GENERAL

December 2023

Principles for responsible, trustworthy and privacy-protective generative AI technologies

Office of the Privacy Commissioner of Canada
Commissariat à la protection de la vie privée du Canada

Information and Privacy Commissioner of Ontario
Commissaire à l'information et à la protection de la vie privée de l'Ontario

Commission d'accès à l'information du Québec

OIPC Office of the Information & Privacy Commissioner Nova Scotia

ombud NEW BRUNSWICK NOUVEAU-BRUNSWICK

Manitoba Ombudsman

oipc OFFICE OF THE INFORMATION & PRIVACY COMMISSIONER FOR BRITISH COLUMBIA

OFFICE OF THE INFORMATION & PRIVACY COMMISSIONER for Prince Edward Island

Office of the Saskatchewan Information and Privacy Commissioner

Office of the Information and Privacy Commissioner of Alberta

OFFICE OF THE INFORMATION AND PRIVACY COMMISSIONER NEWFOUNDLAND AND LABRADOR

OFFICE OF THE INFORMATION AND PRIVACY COMMISSIONER NORTHWEST TERRITORIES

Yukon Information and Privacy Commissioner

OFFICE OF THE INFORMATION AND PRIVACY COMMISSIONER OF NUNAVUT

Introduction

Within the overall context of ongoing advances in AI technologies, one version of the technology has seen particularly rapid development, proliferation of use cases, and increase in adoption of late: generative AI. Generative AI is a subset of machine learning in which systems are trained on massive information sets – often including personal information – to generate content such as text, computer code, images, video, or audio in response to a user prompt. This content is probabilistic, and may vary even in response to multiple uses of the same or similar prompts.

Authorities across multiple fields around the world are recognizing the potential risks posed by this technology, including the June 2023 publication of a joint statement from G7 data protection and privacy authorities on generative AI technologies[1], the November 2023 G7 Leaders Statement which included guiding principles and a code of conduct for organizations developing advanced AI systems[2], and the October 2023 Global Privacy Assembly resolution on generative AI systems.[3] The Office of the Privacy Commissioner of Canada (OPC) and its counterparts in British Columbia, Quebec and Alberta also have an open investigation relating to a particular generative AI service. [4] Privacy authorities from countries around the world have recently called on organizations to exercise great caution before scraping[5] "publicly accessible" personal information, which is still subject to data protection and privacy laws in most jurisdictions.[6] Such scraping is common practice when training generative AI systems. Privacy authorities have also been working with their counterparts in related fields – such as human rights commissioners – to call for strong guardrails that ensure AI systems are safe, privacy protective, transparent, accountable, and human rights affirming.[7]

While generative AI tools may pose novel risks to privacy and raise new questions and concerns about the collection, use and disclosure of personal information, they do not occupy a space outside of current legislative frameworks. Organizations developing, providing, or using generative AI are obligated to ensure that their activities comply with applicable privacy laws and regulations in Canada. Organizations should also recognize that to build and maintain a digital society in which innovation is socially-beneficial and human dignity is protected, AI development and use have to be responsible and trustworthy.

[1] Statement on Generative AI – Roundtable of G7 Data Protection and Privacy Authorities. June 21, 2023.

[2] G7 Leaders' Statement on the Hiroshima AI Process. October 30, 2023.

[3] Global Privacy Assembly – Resolution on Generative Artificial Intelligence Systems. October 2023.

[4] Office of the Privacy Commissioner of Canada – OPC to investigate ChatGPT jointly with provincial privacy authorities. May 25, 2023.

[5] Data scraping refers to automated extraction of data generally from one or more websites.

[6] Joint statement on data scraping and the protection of privacy. August 24, 2023.

[7] Joint statement by the Information and Privacy Commissioner of Ontario and the Ontario Human Rights Commission on the use of AI technologies. May 25, 2023.

About this document

When applying the principles set out below, developers, providers and organizations using generative AI should give particular consideration to their mutually-shared responsibility to identify and prevent risks to vulnerable groups, including children and groups that have historically experienced discrimination or bias.

