METROLAB NETWORK

# GenAI for Local Governments: Cybersecurity and Privacy Concerns

A MetroLab Network White Paper

# GenAI for Local Governments: Cybersecurity and Privacy Concerns

METROLAB NETWORK

## Table of Contents


About This White Paper ..... 2

Explainer and Guidance ..... 3

Cybersecurity and Privacy: Key Policy Considerations ..... 4

I. Data ..... 4

II. GenAI Data Lifecycle ..... 5

GenAI Risks to Privacy ..... 8

GenAI Risks to Cybersecurity ..... 12

Cybersecurity Policy Requirements for GenAI Systems ..... 15

Privacy Policy Requirements for GenAI Systems ..... 17

# About This White Paper

MetroLab Network is a nonprofit in Washington DC that aims to equip local governments with science and research. It serves as a convener of an emerging academic practice focused on integrative, use-inspired, community-focused research, done in partnership with local government and communities.

The "In the Lab" program supports a national effort for practitioners, by practitioners, to produce policy guidance on emerging policy needs. In August 2023, when governments had barely begun the conversations on the concept and use of the "new AI" in the market, MetroLab Network launched the GenAI (Generative AI) for Local Governments Task Force as a new edition of its "In the Lab" program. It convened 130+ individuals to develop policy guidance that includes process recommendations, and a resources library. The task force comprised 45 unique local governments, 16 universities, 25 private sector companies, and 16 other stakeholder groups including four federal agencies, nonprofits, and coalition organizations.

As a result of this work, MetroLab is publishing three separate deliverables that together, represent a complete body of work. First, we are publishing a policy guide specific to community engagement. This guide includes ways to proactively engage communities to proactively shape GenAI policies such as community education, ways in which local governments can increase transparency, and more. Second, a call to the research community. At every turn, we asked what research is needed to further efforts on understanding the technology and its impact. We are providing an updated list of research questions as identified by the dozens of stakeholders involved with the task force. And finally, a white paper on cybersecurity and privacy. This whitepaper identifies the unique ways AI is impacting cybersecurity and privacy protections, for example, citing the increasing difficulty in enforcing consent and the right to be forgotten mechanisms.

***Intended Uses.*** These publications intended to be several things:

a. A useful tool for practitioners, co-developed by practitioners (with the guidance and input of expertise from individuals from academia and other organizations with relevant expertise).
b. A living guide housed on the MetroLab Network website, and curated, updated, and refined there through a multi-functional online platform to help local governments as they keep pace with rapidly evolving AI policy guidance and regulatory best practices.
c. A reminder of several legal considerations that permeate data governance—while a few lawyers were involved in this project, it is incumbent upon us to tell you that this is not legal advice.

This white paper has been authored by Nitisha Tripathi, the Generative AI Policy Fellow at MetroLab Network. It is a result of deliberations of stakeholder consultations organized through virtual meetings of a "Cybersecurity and Privacy Subcommittee" and primary qualitative research. Throughout the last year, the subcommittee met frequently to identify the scope of discussion, specific themes that need attention and risks that need to be addressed. Following initial discussions and research, draft outlines of this paper were prepared. Through a series of online meetings, extensive feedback was received leading to the development of the final text of the guide that was presented to the subcommittee, inviting edits and suggestions.

It must be noted that AI and GenAI are rapidly evolving technologies. As their nature of operation changes, so will the cybersecurity and privacy risks associated with them. There is a need to stay constantly vigilant about the growth of these technologies. We hope that this paper can serve as a foundation block for local governments to responsibly begin and continue their AI and GenAI adoption journey.

# Explainer and Guidance

*Privacy is about people; cybersecurity is about infrastructure. Privacy is about rights; cybersecurity is about protection.*

Historically, cybersecurity has been primarily concerned with protecting digital systems and infrastructure, while privacy has been focused on guarding people's data and information. However, the sophisticated evolution of technology has expanded the scope of understanding these terms exponentially. While it may be premature to claim that AI and Generative AI have fundamentally altered the conventional understanding of cybersecurity and privacy, NIST has provided the following definitions to assess their scope comprehensively:

- **Cybersecurity:**[1] "Prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication, and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation."

  The dominant focus of this definition is on fortifying the infrastructure.

[1] NIST AIRC. (n.d.). NIST AIRC - Glossary. NIST Trustworthy & Responsible AI Resource Center. Retrieved from https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

- **Privacy, Data and Model Privacy:**
  i. "Attacks against Machine Learning models to extract sensitive information about model and training data"[2]
  ii. "Freedom from intrusion into the private life or affairs of an individual"[3]
  iii. "Freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual"[4]

The dominant focus of these definitions is on safeguarding the private information of individuals.

