Report incident as set out in paragraphs 55 – 69, OVIC determined that the contraventions were "serious"[32] for the purposes of section 78(1)(b)(i).

4. Whether to issue a compliance notice

95. A compliance notice may be issued where it appears to OVIC that there has been a serious contravention of the IPPs. A compliance notice requires that, within a specified timeframe, an organisation must take action specified by OVIC in order to ensure compliance with the IPPs.

96. The decision on whether to issue a compliance notice therefore required OVIC to move on from looking back at the PA Report incident, to considering the present circumstances and whether DFFH currently has reasonable controls in place to prevent similar breaches of IPP 3.1 and IPP 4.1. As part of this, OVIC analysed:

    - what controls DFFH currently has in place to regulate the use of ChatGPT and other GenAI tools by Child Protection staff

[32] See OVIC, Regulatory Action Policy p. 18-19 for a discussion of factors considered when determining whether a contravention of the IPPs is 'serious'.

- whether these controls are reasonable to protect against the kinds of inaccurate personal information and unauthorised disclosures that arose in the PA Report incident.

What controls are in place now to prevent further inappropriate use of ChatGPT by Child Protection staff?

97. Since the PA Report incident occurred, DFFH has introduced additional controls to mitigate the privacy risks associated with GenAI tools like ChatGPT. These are described below.

Generative AI Guidance

98. DFFH released 'Generative Artificial Intelligence Guidance' (**GenAI Guidance**) to all staff on 25 October 2023. The guidance seeks to "help employees understand the risks, limitations and opportunities of using GenAI tools such as ChatGPT".

99. The guidance identifies two "critical rules" applicable to GenAI tools:

    - Employees should be able to explain, justify and take ownership of their advice and decisions
    - Employees should assume that any information they input into public GenAI tools could become public. They must not input anything that could reveal classified, personal or otherwise sensitive information.

100. The guidance refers to existing DFFH policies; includes a list of "dos and don'ts"; and covers risks and limitations associated with GenAI tools. In total, the guidance identifies and cautions against a comprehensive range of risks. Some examples include the following actions which it directs staff to avoid:

    - entering personal information as well as "any client or case information" into public facing web-based applications
    - relying on GenAI as the only input to work because "it should not replace your own research, analysis and content development"
    - using these tools for "any query that is complex or sensitive, or where local context and nuance is critical", including "healthcare, housing or child protection queries".

101. The directive style of these "dos and don'ts" is somewhat watered-down, however, by overarching advice that staff must "conduct their own risk assessment which includes balancing any potential benefit and risks of using an GenAI tool".

102. The guidance lacks real-life examples or case studies of appropriate and inappropriate use of GenAI tools. In particular, while the guidance permits their use, it does not provide examples of appropriate and beneficial use cases.

Employee awareness

103. DFFH promoted its GenAI Guidance through awareness raising activities which involved:

- the guidance being listed in a whole-of-department newsletter on 20 November 2023
- the guidance being listed in a newsletter to all DFFH people managers on 29 January 2024
- the guidance being promoted by two Deputy Secretaries in an email to all staff in their units in December 2023 and February 2024 respectively.

Training

104. DFFH's GenAI Guidance identifies key knowledge and skills required to appropriately use GenAI tools. It says that this requires "the individual to ask the right questions or prompts, to recognise what to trust or use, and to assess quality and bias. Having the right domain expertise and the skills to work with AI generated outputs is critical to ensuring adequate human oversight and accountability".

105. However, there is no evidence that DFFH has provided any training to equip staff with such skill and knowledge, and it is unclear whether or not DFFH staff have the training in risk assessment to adequately assess what constitutes a high-risk use case for LLM use.

106. DFFH noted that it held 'Lunch and Learn' sessions in October 2023 and March 2024 that were attended by 550 and 500 staff respectively. However, these did not engage with privacy risks and related mitigations strategies in a meaningful way.

107. The October session was titled 'Artificial intelligence and phishing' which gave a brief explanation of GenAI with a description of positive use cases and one bullet point covering "risks and limitations". The majority of the associated slides focussed on external threats from AI through phishing.

108. The March session was titled 'Impersonation and Artificial Intelligence from a hacker's perspective' which also focussed largely on external threats posed by artificial intelligence. It did, however, cover the "dos and don'ts" from the GenAI guidance – with a real-life example of inappropriate input of personal information – and provided basic tips on adjusting privacy and security settings on ChatGPT.

Identification of known users of ChatGPT and understanding of uses cases

109. As explained at paragraph 77, DFFH's CIO targeted almost 900 users who had accessed the ChatGPT website (based on a review of users between July and December 2023) with an email in April 2024. This email referred recipients to relevant policies, and included appropriate messaging that aimed to reinforce the GenAI Guidance and employee responsibilities when using ChatGPT.

110. However, the email sought responses from staff to provide DFFH with an understanding of current use cases of ChatGPT. Given that only 10 people responded, it must be concluded that DFFH is unaware of the nature, extent and appropriateness of ChatGPT usage throughout the department.