Developers, providers, and organizations using generative AI systems must all actively work to ensure the fairness of these systems. When developing a generative AI system, this means evaluating the training data sets to ensure that they do not replicate, entrench, or amplify historical or present biases – or introduce new biases. When deploying such a system, this could mean establishing additional oversight and review of outputs, or enhanced monitoring for potential adverse effects. Without these steps, the use of generative AI models and applications may be more likely to result in discriminatory outcomes based on race, gender, sexual orientation, disability, or other protected characteristics, particularly where they are used as part of an administrative decision-making process (whether or not that process is fully automated) or in highly impactful contexts such as health care, employment, education, policing, immigration, criminal justice, housing or access to finance.

Children are particularly at high risk of significant negative impact by AI technologies, including generative AI. They may be less able than adults to identify or challenge biased or inaccurate information, or be more prone to having their agency limited by an AI that generates information based on a restricted world view. Children should be able to benefit from technology safely and free from fear that they may be targeted, manipulated, or harmed. Young people are also generally less able to understand and appreciate the long-term implications of data collection, use and disclosure which is why they need even greater privacy safeguards.

Developers, providers and organizations using generative AI tools should work together to ensure that risks to vulnerable populations are mitigated, including through important protective measures such as privacy impact assessments.

In this document, we identify considerations for the application of key privacy principles to generative AI technologies.[8] We recognize that generative AI is an emerging field, and that our understanding of it will evolve over time. Canada's federal, provincial, and territorial privacy commissioners will continue to explore this complex topic and may provide guidance or other resources as we learn more about the technology and its potential risks, including as formal investigations related to the technology are completed.

Obligations under privacy legislation in Canada will vary by nature of the organization (such as whether it is in the private, health, or public sector) as well as the activities it undertakes. As such, while we use "should" throughout this document many of the considerations listed will

[8] Noting that many of these considerations will be equally applicable to most, if not all, AI technologies.

be required for an organization to comply with applicable privacy law. Organizations are responsible for understanding, and complying with, these laws. We also note that the principles set out below do not exhaustively reflect all compliance requirements under privacy and other laws and do not bind any specific investigations or adjudications carried out by Canada's federal, provincial, or territorial privacy commissioners, depending on the individual circumstances of each case.

Intended audience for this document

This document is intended to help organizations developing, providing or using generative AI apply key Canadian privacy principles. For this document, we use the following terminology:

- Developers and Providers[9]: Individuals or organizations that develop (including training) foundation models or generative AI systems, or that put such services onto the market. In short, those organizations that determine how a generative AI system operates, how it is initially trained and tested, and how it can be used.
- Organizations using Generative AI: Organizations (or individuals acting on behalf of an organization) using a generative AI system as part of their activities. This could include both public-facing uses (i.e. a generative AI-based chatbot to interact with clients) or private use (i.e. the use of generative AI as part of a decision-making system). Organizations that refine a foundation model for a specific purpose (such as by further training it on a dataset proprietary to the organization) are included in this category.

An organization might shift between or play multiple roles at once. The activities undertaken (including collection, use, or disclosure of personal information) by an organization will also vary within each group. However, the division into 'developers and providers' and 'organizations using generative AI' is a useful way to examine the application of privacy principles to multiple elements of the generative AI ecosystem[10].

For clarity, these Principles focus on privacy legislation and regulation, and how they may apply to organizations. However, we recognize that individuals or organizations may have further obligations, restrictions, or responsibilities pursuant to other laws, regulations, or policies.

[9] This encompasses "AI provider", "AI platform provider", "AI device or product provider", and "AI developer" as defined by ISO/IEC 22989:2022, *Artificial intelligence concepts and terminology*.

[10] Noting that these are not the only actors in the AI ecosystem, which can include among many others data subjects, data brokers or other organizations who collect and sell data for training purposes, and individual users of public-facing generative AI systems.

Special consideration: The unique impact on vulnerable groups

When applying the principles set out below, developers, providers and organizations using generative AI should give particular consideration to their mutually-shared responsibility to identify and prevent risks to vulnerable groups, including children and groups that have historically experienced discrimination or bias.