# Cybersecurity and Privacy: Key Policy Considerations

From social security numbers to tax and voting records, local governments are a custodian of a range of sensitive citizen data. However, dependence on old or dated tech systems and infrastructure that lack sufficient security and privacy measures creates an imminent risk of a cyber-attack or data theft operation.

1. Has Generative AI modified the existing understanding of 'Data', 'Cybersecurity' and 'Privacy'?

## I. Data

Data is the foundation of all AI systems. While the fundamental understanding of 'data' hasn't materially changed, the scope of its governance and responsible use has expanded dramatically. The distinguishing feature of a Generative AI (GenAI) system from an AI system i.e. its ability to create or generate new content based on the data it is trained on, is reliant on collection of millions of publicly available data sources. Innovation of Generative AI has introduced modifications in the way data is understood and governed. Let's try to understand this from the perspective of the 'Data Lifecycle'.

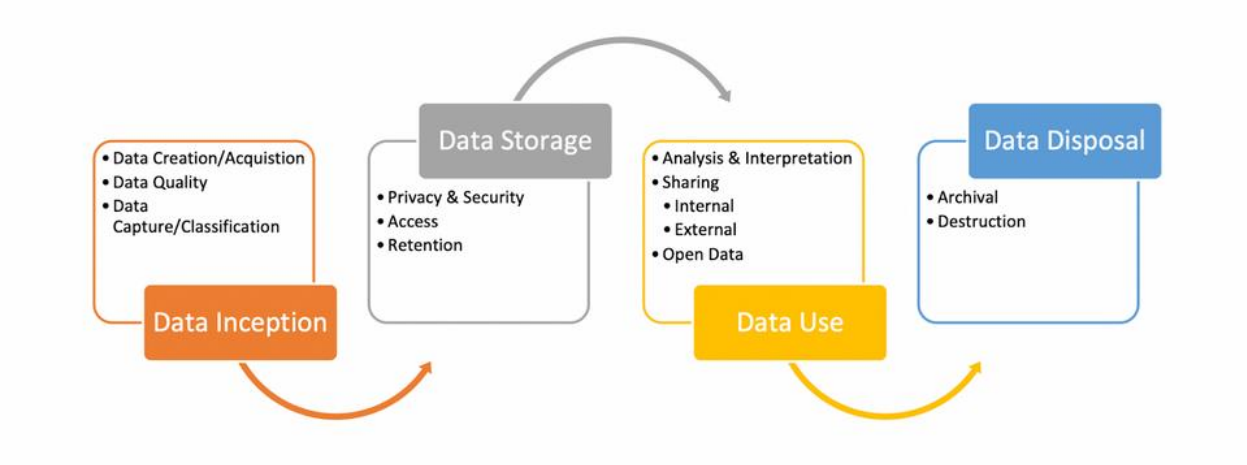
The Data Life Cycle, as envisioned by the MetroLab Data Governance Guide (Fig.1.), is an

[2] Bandy, J., & Vincent, N. (2021). Addressing Documentation Debt in Machine Learning Research: A Retrospective Datasheet for BookCorpus. arXiv:2105.05241. Retrieved from https://arxiv.org/pdf/2105.05241v1
[3] NIST AIRC. (n.d.). NIST AIRC - Glossary. NIST Trustworthy & Responsible AI Resource Center. Retrieved from https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary (Refer to ISO/IEC_TS_5723:2022(en) in the sheet)
[4] NIST AIRC. (n.d.). NIST AIRC - Glossary. NIST Trustworthy & Responsible AI Resource Center. Retrieved from https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary (Refer to aime_measurement_2022, citing ISO/IEC TR 24029-1)

iterative process that encapsulates the "what" of operationalizing data governance. Keeping the spirit of 'iteration' intact, through the 'Data Lifecycle of a GenAI system', we intend to add new aspects to this discussion. It is important to note that GenAI development Lifecycle differs from the existing AI development lifecycle due to the addition of fine-tuning services.

To make the distinction between GenAI, foundation models and fine-tuning clear, while GenAI comprises any AI system that can generate content, foundation models[5] are often large language models or pre-trained models that are trained on massive datasets and fine-tuning is a technique through which these models are trained for specific tasks.

*Figure 1: Data Lifecycle*




## II. GenAI Data Lifecycle

**Stage 1: Data Inception: How are you collecting, validating and preparing your data?**

1. Data composition and collection
2. Data Validation
3. Data Preparation (Quality): Cleaning, Standardizing, Labeling

***Key questions to consider[6]:***

[5] Stanford University defined a foundation model and fine-tuning as "A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks": Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258. Retrieved from http://arxiv.org/pdf/2108.07258; According to Section 3(k) of the "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", the term "*dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters*..."

