Response by the Centre for Information Policy Leadership (CIPL) to the National Institute of Standards and Technology (NIST)'s Request for Comment on the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Submitted May 31, 2024

The Centre for Information Policy Leadership (CIPL) welcomes the opportunity to respond to the National Institute of Standards and Technology (NIST)'s *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,* which aims to assist organizations in identifying, managing, and mitigating risks associated with the development and deployment of generative AI. For over 20 years, CIPL has been a key thought leader and advocate for organizational accountability and a risk-based approach to smart regulation, responsible governance and use of data, and accountable development and deployment of AI. Notably, our recent report, *Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework*[1], identifies best practices and shares case studies of how leading organizations are responsibly developing and deploying AI through the lens of CIPL's Accountability Framework. In addition, CIPL's *Ten Recommendations for Global AI Regulation*[2] proposes a three-tiered approach to AI regulation that would protect fundamental human rights and minimize the potential risks of harm to both individuals and society, while enabling beneficial development and use of AI.

CIPL commends NIST's efforts to build upon its AI Risk Management Framework (AI RMF) and President Biden's landmark Executive Order 14110 on Safe, Secure, and Trustworthy Artificial Intelligence and supports NIST's sociotechnical methodology in categorizing generative AI (GAI) risks. The rapid advancement, democratization, and widespread use of GAI has sparked numerous questions regarding existing risks that are common across AI technologies (e.g., data privacy, information security, human-AI configuration), as well as novel ones that require greater considerations (e.g., intellectual property and confabulation). CIPL agrees that a holistic and comprehensive approach is necessary to appropriately manage and mitigate risks raised by GAI. CIPL's Accountability Framework follows a similar, systematic approach that has enabled organizations to build comprehensive accountability programs that implement relevant legal requirements and standards, as well as internal corporate or ethics principles. CIPL's Accountability Framework contains seven core elements of accountability: leadership and oversight; risk assessment; policies and procedures; transparency; training and awareness; monitoring and verification; and response and enforcement (Figure 1). The Framework has been used by numerous organizations to effectively build and implement comprehensive privacy and data governance programs that enable not only compliance with

[1] CIPL, "Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework", February 2024, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf

[2] CIPL, "10 Recommendations for Global AI Regulation", October 2023, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ten_recommendations_global_ai_regulation_oct2023.pdf

applicable legal requirements, but also business operations that are grounded in accountability and foster customer trust.

Figure 1 - CIPL Accountability Framework (Source: CIPL)

CIPL has long encouraged the adoption of organizational accountability principles to effectively manage and regulate the development and deployment of AI technologies.[3] This approach promotes the implementation of protective measures that are proportionate to the likelihood and severity of the risks of harm in the context of development and deployment while enabling the technology's benefits. Furthermore, a programmatic, risk-based approach—as opposed to an exclusive focus on evaluating risks associated with individual technologies —allows organizations to systematically, holistically, and comprehensively assess the risks and benefits of AI products, projects, and deployments on their own and in interaction with each other, or mitigate any identified risks, while still enabling adaptation over time based on continuous assessment of risks to individuals, society, and the organization. Such an approach also provides organizations with the flexibility to calibrate their programs to new regulatory requirements, changes in risk profile, and developments in technology.

Feedback on the Risk List: In the Profile, NIST provided a set of risks that are unique to or exacerbated by GAI and can arise across the entire AI lifecycle (See Appendix A for this set of risks). NIST has requested feedback on whether these identified risks should be further categorized into technical / model risks, risks associated with misuse by humans, or ecosystem / societal risks.