Developers, providers, and organizations using generative AI systems must all actively work to ensure the fairness of these systems. When developing a generative AI system, this means evaluating the training data sets to ensure that they do not replicate, entrench, or amplify historical or present biases – or introduce new biases. When deploying such a system, this could mean establishing additional oversight and review of outputs, or enhanced monitoring for potential adverse effects. Without these steps, the use of generative AI models and applications may be more likely to result in discriminatory outcomes based on race, gender, sexual orientation, disability, or other protected characteristics, particularly where they are used as part of an administrative decision-making process (whether or not that process is fully automated) or in highly impactful contexts such as health care, employment, education, policing, immigration, criminal justice, housing or access to finance.

Children are particularly at high risk of significant negative impact by AI technologies, including generative AI. They may be less able than adults to identify or challenge biased or inaccurate information, or be more prone to having their agency limited by an AI that generates information based on a restricted world view. Children should be able to benefit from technology safely and free from fear that they may be targeted, manipulated, or harmed. Young people are also generally less able to understand and appreciate the long-term implications of data collection, use and disclosure which is why they need even greater privacy safeguards.

Developers, providers and organizations using generative AI tools should work together to ensure that risks to vulnerable populations are mitigated, including through important protective measures such as privacy impact assessments.

Principles for the Development, Provision, and Use of Generative AI systems

1. **Legal Authority and Consent**
   Ensure legal authority for collecting and using personal information; when consent is the legal authority, it should be valid and meaningful.

   All parties should:

   - Know and document their legal authority for collection, use, disclosure and deletion of personal information that occurs as part of the training, development, deployment, operation, or decommissioning of a generative AI system.
   - Ensure that where consent is the legal authority for collection, use or disclosure of personal information, it is valid and meaningful.[11] Consent should be as specific as possible, and deceptive design patterns should be avoided.
   - Ensure that where personal information is sourced from third parties, the third parties have collected it lawfully and have authority to disclose it.
   - Be mindful that the inference of information about an identifiable individual (such as outputs about a person from a generative AI system) will be considered a collection of personal information, and as such would require legal authority.
   - In contexts where information is sensitive and consent (even where provided) may be inappropriate or inadequate, such as in health care, establish a separate review process which takes into account both the privacy and ethics of the proposed use of the information and which is subject to independent oversight.

2. **Appropriate Purposes**
   Collection, use and disclosure of personal information should only be for appropriate purposes.

   All parties should:

   - Ensure that any collection, use or disclosure of personal information associated with a generative AI system be for appropriate purposes. In many Canadian jurisdictions, this means for purposes that a reasonable person would consider appropriate in the circumstances.[12]
   - Consider also the legitimacy of the *manner* of any collection, use and disclosure of personal information in relation to a generative AI system. This includes consideration of whether the use of the generative AI system is appropriate for the specific application.

[11] Office of the Privacy Commissioner of Canada. Guidance for obtaining meaningful consent. Revised: August 13, 2021.

[12] "Quebec's legislation refers to terms such as collection for a 'serious and legitimate reason'."

Developers and providers of generative AI systems should:

- Not develop or put into service generative AI systems that violate "no-go zones" [13] such as profiling that may lead to unfair, unethical, or discriminatory treatment, or creating outputs that threaten fundamental rights and freedoms.
- Use an adversarial or red team[14] testing process to identify potential unintended inappropriate uses of the generative AI system.
- Where potential unintended inappropriate uses are identified, take appropriate steps to mitigate the likelihood of, or potential risks associated with, such uses. This might include establishing technical measures to prevent the inappropriate use or developing appropriate use policies to which individuals or organizations using the generative AI system must agree in advance of use.

Organizations using generative AI systems should:

- Only use generative AI tools that respect privacy laws and best practices, including with respect to the personal information collected or used for training or operation of the system.
- Avoid prompting a generative AI system to re-identify any previously de-identified data.
- Monitor for, and notify developers or providers of, potential inappropriate uses or biased outcomes that have not been disclosed as a potential limitation of the system.
- Avoid inappropriate uses of generative AI tools, including 'no-go zones' such as the collection, use, or disclosure of personal information that is otherwise unlawful; profiling or categorization that may lead to unfair, unethical, or discriminatory treatment that is contrary to human rights law; the collection, use, or disclosure of personal information for purposes that are known or likely to cause significant harm to individuals or groups, or activities which are known or likely to threaten fundamental rights and freedoms.
- If use of a generative AI system is detected to violate a 'no-go zone', cease the activity.