[6] Bandy, J., & Vincent, N. (2021). Addressing Documentation Debt in Machine Learning Research: A Retrospective Datasheet for BookCorpus. arXiv:2105.05241. Retrieved from https://arxiv.org/pdf/2105.05241v1; It is important to note that this research was published in 2021. Since then, a range of commercial LLMs and GenAI systems have been released in the

☐ What types of data does the AI system represent?
☐ How many different types of data have been included?
☐ Have all possible types of data been considered or is this a sample of the data considered? What are the details of the sample data (size, type)?
☐ Is each data labeled?[7]
☐ Any information is missing from the data sources?
☐ Is there any explicit connection between the data sources? If yes, has it been clearly stated?
☐ Are there any errors in the data?
☐ How did the data collection process take place?
☐ Who collected the data to train the system?
☐ How long did it take for the data to be collected?
☐ Did the data undergo cleaning? If yes, how? Which software was used?
☐ Is the data being enriched or joined with other datasets?

**Stage 2: Data Storage: What mechanisms are critical to storing data responsibly?**

1. Privacy and Security of data stored as parameters
2. Access
3. Retention
4. Deletion

***Key questions to consider[8]:***

☐ Any confidential data included in the training dataset? Any sensitive personal data?
☐ Can a direct or indirect identification of individuals take place from the data?
☐ Has it been considered whether the data is representative of the demography? If so, give details about the extent of its representation.
☐ Was the consent of the authors of data sources obtained? Can the AI system revoke consent if needed?
☐ Was the impact on different communities assessed and considered?

**Stage 3: Data Use: How is the data being used to develop the AI system?**

***Data Experimentation***

1. Model Training on Data: Training, re-training and iterative fine-tuning of model

market including ChatGPT, creating a new set of questions to consider. However, in our opinion, despite the lapse in time, these questions can serve as a robust foundation to develop an understanding of the idea "GenAI Data Lifecycle."

[7] GenAI systems, like large language models, can often learn from data without needing humans to label or categorize it first. This is called "unsupervised learning." It's different from "supervised learning," where the AI is trained on data that humans have carefully labeled. Many AI systems that process documents don't need the information to be neatly organized. Some AI tools can even automatically sort through messy data on their own. This matters because it affects how we prepare data for AI and how we think about potential biases in these systems.

[8] Bandy, J., & Vincent, N. (2021). Addressing Documentation Debt in Machine Learning Research: A Retrospective Datasheet for BookCorpus. arXiv:2105.05241. Retrieved from https://arxiv.org/pdf/2105.05241v1

2. Model Evaluation: Testing the data for its performance, bias

***Data Monitoring***

1. Model Deployment: Ingesting new/live data entered by the user.
2. Model Monitoring: Ensuring the data processing to generate new content does not result in incorrect or inaccurate results.

***Key questions to consider[9]:***

☐ What is the specific task/set of tasks for which this data is being used?
☐ Have the sources of data been included in a central repository? Have the errors in the data been published?
☐ What are some of the other possible uses of this data? Include the prohibited uses, if any.
☐ Who is responsible for supporting or maintaining the system?
☐ Will the system be updated? By whom?
☐ How can anyone else build on or improve the system's capabilities?
☐ Do any IP/licensing/terms of use apply? If yes, what?
☐ Are there any export/regulatory restrictions? If yes, what?

**Stage 4: Data Disposal: How are you retiring the AI system?**

1. Archival
2. Data Portability
3. Destruction

***Key questions to consider:***

☐ When migrating AI systems between vendors, consider the following:
(1) transferability of training data and outputs,
(2) compatibility of model weights[10] across different frameworks, and
(3) protocols for securely retiring the old system while preserving necessary information.

2. Are there any new Generative AI risks to privacy and cybersecurity that are expected to inform the solutions? What are these new risks?

[9] Ibid.
[10] Section 3(u) of the Executive Order 14110 defines model weights as *"a numerical parameter within an AI model that helps determine the model's outputs in response to inputs."* In simpler terms, these are the numbers that are used by an AI system to calculate and respond to the questions asked in the form of prompts.

AI risks can differ from or magnify traditional software risks. Likewise, Generative AI can increase existing AI risks and create unique risks as well.

Following President Biden's Executive Order 14110, NIST released a companion document for its previously released AI Risk Management Framework. It particularly highlights the new risks posed by the most recent version of AI in the market: Generative AI. While the document comprehensively assesses the GenAI realm to identify new risks that stakeholders should be aware of, our objective is to draw the attention of local governments to some of the most critical policy considerations or risks that they should be aware of, particularly as this technology continues to evolve with every passing day.

# GenAI Risks to Privacy

Privacy is an exercise of how an organization can efficiently control, protect and share data. The following are some of the Generative AI-induced new and significant risks to privacy of an AI system:

1. **Lack of transparency into GenAI models' training data sources and their explainability:**

   ***"Transparency in AI allows us to understand how the system works, which is essential for building trust"***[11]

   Understanding how the system works also includes the 'what' and 'why' of the working of the system. Imagine opening an AI system like a box and seeing several lights flashing but not being able to figure out what those lights mean and why they are flashing[12]. So, transparency is critical but so is explainability. When this clarity is given to the users, it enables them to prevent issuing consent for using their data for scenarios, individuals or organizations that they are not comfortable sharing their data with. It gives the power back to the average user, balancing the scale of innovation with responsibility.