111. DFFH has since commenced ongoing monitoring of logs to identify staff use of specific GenAI tools and send automated reminders of departmental guidelines to those staff.

Technical controls

112. DFFH advised that it has no technical means of verifying if personal information has or is being input into ChatGPT by employees.

113. It also stated that it has "the ability to monitor or prevent users accessing nominated websites (Blacklisting)" but there was no evidence that this capability is applied to any GenAI tools.

Are DFFH's controls sufficient to ensure compliance with the IPPs?

114. From the above, it can be summarised that DFFH has attempted to mitigate the privacy risks associated with GenAI tools like ChatGPT solely by way of its GenAI guidance along with existing policies.

115. While the content of the GenAI 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. It has no way of ascertaining whether personal information is being entered into GenAI tools and how GenAI-generated content is being applied. Further, as is always the case with policy and guidance, there is no way of guaranteeing that all staff will properly read, understand, and apply these.

116. This situation is made all the more concerning when considering that the use of ChatGPT and other GenAI tools is reasonably common across DFFH. Many areas of DFFH work with personal and sensitive information.

117. In these circumstances, the controls that DFFH has in place are insufficient when considering the child protection context, and the level of risks posed by the collection, use, and disclosure of inaccurate personal information and the unauthorised disclosure of personal information.

118. As the PA Report incident illustrates, the inherent limitations of GenAI tools can lead to inaccurate personal information in child protection cases. This has the potential to have significant negative impacts affecting decisions about risks to a child's safety and whether they require protection. Similarly, the volume and sensitivity of the personal information involved in child protection cases means that unauthorised disclosure can be very harmful.

119. Simply put, in child protection matters the risks of harm from using GenAI tools are too great to be managed by policy and guidance alone. At present, there are insufficient controls in place regarding staff access to GenAI tools coupled with a lack of assurance capabilities to verify that such use is appropriate. In other words, these controls are insufficient to prevent a re-occurrence of incidents like the PA Report incident.

120. Given this, OVIC considers that a major gap in DFFH's controls is the use of technical solutions to manage employee access to ChatGPT. Specifically, OVIC considers that ChatGPT and similar GenAI tools should be prohibited among Child Protection employees.

OVIC's decision to issue a compliance notice

121. Based on the seriousness of the confirmed PA Report incident and the contraventions of IPP 3.1 and 4.1, as well as the gaps in DFFH's current controls, OVIC decided to issue a compliance notice, a copy of which is included at **Annexure A.** The compliance notice requires DFFH to:

Specified action 1

DFFH must 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 2024[33].

Specified action 2

DFFH must 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.

Specified action 3

DFFH must 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.

Specified action 4

DFFH must 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.

Specified action 5

The Organisation must provide notification to OVIC upon the implementation of each of Specified Actions 1 – 4 explaining the steps taken to implement the respective Specified Actions.

[33] The compliance notice that was issued to DFFH specified a deadline of 17 September 2024. However, DFFH subsequently sought an extension to comply with this specified action. OVIC agreed to extend the deadline to 24 September 2024.

Specified action 6

> The Organisation must 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.

122. OVIC recognises that technology relating to LLMs and GenAI is evolving. The position that these tools should not be used by Child Protection staff is based on an assessment of their current limitations, the associated privacy risks, and DFFH's current controls.

123. In the event significant advances are made to enable LLMs to better understand context, it is possible that the risk environment will change in the future. Similarly, it is open to DFFH to demonstrate that improvements it has made to its controls would ensure compliance with IPP 3.1 and IPP 4.1 without the need for the above specified actions.

124. It may be that DFFH wishes to revisit the matter of the use of LLMs and GenAI in the child protection context within the two-year compliance notice period, in the event of such changes.

125. Should DFFH wish to depart from any of specified actions 2 – 4 within the two-year compliance notice period, it may apply to OVIC to amend the compliance notice by removing one or more of these actions. Any such application must include details and evidence of additional controls that DFFH has implemented to mitigate the privacy risks associated with the use of GenAI tools by Child Protection staff and ensure compliance with IPP 3.1 and IPP 4.1

126. The Deputy Commissioner believes there may be some specific use cases where the risk is less than others, but that child protection, by its nature, requires the very highest standards of care. Any application to vary the specified actions in relation to Child Protection staff, information, or activities would need to be accompanied by the highest standards of verifiable evidence.

Compliance notice

Under section 78 of the *Privacy and Data Protection Act 2014 (Vic)*

Office of the Victorian Information Commissioner

To: **Department of Families, Fairness and Housing**
50 Lonsdale Street
Melbourne VIC 3000
(the **Organisation**)

I, Rachel Dixon, pursuant to section 78(1) of the *Privacy and Data Protection Act 2014 (Vic)* (the **PDP Act**), serve this compliance notice under Division 9 of Part 3 of the PDP Act.