[13] Office of the Privacy Commissioner of Canada. Guidance on inappropriate data practices: Interpretation and application of subsection 5(3). May 2018.

[14] In summary, adversarial and red-team testing refers to processes in which the tester (potentially part of the organization, but generally not part of the development team) attempts to defeat or bypass system protections or policies, with the goal of identifying required improvements or changes.

Potential emerging no-go zones

Firm rulings on the legality of certain practices – such as through investigative or legal findings – have not yet been made in the context of generative AI, nor have Canada's federal, provincial or territorial privacy commissioners issued policy positions on generative AI no-go zones.

However, we anticipate (without binding future investigations, legal findings or policy positions) that such no-go zones may include purposes such as:

- the creation of AI content (including deep fakes) for malicious purposes, such as to bypass an authentication system or to generate intimate images of an identifiable person without their consent;
- the use of conversational bots to deliberately nudge individuals into divulging personal information (and, in particular, sensitive personal information) that they would not have otherwise; or,
- the generation and publication of false or defamatory information about an individual.

3. Necessity and proportionality

Establish the necessity and proportionality of using generative AI, and personal information within generative AI systems, to achieve intended purposes.

All parties should:

- Use anonymized, synthetic, or de-identified data rather than personal information where the latter is not required to fulfill the identified appropriate purpose(s).

Organizations using generative AI systems should:

- Consider whether the use of a generative AI system is necessary and proportionate, particularly where it may have a significant impact[15] on individuals or groups. This means that the tool should be more than simply potentially useful. This consideration should be evidence-based and establish that the tool is both necessary and likely to be effective in achieving the specified purpose.
- Evaluate the validity and reliability of the generative AI tool for the intended purpose. [16] Tools must be accurate throughout the intended lifecycle of the tool and across the variety of circumstances in which they are used.

[15] For this joint statement, the term 'significant impact' is indicative as opposed to having a specific legal definition.

[16] For more information on validity and reliability in AI systems, see the NIST AI Risk Management Framework.

- Consider whether there are other more privacy-protective technologies that can be used to achieve the same purpose.

4. Openness

Be open and transparent about the collection, use and disclosure of personal information and the potential risks to individuals' privacy.

All parties should:

- Inform individuals what, how, when, and why personal information is collected, used or disclosed throughout any stage of the generative AI system's lifecycle (including development, training and operation) for which the party is responsible. This includes stating the appropriate purposes for these collections, uses and disclosures. Ensure that system outputs that could have a significant impact on an individual or group are meaningfully identified as being created by a generative AI tool.
- Ensure that all information communicated about a generative AI system is designed to be understandable by the intended audience, and made readily available both before, during and after use of the system.

Developers and providers of generative AI systems should:

- Inform organizations using, and any individuals interacting with, a generative AI system about both the primary purpose(s) and any secondary purpose, such as where personal information collected from prompts is used for further training or refining of an AI model.
- Ensure that organizations using a generative AI system are made aware of any known or likely risks associated with that system, including any known or reasonably expected failure cases (such as inputs or contexts in which the system may produce incorrect information, particularly if that system will foreseeably be used in a process to make decisions about individuals).
- Inform organizations using a generative AI system about any known policies and practices that could reasonably be used to mitigate identified privacy risks, where the developer or provider cannot implement those policies or practices themselves.
- Maintain and publish documentation about the datasets used to develop or train the generative AI tool, including the sources of the datasets, the legal authority for its collection and use, whether there are any licensing agreements or other restrictions on the acceptable uses of the datasets, and any modification, filtering or other curation practices applied to the datasets.

Organizations using generative AI systems should:

- Clearly communicate to any affected party whether a generative AI tool will be used as part of a decision-making process, and if so, in what capacity, with what safeguards, and what options or recourse are available to the affected party (particularly where a decision may have a significant impact on an individual). This explanation should also include a general description of the functioning of the system, how it is used to make a decision or take an action, and an overview of potential outcomes.