2. **Re-identification of sensitive data:** Modern AI like chatbots and language models are incredibly powerful because they learn from huge amounts of data from the

[11] Observed in the MetroLab Cybersecurity and Privacy Sub-committee Meetings.
[12] Josh Batson, a researcher at the A.I. startup Anthropic shared this analogy in an interview with Kevin Roose and Casey Newton.

internet and other sources. However, this training data often includes people's personal information that is available online. To protect the privacy of such data, anonymization is a technique used to remove any direct link between an individual and their personal data. This is supposed to eliminate the risk of re-identifying such data when used by others. While there is already an existing concern around it being impossible to anonymize large excerpts of data,[13]Generative AI has added another complication to this issue.

The new risk is that the AI system could potentially re-identify the anonymized data and expose this private data when answering questions asked by subsequent users. This could happen due to the AI unintentionally keeping the data in its memory, making it an easy target for cyber-attacks or data theft, resulting in damaging consequences.

Key examples of this risk include:

- **Disclosing Personal Details and Sensitive Information:** The AI may directly release data like someone's name, address, contact information or identification numbers that was contained in its training data. When discussing certain topics, the AI could reproduce extremely sensitive private information it learned, like personal conversations, medical history, or financial details. Leakage of this kind of information can enable crimes like blackmail, identity theft, financial fraud, stalking or harassment against that individual.
- **Easy Identification/Re-identification of Individuals:** Bad actors may analyze the AI's responses to figure out if a specific person's data was used to train the system, violating public expectations and trust regarding responsible use of personal data.
- **Use of data without Consent:** People need to know and give their consent to how and why their data is being used. The data used to train AI systems often includes personal information collected from websites, databases without getting explicit permissions to use. If such data is revealed, it can invite damaging attention for the concerned individual.
- **Direct and Indirect Identification via Combination of Information[14]:** Even if the direct link between text and its original source is removed, the data itself may still exist in other publicly accessible training data sources.

[13] Rogers, A., Baldwin, T., & Leins, K. (2021). 'Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. arXiv preprint arXiv:2109.06598. Retrieved from https://arxiv.org/pdf/2109.06598
[14] Ibid.

By combining this with existing metadata[15], even small pieces of text could be easily re-identified by an AI system. Alternatively, re-identification of the data could occur through other information an individual has published or shared online, as AI systems have the capability to integrate available information and learn from it.

3. **The Other Side of Privacy Checks:** A major challenge in protecting privacy with powerful AI systems is how difficult it can be to identify and filter out the private personal information that the AI has already memorized from its training data.

- AI can analyze private records by combining scattered personal details from across its training data sources.
- If privacy filters are too aggressive, they risk removing or altering non-sensitive content, negatively impacting the AI's performance.
- Upgradation, fine-tuning or refinement of data can also change the nature of privacy risks involved.

4. **Difficulty in enforcing consent and right to be forgotten mechanisms:**

- **Inability to Revoke Consent:** Once personal data gets absorbed into an AI's training dataset, it becomes extremely difficult, if not impossible, for individuals to revoke any prior consent they may have given for use of their data. There are no clear mechanisms to identify and remove specific individuals' data from these massive training datasets on requests to revoke the consent.
- **Difficulty Enforcing "Right to Be Forgotten:"** Several data privacy laws include a "right to be forgotten" and have personal data deleted on request. However, once ingested by large AI models, this personal data gets decentralized and disseminated in ways that make deleting or forgetting the original data sources technically infeasible.

5. **Negative impact of using synthetic data:** Synthetic data is a kind of ‘fake’ data that imitates real-world data while hiding the personal information. Training AI models with such data can create its own set of risks that can deteriorate trust and reduce the accuracy of the data used. Following are some of the pitfalls:

- **Inaccuracies:** Since the data is artificially generated, it may not fully capture all the complexities of real-world data. This could lead to inaccuracies in the performance of the AI model.
- **Re-identification Concern:** Depending on how it's generated, it may still be possible in some cases for synthetic data to get reverse-engineered or decoded to

[15] This includes any additional information about a piece of data. For example, if an email address is available on the internet, the metadata would include the name of the sender, time when it was sent etc.

reveal connections back to the real individuals it originates from, threatening the exposé of personal data.