---

[3] For example, please see the following papers: CIPL, "How the GDPR Regulates AI", March 2020, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf; CIPL, "How the GDPR Regulates AI", March 2020, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf; CIPL, "Artificial Intelligence and Data Protection in Tension", October 2018, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_first_ai_report_-_ai_and_data_protection_in_tension__2_.pdf

Further categorizing the identified risks

CIPL supports NIST's proposal to further sort the 12 risks identified in the Profile between technical / model risks, misuse by humans, and ecosystem / societal risks. We believe that organizing the risks in this manner can help organizations more clearly identify actionable steps they can take to mitigate each type of risk as each broader category will require its own, unique mitigations throughout the AI life cycle. For example, a model risk may require greater technical support and mitigations closer to the development stage of the model, while a risk of misuse by humans may require greater human oversight at the deployment stage. CIPL also encourages NIST to ensure that these risks are aligned to those presented in the NIST AI RMF as best as possible to foster interoperability.

According to the identified risks and categories, we propose the following general breakdown:

Technical / Model risks	**Misuse by humans**	**Ecosystem / societal risks**
• Confabulation • Dangerous or Violent Recommendations • Data Privacy • Value Chain and Component Integration	• CBRN Information • Human-AI Configuration • Obscene, Degrading, and/or Abusive Content	• Environmental • Information Integrity • Information Security • Intellectual Property • Toxicity, Bias, and Homogenization

It is important to emphasize that the technical / model risks and those associated with human misuse can also create ecosystem / societal risks. For example, the risks of GAI models leaking personal data about individuals, releasing CBRN information, or making dangerous or violent recommendations pose a grave risk to society and can be utilized by nefarious actors to do broad harm. Furthermore, risks can manifest across multiple categories. For example, data privacy risks can result from technical design flaws, but can also arise due to deliberate, human misuse of otherwise technologically sound systems. And privacy harms can have impacts at a societal level, e.g. if they become so widespread as to weaken trust in certain technologies across society.

Additional Feedback for Section 2.4, Data Privacy

CIPL offers additional feedback for Section 2.4 (Data Privacy) that we believe is integral to addressing data privacy as a GAI risk, drawing upon our previous work on the applying the GDPR to AI, identifying tensions in applying data protection principles to AI, and developing practical solutions to resolve the tensions (Figure 2).[4]

[4] See footnote 3 above for examples of CIPL's relevant works.