- Describe what, if any, personal information was used to re-train or refine the generative AI system for their specific use.
- Where a generative AI tool is public-facing, ensure that individuals interacting with the tool are aware that they are interacting with a generative AI tool and that they are informed about both privacy risks and any mitigations available to them (such as not entering personal information into a prompt, unless necessary).

5. Accountability

Establish accountability for compliance with privacy legislation and principles and make AI tools explainable.

All parties should:

- Recognize that they are responsible for compliance with privacy legislation, and should be able to demonstrate this compliance.
- Have a clearly defined internal governance structure for privacy compliance, including defined roles and responsibilities, policies and practices establishing clear expectations with respect to compliance with privacy obligations.
- Establish a mechanism by which the organization can receive and respond to privacy-related questions or complaints.
- Undertake assessments, such as Privacy Impact Assessments (PIAs) and/or Algorithmic Impact Assessments (AIAs), to identify and mitigate against potential or known impacts that the generative AI system (or proposed use thereof, as applicable) may have with respect to privacy and other fundamental rights.
- Regularly re-visit and re-evaluate accountability measures (including bias testing and assessments), given the evolving nature of both generative AI systems and AI regulation.

Developers and providers of generative AI systems should:

- Take appropriate steps to make the outputs from generative AI systems traceable and explainable. In summary, this includes a complete account of how the system works (traceability) and a rationale for how an output was arrived at. Where a developer or provider is of the opinion that outputs from a generative AI tool *are not* explainable, this should be made explicit to any organization using or individual interacting with the tool to allow them to determine whether the tool is appropriate for use for their intended purpose.

- If revealing a generative AI's training data would impact individuals' privacy, ensure that testing is done on the system's vulnerability to data extraction and other methods by which training data could be revealed to a third party.
- Undertake independent auditing to assess the validity and reliability of the system, confirm compliance with privacy legislation, test outputs for inaccuracies and biases, and recommend effective guardrail measures to mitigate potential risks. Developers and providers are also encouraged to allow independent researchers, data protection authorities, and other relevant oversight bodies to assess and audit their generative AI systems (or foundation model) for potential risks and impacts.

Organizations using generative AI systems should:

- Know that accountability for decisions rests with the organization, and not with any kind of automated system used to support the decision-making process.
- Ensure that impacted individuals are provided with an effective challenge mechanism for any administrative or otherwise significant decision made about them. This includes maintaining and providing on request sufficient information for that person to be able to understand how a decision was reached, and allowing them the opportunity to request human review and/or re-consideration of the decision.
- If the outputs of a generative AI system are not meaningfully explainable, consider whether the proposed use is appropriate.

6. Individual Access

Facilitate individuals' right to access their personal information by developing procedures that enable it to be meaningfully exercised.

All parties should:

- Ensure that procedures exist for individuals to access and correct any information collected about them during their use of the system.
- Develop processes to permit individuals to exercise their ability to access or correct personal information contained in an AI model, particularly where that information may be included in outputs generated in response to a prompt.

Organizations using generative AI systems should:

- Where a generative AI system is used as part of a decision-making process, maintain adequate records to allow for requests for access to information about that decision to be meaningfully fulfilled.

7. Limiting Collection, Use, and Disclosure

Limit the collection, use, and disclosure of personal information to only what is needed to fulfill the explicitly specified, appropriate identified purpose.

All parties should:

- Ensure that the collection and use of personal information for training AI tools is limited to what is necessary for the purpose, and use anonymized or de-identified data where possible. This can include the use of synthetic data.
- Avoid function creep, and only use personal information for purposes identified at the time of collection or (where permissible) for purposes that are consistent with the purpose for collection.
- Avoid indiscriminate collection of personal information based on assertions about the breadth of potential purposes for a generative AI system.
- Recognize that the public accessibility of data does not mean that it can be indiscriminately collected or used. Personal information that is accessible online is still subject to Canadian legislation or other regulatory instruments – even where that information is defined as being 'publicly available'.
- Establish and abide by appropriate retention schedules for personal information, including (as applicable) that contained within training data, system prompts, and outputs. These schedules should both (i) limit retention for information that is no longer required, and (ii) ensure that information is retained long enough for individuals to exercise their right to access (particularly where a decision has been made about them).