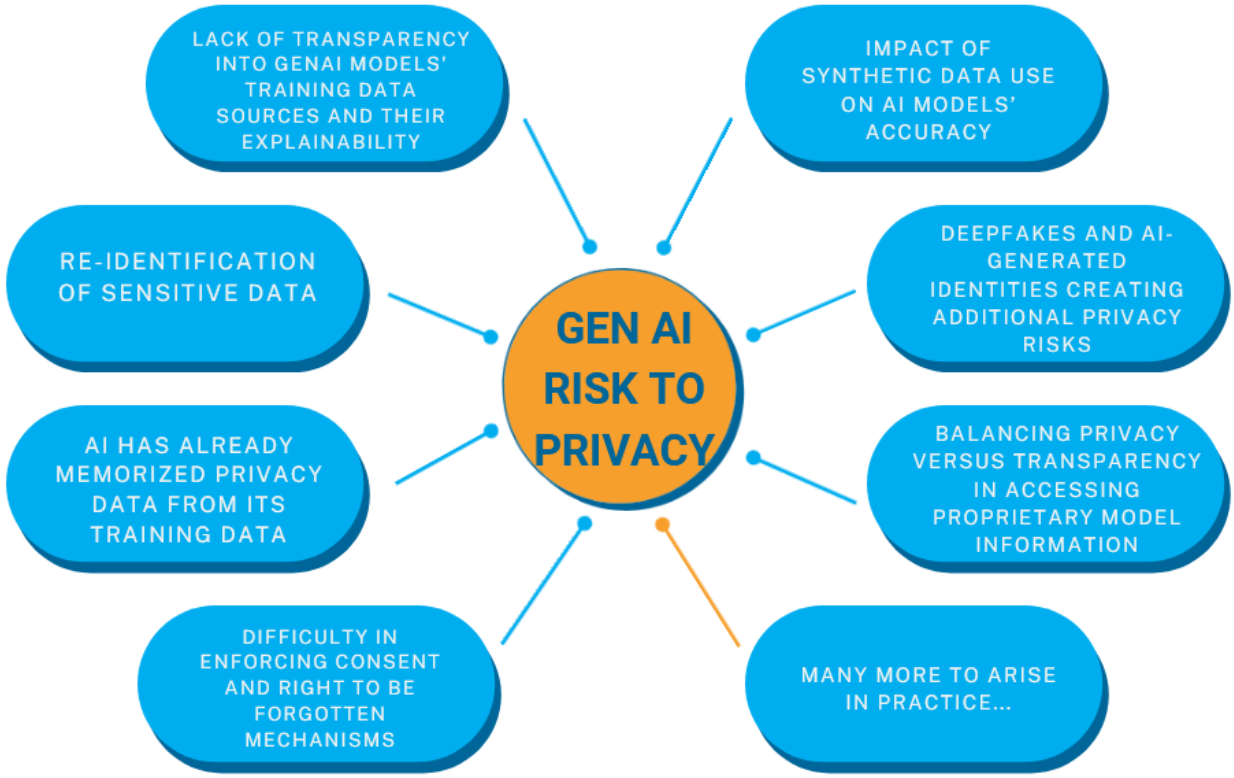
- **Public Distrust:** If it becomes known that an AI assistant or other system is trained on synthetic versus real data, some people may not find it to be trustworthy and doubt its credibility.
    - While using synthetic data can help protect individual privacy during training, there is also a risk of making AI systems seem untrustworthy or inaccurate. Organizations should carefully weigh this tradeoff.

6. **The 'Deepfake' Risk:** Deepfakes are synthetic media that can be created by AI. These include doctored but realistic videos, images or audio content where individuals can be made to say words or do actions that were never said or done. In the past few years, the rampant use of this kind of synthetic media by bad actors has emerged as a legitimate threat to privacy.

- **Use of personal data without permission:** Bad actors can misuse individuals' personal data, such as biometric details including facial features/expressions, voice etc. Such data can be easily used to create deep fakes even without obtaining the consent of the concerned individuals.
- **Social engineering attacks/operations:** Impersonation is a classic use of deep fakes that can result in severe privacy violations. This tactic can result in a kind of attack that can easily deceive unsuspecting people and make them reveal personal details that they would not disclose otherwise.
- **Difficulty in proving privacy violations:** Due to the synthetic nature of deep fakes, it has been found to be difficult to prove it in a court of law, particularly as a privacy violation[16].
- **Threat to Elections[17]:** Deepfakes can be in either of these forms: audio, video or imagery. Comparatively, audio is easier to create and target for manipulative agendas. In such a scenario, fake audio calls to local election officials can serve as a legitimate threat to an election process. Similarly, circulation of fabricated videos can exacerbate this threat further.