Background

1. Having conducted an investigation pursuant to my function under section 8C(2)(e) in conjunction with section 8B(1)(a) of the PDP Act, I am satisfied that:

   a. In or around September 2023, a Child Protection worker (**CPW1**) of the Organisation used ChatGPT, a web-based generative Artificial Intelligence (**GenAI**) tool, when drafting a Protection Application Report. This report was later submitted as part of Children's Court proceedings related to the Child Protection matter.

   b. CPW1's use of ChatGPT in this way resulted in:

      i. the collection, use, and disclosure of inaccurate personal information; and

      ii. the unauthorised disclosure of personal and sensitive information to OpenAI, the company which operates ChatGPT.

   c. The Organisation contravened Information Privacy Principle (**IPP**) 3.1 by failing to take reasonable steps to mitigate risks that ChatGPT use by its staff would result in the collection, use and disclosure of inaccurate personal information.

   d. The Organisation contravened IPP 4.1 by failing to take reasonable steps to mitigate risks that ChatGPT use by its staff would result in the unauthorised disclosure of personal information.

   e. The contraventions of IPPs 3.1 and 4.1 were serious for the purposes of section 78(1)(b)(i) of the PDP Act.

Specified Actions and Specified Periods

2. In accordance with section 78(2) of the PDP Act, this compliance notice requires the Organisation to take the below specified actions within the specified periods for the purpose of ensuring compliance with IPPs 3.1 and 4.1.

Specified Action 1

The Organisation must 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 17 September 2024.

Specified action 2

The Organisation must 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

Specified action 3.

The Organisation must 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.

Specified action 4

The Organisation must 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.

Specified action 5

The Organisation must provide notification to OVIC upon the implementation of each of Specified Actions 1 – 4 explaining the steps taken to implement the respective Specified Actions.

Specified action 6

The Organisation must 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.

Enforcement of this compliance notice

3. The Organisation must comply with this compliance notice.
4. It is an indictable offence not to comply with a compliance notice served that is in effect. The penalty for this offence is:
    a. 600 penalty units, in the case of an individual; and
    b. 3000 penalty units, in the case of a body corporate.

Application for review

5. An individual or organisation whose interests are affected by my decision to serve this compliance notice may apply to the Victorian Civil and Administrative Tribunal for review of my decision.

6. An application for review must be made within 28 days after the later of -

   (a) the day on which I made this decision; or

   (b) if, under the *Victorian Civil and Administrative Tribunal Act 1998*, the individual or organisation requests a statement of reasons for the decision, the day on which the statement of reasons is given to the individual or organisation or the individual or organisation is informed under section 46(5) of that Act that a statement of reasons will not be given.

Rachel Dixon
Privacy and Data Protection Deputy Commissioner
3 September 2024

Annexure B

Department of Families, Fairness and Housing

50 Lonsdale Street
Melbourne Victoria 3000
Telephone: 1300 475 170
GPO Box 1774
Melbourne Victoria 3001
www.dffh.vic.gov.au
DX 210319

BAC—EOB-535

Rachel Dixon
Deputy Commissioner, Privacy and Data Protection
Office of the Victorian Information Commissioner
PO Box 24274
Melbourne VIC 3001

Dear Ms Dixon

Thank you for your letter of 3 September 2024 providing the final report (the Report) for your investigation into the department's use of generative artificial intelligence (GenAI) application ChatGPT.

I accept that there was an unauthorized use of ChatGPT by a Child Protection Practitioner when developing a Protection Application Report for court and that the department should have done more to ensure greater data quality and better data security. I acknowledge that, as a result, there has been a breach of Information Privacy Principle (IPP) 3.1 (data quality) and IPP 4.1 (data security).

The department takes seriously its responsibility for protecting each person's human right to privacy. This includes ensuring that the personal, sensitive, and delicate information the department holds about some of our most vulnerable Victorians is collected and used in accordance with the IPPs at all times.

Response to the Investigation Report findings

The Department agrees with the Report (paragraphs 49, 68 and 77) that it is not possible to definitively confirm that personal information has ever been entered into ChatGPT by the Child Protection Practitioner who no longer works in the department. The Report did not find that any staff had used GenAI to generate content for sensitive work matters, in line with the requirements of the *Generative Artificial Intelligence Guidance* and *Acceptable use of the department's technology policy*. The department contends that this was an isolated incident and the use of GenAI is not prolific.

Further, the Report confirms that (as per paragraph 70) no decisions were changed for the child in this Child Protection case, nor were any decisions impacted by potential use of

ChatGPT following a broader review of cases of the former Child Protection Practitioner's team.

The department agrees with the significant risk to data quality (IPP3.1) that insufficiently checked outputs of even secure and sanctioned GenAI can present, including internally tenanted Microsoft 365 Copilot.

The department accepts the findings of the Report and commits to all reasonable efforts to address Specified Actions in the required timelines. The department also notes the process to seek advice from OVIC on removal of the Specified Action requirements.

Should you wish to discuss this matter further, please contact Michael Mefflin, Executive Director, Service Agreement and Quality Systems at the Department of Families Fairness and Housing on (03) 8633 4553 or <michael.mefflin@dffh.vic.gov.au>.

Yours sincerely

Peta McCammon
Secretary

18/09/2024