Developers and providers of generative AI systems should:

- Where possible and appropriate, use a filter or other process to remove personal information from data sets in advance of using them for training.
- Ensure that the outputs of AI products and services disclose only personal information that is necessary to achieve the request in the prompt.

Organizations using generative AI systems should:

- Ensure that any inferences created about individuals are for specified and disclosed purposes, and that their accuracy can be reasonably assessed and validated.
- Treat any inferences generated about an identifiable individual as personal information.
- Where possible and reasonable, use anonymized or de-identified information within prompts to a generative AI system rather than personal information.
- Where personal information (and, in particular, sensitive or confidential information) must be entered into a prompt, only do so where authorised.
- Unless otherwise required, prompts should not be retained, used for secondary purposes, or disclosed.

8. Accuracy

Personal information must be as accurate, complete, and up-to-date as is necessary for purposes for which it is to be used.

Developers and providers of generative AI systems should:

- Ensure that any personal information used to train their generative AI models is as accurate as necessary for the purposes. This may require a detailed consideration; for instance, the introduction of 'inaccuracies' by modifying a dataset to address a known bias (such as by enhancing it with synthetic data) may be preferable to use of the original 'accurate' dataset.
- Have a process by which a generative AI system can be updated (for instance, by refining or retraining the model) where it becomes known that the information on which it was trained is inaccurate or out-of-date.
- Inform organizations using generative AI about any known issues or limitations about the accuracy of generative AI outputs. This may include where the training dataset is time bounded (i.e. only contains information up to a certain date); where it is from a single, non-representative source; or where there are particular use cases or inputs that tend to lead to inaccurate outputs.

Organizations using generative AI systems should:

- Ensure that personal information is as accurate, complete and up-to-date as necessary for the purpose whenever it must be entered into a generative AI prompt or is used to train a bespoke generative AI model.
- Evaluate the impacts of any accuracy issues or limitations disclosed by the provider or developer of the generative AI system, such as time-bounded or single-source training data, on the use of the system. If this has not been disclosed and is not otherwise available, consider whether the use of the system remains appropriate and/or legally authorized.
- Take reasonable steps to ensure that any outputs from a generative AI tool are accurate as necessary for the purpose, especially if those outputs are used to make or assist in decisions about an individual or individuals, will be used in high-risk contexts, or will be released publicly.
- If the proposed use of a generative AI system relates to a specific group, take appropriate measures to ensure that that group is adequately and accurately represented in the system's training data.
- Be aware that issues regarding the accuracy of training data or outputs may make a generative AI system inappropriate for use (either in general or where such use could have significant impacts on an individual).

9. **Safeguards**
   Establish safeguards to protect personal information and mitigate potential privacy risks.

   All parties should:

   - Safeguard any personal information collected or used throughout the lifecycle of a generative AI tool with measures commensurate to the sensitivity of the information.
   - Maintain ongoing awareness of, and mitigations against, threats that are of particular concern when using generative AI, which include but are not limited to prompt injection attacks (in which carefully crafted prompts bypass filters or make the model perform unanticipated actions); model inversion attacks (in which personal information contained in the model's training data is exposed); and jailbreaking (in which privacy or security controls in the tool are overridden).

   Developers and providers of generative AI systems should:

   - Design products and services to prevent the inappropriate use of their tools and limit or prohibit the creation of illegal or harmful content. This includes safeguards and guardrails that prevent inappropriate uses that may lead to unfair, unethical, or discriminatory treatment, and threats to fundamental rights and freedoms.
   - Monitor for instances of the generative AI tool being used inappropriately and amend or correct systems to address these issues.

   Organizations using generative AI systems should:

   - Confirm that when using data under their control in the course of preparing, using, or deploying a generative AI system, the use of that data does not negatively impact model safeguards such as by creating or exacerbating biases, increasing the ability to undertake prompt injections, model inversions, or jailbreaks, or otherwise resulting in unauthorized parties being able to extract personal information in the course of using a generative AI system.

These guidelines are for information purposes only and do not constitute a decision or finding by the Office of the Information and Privacy Commissioner for British Columbia. These guidelines do not affect the powers, duties, or functions of the Information and Privacy Commissioner regarding any complaint, investigation, or other matter under FIPPA or PIPA.