[16] Reuters. (n.d.). Manipulating Reality: The Intersection of Deep Fakes and the Law. Retrieved from https://www.reuters.com/legal/legalindustry/manipulating-reality-intersection-deepfakes-law-2024-02-01/
[17] Brennan Center for Justice. (n.d.). The Danger of Deep Fakes to Democracy. Retrieved from https://www.brennancenter.org/our-work/analysis-opinion/danger-deepfakes-democracy

*Figure 2: Gen AI Risks to Privacy*




# GenAI Risks to Cybersecurity

This is an exercise of how well an organization can implement measures and protocols to protect its devices, networks and information from manipulation by bad actors. There are two kinds of cybersecurity risks that need to be considered here:

1. Internal: The risk we face due to internal use of GenAI (daily operations), and;
2. External: The risk we face from external cybercriminals using GenAI, that includes imparting education about the kind of risks created by GenAI and bolstering existing defensive activities rather than supporting new initiatives.

The following are some of the risks that have been uniquely created by GenAI:

1. **Risks of data poisoning attacks compromising integrity of AI models[18]:** This is a novel risk created by GenAI systems that has the power to poison the dataset on which a system is trained through injection of corrupted data, modification or

[18] CrowdStrike. (n.d.). What Is Data Poisoning?. Retrieved from https://www.crowdstrike.com/cybersecurity-101/cyberattacks/data-poisoning/

deletion of existing data to derail its intended behavior and force it to act maliciously. Let us take a closer look at some of the different types of such attacks:

- Membership Inference Attacks: In this kind of attack, the bad actor makes attempts to assess if a particular data point exists in the training data. For example, trying to find the personal details of a specific individual.
- Model Extraction: For such an attack, the bad actor analyzes the responses given by the AI system to recreate it or steal it to achieve their agenda. For example, employees of an organization trying to analyze the responses to recreate the objectives of an AI systems' behavior.

2. **Hallucinations[19]:** GenAI models/systems have been found to exhibit a unique risk - the ability to provide responses that are fabricated, factually incorrect, or unrelated to the original query. This phenomenon is being termed as 'Hallucinations'. The internet, a prominent source for AI training datasets, is riddled with biased, fabricated, or outdated data. Furthermore, unlike developing AI models for specific tasks, most modern models are designed for multiple purposes involving analysis of various languages and texts, increasing the likelihood of hallucinated responses. Let's explore three types of hallucinations:

- **Intrinsic Hallucinations (Input-Conflicting):** When the AI system's response differs from the information provided in the input data. For instance, if an AI assistant is asked, "What is the city of Washington's cybersecurity policy?" and the answer is, "Here is some information on the City of Washington's workplace harassment policy." This indicates a missing connection between the question and the response.
- **Extrinsic Hallucinations (Context and Fact- Conflicting):**
    a. **Context-Conflicting:** When the AI system generates responses that contradict information it provided earlier. For example, if the AI system answers the above question by saying, "Here is some information on the City of Washington's workplace harassment policy. I hope the measures adopted in the AI procurement policy can provide you effective guidance." The mention of "procurement policy" directly contradicts the previously stated policy.
    b. **Fact-Conflicting:** When the AI system provides factually incorrect responses. For instance, if the AI system answers the above question by stating, "The city of Washington does not have any cybersecurity policy," this would be inaccurate if such a policy exists. Interestingly, so far, fact-related hallucinations have been witnessed with a higher frequency as

[19] Zhang, Y., Li, Y., Fu, T., & Zhang, Y. (2023). Siren's Song in the AI Ocean: A Survey of Hallucination in Large Language Models. arXiv:2309.01219. Retrieved from https://arxiv.org/pdf/2309.01219

opposed to others due to their emphasis on the need for authoritative sources of information.

Example:

**AI Package Hallucination:** This is an example of extrinsic hallucination that makes it easy for attackers to develop malicious code that looks legitimate. They can generate fake versions of popular code libraries using AI. Then, they can distribute these malicious packages openly without trying to hide their true malicious nature. The AI-generated code gives the illusion of being safe and authentic. This allows cyber criminals to bypass security checks and distribute attack code more easily by taking advantage of generative AI's ability to produce convincing but fake software components. Users may inadvertently install these malicious packages, compromising their systems and data.

3. **Easy availability of powerful AI tools lowering barriers for cybercriminals[20]:** With the rampant development of AI systems/tools, it is becoming increasingly easy for bad actors to gain access to such tools and use them to explore "new and hard to forecast" cyber risks and offensive capabilities. Key examples of this risk are:
    - **Discovery of vulnerabilities in hardware, software, data of AI systems:** GenAI systems have the capability to find new weaknesses or flaws in a computer system/digital product and use it to inflict damage. It can write new code to achieve this and use cyber tactics such as leaving bugs in the system to enable future exploitation.
    - **Create phishing mails at scale in lesser time:** GenAI systems have the capacity and ability to create deceptive information with ease and at scale. They can personalize the writing style and communication patterns to launch phishing attacks in a manner that can easily escape detection. More importantly, as opposed to the manual and time-taking processes employed earlier to send phishing emails or release such campaigns, GenAI systems have automated this process, simplifying the process further.
    - **Prompt-Injection[21]:** This is another example of how a GenAI system can manipulate other AI systems to behave in a way it is not supposed to. Within this kind of attack, a bad actor can remotely give an AI system a set of

[20] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... & Dafoe, A. (2023). Model Evaluation for Extreme Risks. arXiv preprint arXiv:2305.15324. Retrieved from https://arxiv.org/pdf/2305.15324

[21] Greshake, K., Endres, C., Abdelnabi, S., Holz, T., Mishra, S., & Fritz, M. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv preprint arXiv:2302.12173. Retrieved from https://arxiv.org/pdf/2302.12173

carefully designed instructions through a prompt. Once the AI system processes the input data given in the form of the prompt, unintended and damaging behavior is shown by the system, making it a threat to the security structure. These kinds of prompt-injection attacks can also easily leak the data, making it a risk to privacy as well. Such prompts can also act like malwares and spread the harmful behavior to other users through emails or other content.

4. **Weaponization of GenAI to generate misinformation, disinformation or mal-information at scale[22]:** Let's understand the difference between these concepts. Misinformation includes the dissemination of content without any harmful intention, disinformation includes the sharing of information with the intent to harm and mal-information comprises sharing of genuine information to cause harm, particularly by releasing information that was expected to remain private.

   The advent of Generative AI (GenAI) has streamlined and made creating such information economically viable, resulting in heightened threats of fraud and manipulation of public opinion via social media platforms, potentially influencing critical events such as elections. Furthermore, the increased sophistication has facilitated the creation of targeted content tailored to specific communities, rendering local governments' law and order situations more vulnerable.

# Cybersecurity Policy Requirements for GenAI Systems

1. **Clearly define 'cybersecurity' for your organization, accommodating the shared concerns of key stakeholders:**
   a. It should be a comprehensive, modern definition of cybersecurity that incorporates potential threats arising from the use of Generative AI systems.
   b. This step will ensure that different public agencies within the local government can interpret cybersecurity in a clear and standardized manner.
   c. This will avoid any confusion for the agencies, and they will refer to the concerned policies and technologies with confidence.

2. **Develop and establish a Cybersecurity Incident Response Plan:**

[22] Communications Security Establishment Canada. (2022, February 23). How to Identify Misinformation, Disinformation, and Malinformation (ITSAP.00.300). Canadian Centre for Cyber Security. Retrieved from https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300

a. Formulate a step-by-step process to identify, report and respond to incidents impacting the cybersecurity posture of your organization. For clarity, incidents refer to events that have harmed or have the potential to compromise the security of GenAI systems. These can include hacking operations or other kinds of cyberattacks.
b. For rapid and efficient response to an incident, clearly define the responsibilities of each team member to ensure there is clarity in terms of who is expected to do what.

3. **Train and Educate Your Staff:** Train all staff involved with GenAI systems regularly on how to handle cybersecurity issues. This keeps everyone prepared and aware of potential threats. The State of California has developed a comprehensive GenAI toolkit that provides tailored training modules based on staff levels, accounting for legal, privacy, and security considerations, as well as a range of general topics. Following this model, developing a specialized module addressing the question "How to respond to a cybersecurity incident" could prove to be helpful for local government employees.

4. **Perform Risk Assessments:**
   a. Identify Risks: Conduct thorough risk assessments at least once a year. These assessments look for potential cybersecurity and privacy risks, both within and outside the organization.
   b. Evaluate Impact: Assess the impact and likelihood of these risks. This includes considering how severe the damage could be, how long it might last, and how widespread it could become.
   c. Conduct Red-teaming Exercises: Prior to deploying a GenAI tool, simulate common cyber threat attacks on a tool to stress test the technical and organizational safeguards put in place for a GenAI tool.

5. **Develop GenAI Procurement Guidelines:**
   a. Set Requirements for Vendors: Establish clear cybersecurity and privacy criteria that vendors must meet when providing GenAI systems.
   b. Vendor Accountability: Require vendors to provide detailed information on their cybersecurity practices and how they handle data.
   c. Due Diligence: Thoroughly check and assess the risks before purchasing GenAI systems to ensure they comply with security standards.

6. **Issue Transparency Reports[23]:**

[23] Knight First Amendment Institute. (2023, June 26). Generative AI Companies Must Publish Transparency Reports. Retrieved from https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports

- Define Harmful Content and its detection process: Explain what constitutes harmful content, preferably through examples and lay down the process that is used to detect such content.
- State the Frequency of Identification of Harmful Content: Within the period of assessment and reporting, state how often the content was found.
- Describe the Enforcement Mechanism: Identify if there has been a violation of Terms of Service of the AI system and accordingly give details about the enforcement process along with an analysis of its impact.
- Provide Details of the Mitigation Strategy: Include information about the steps taken to avoid violation of safety requirements.

---

# Privacy Policy Requirements for GenAI Systems

***We highly recommend you see Section 2 in our Data Governance guide for a comprehensive list of recommendations related to privacy and strengthening your data governance processes.***

1. **Establish and Adopt Privacy Principles:**
   In accordance with the AI Bill of Rights:
   a. Get informed consent from individuals when using their personal data with GenAI systems.
   b. Offer human alternatives and allow people to challenge decisions made by the AI. This ensures transparency and fairness.
   c. Offer a justification for the use and collection of personal and sensitive information.

2. **Clearly define or adopt the definition of 'informed' and 'revoked' consent by providing a checklist:**
   a. For Informed Consent -
      i. Proactively notify people that their data may get used to train AI systems and provide following information to clarify how it will be used:
         (i) The nature of the task/operation
         (ii) Information about use of Generative AI
         (iii) Purpose of the use and its potential implications
      ii. Explain the privacy risks.
      iii. Get explicit permission from them before their data can be included.
   b. For Revoked Consent -

i. Clearly communicate that people can revoke any prior consent to have their data used in AI training.
ii. Explain the process to remove or exclude their specific data from future AI training datasets and models.

3. **Combat Re-identification Concerns:**
    a. Implement robust de-identification and anonymization techniques for sensitive data used in training generative AI models.
    b. Develop safety checks on GenAI tool outputs that can scan and filter out personal identifiable information before it reaches end users.
    c. Guard against reverse engineering attempts that could re-identify training data. Reverse engineering includes unpacking the functioning of an AI system to recreate it and this may create the risk of re-identifying anonymized data.

4. **Implement Consent Mechanisms:**
    a. Obtain informed consent from individuals for using their data to train generative AI models, with ability to revoke consent.
    b. Provide transparency into the training datasets and processes used for developing generative AI systems, to the extent possible without compromising intellectual property protections.

5. **Data Rights and Regulatory Compliance:**
    a. According to available resources, explore the right to be forgotten to model training data and weights, with appropriate techniques for data deletion from models.
    b. Develop processes to fulfill other data rights like access and correction for generative AI systems.
    c. If generative AI uses biometric data, ensure strict compliance with biometric data protection laws.

6. **Synthetic Data:**
    a. When using synthetic data for training, check the output for accuracy, especially on sensitive topics that require human oversight.

7. **Deepfakes and Identity Fraud[24]:**

[24] In October 2022, the US Department of Homeland Security released a guide prescribing deep fake mitigation measures. It includes a list of best practices that should be followed by an organization along with a checklist to assess the 'Deep Fake Preparation' of an organization: Brooks, T., Daniel, C. P., Jesse, H., Kim, S., Maureen, R., Sahin, B., James, S., Oliver, T., & Federal Bureau of Investigation. (2021). Increasing Threats of Deep Fake Identities – Phase 2: Mitigation Measures.

a. Have clear guidelines distinguishing privacy violations from cybersecurity issues in cases like AI-generated identity fraud.
b. Establish technical and policy safeguards to prevent misuse of generative AI for creating deep fakes:
   - ☐ Define acceptable behavior boundaries
   - ☐ Clarify what counts as manipulated media and its various forms
   - ☐ Establish consequences for policy violations
   - ☐ Use dedicated content review teams
   - ☐ Collaborate with external experts
   - ☐ Verify the original source of content
   - ☐ Require documentation showing the content's origins and any changes made
   - ☐ Publicly promote trusted, authoritative sources
   - ☐ Teach staff and the public about media literacy
   - ☐ Build partnerships with universities, non-profits, and industry stakeholders

---

*About MetroLab: MetroLab Network is a nonprofit in Washington DC that aims to equip local governments with science and research. It serves as a convener of an emerging academic practice focused on integrative, use-inspired, community-focused research, done in partnership with local government and communities. We cultivate partnerships between universities and local governments to drive research-informed, evidence-based policy and enable data and technology transformation; we foster a peer network of stakeholders from academia and local government that constitute an applied, interdisciplinary field of research and practice; and we connect to an ecosystem of federal, philanthropic, and civic partners with a shared interest in the promise of civic research and innovation.*

Department of Homeland Security. Retrieved from https://www.dhs.gov/sites/default/files/2022-10/AEP%20DeepFake%20PHASE2%20FINAL%20corrected20221006.pdf

# Thank you to the MetroLab Local Government for GenAI Task Force, and the Cybersecurity subcommittee for their time, attention, and expertise.

## Acknowledgements

We especially want to thank Nitisha Tripathi, MetroLab Innovation Fellow, for her work and contribution to the AI Task Force. Nitisha brought organization and expertise to this effort, and this document would not be possible without her contribution.

**MetroLab Network**
1701 Rhode Island Ave, NW
3rd Floor
Washington, DC 20036
www.metrolabnetwork.org
info@metrolabnetwork.org