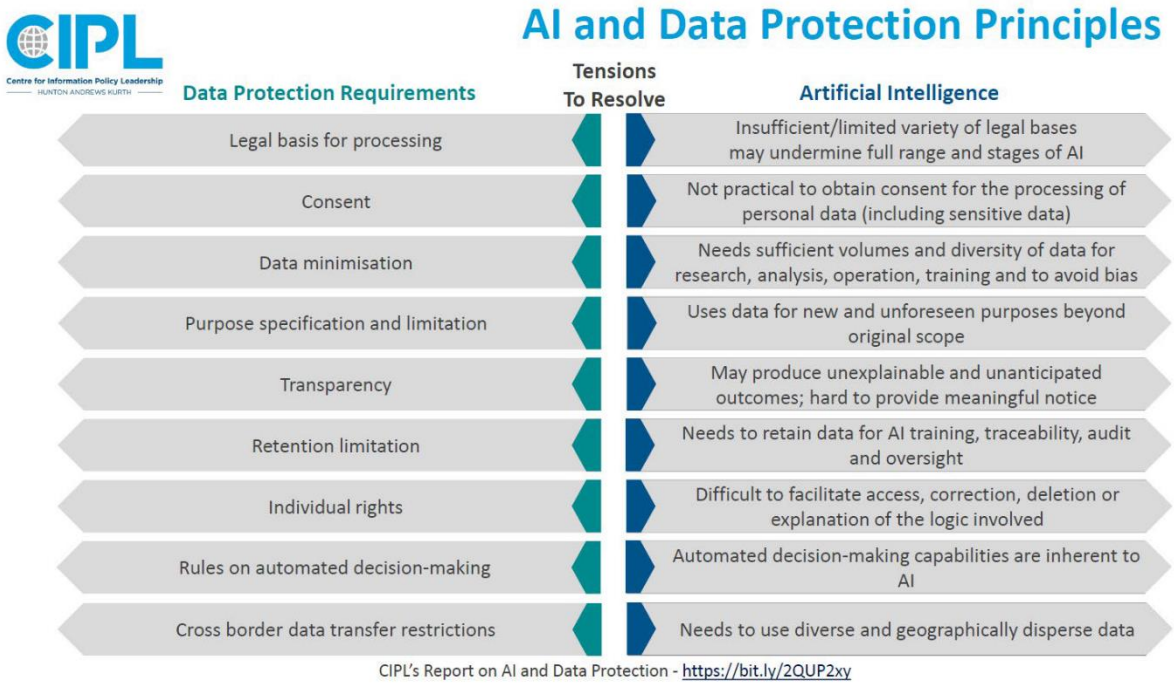

Figure 2 - AI and Data Protection Principles in Tension (Source: CIPL)

Our recent responses to the UK Information Commissioner's Office's (ICO) GAI consultation series further demonstrate CIPL's thinking regarding how principles from data protection can be applied and adapted to responsible GAI development and deployment:

- The first consultation response covered the lawful basis for web scraping to train GAI models.[5] CIPL agrees that while legitimate interest should be recognized under the UK GDPR as a valid and appropriate legal basis for scraping and processing publicly available data to train GAI models, controllers must bear the responsibility of demonstrating that there is a specific valid interest to process personal data for this purpose, and organizations utilizing GAI should implement appropriate, proportionate guardrails throughout the lifecycle of the model to ensure responsible data use and safeguarding of all fundamental rights. CIPL believes that this principle is important and applicable even outside of jurisdictions where the concept of a legitimate interest basis for processing exists in law.
- The second response covered how the purpose limitation principle should be applied at different stages in the GAI lifecycle.[6] While this principle was originally implemented to prevent organizations from a "free-for-all" use and re-use of individuals' personal data, the principle should be applied in a manner that allows model developers to articulate purposes

[5] CIPL, "Response to ICO Consultation on the Lawful Basis for Web Scraping to Train Generative AI Models", March 2024, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_response_-_ico_consultation_on_the_lawful_basis_for_scraping_data_for_generative_ai__mar_2024_.pdf

[6] CIPL, "Response to ICO's 2nd Consultation on Purpose Limitation in the Generative AI Lifecycle", April 2024, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_response_-_ico_consultation_on_purpose_limitation_in_the_generative_ai_lifecycle__apr_2024_.pdf

that are sufficiently broad and flexible for the range of potential applications for which they may be used. Not only do AI technologies require extensive amounts of data for training, development, and operation, but it may also be difficult for model developers to predict all potential applications of their released models to satisfy the purpose limitation principle. Thus, transparency from model developers is crucial to provide meaningful communication to both model deployers and users. In addition, deployers should provide clear explanation of how and why personal data is used to operate the GAI application.

Compliance with data protection principles is an important element of providing comprehensive, appropriate protections against the potential risk of harms from GAI development and deployment. It is imperative for organizations to put in place safeguards that satisfy all elements of accountability, including organization-wide ethical principles, comprehensive risk assessments, reasonable transparency measures, demonstratable policies and procedures, holistic AI-related trainings, ongoing monitoring procedures, and robust oversight and enforcement measures.

CIPL also supports the deployment of emerging technical solutions to safeguard individual privacy while obtaining the value of collected data, such as privacy-enhancing and privacy-preserving technologies (PETs/PPTs). Our report, *Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age*[7] provides insight and analysis into the different types of PETs available, outlines practical applications and case studies, and explores how PETs can help preserve data protection principles and support innovation. At the same time, CIPL supports NIST's caution against overreliance on such technologies, especially as research on their benefits and potential limitations continues.[8] We encourage greater discussion amongst stakeholders to increase awareness of successful adoption and uptake of such technology and to show how they can complement regulation by supporting the implementation of privacy protection principles.

Feedback on the Actions to Manage GAI Risks: NIST has provided a set of actions to help organizations govern, map, measure, and manage GAI risks, and is requesting feedback on whether certain actions could be combined, condensed, or further categorized, as well as feedback on the risks associated with certain actions.

CIPL supports NIST taking steps to consolidate actions. As they are currently presented in the Profile, the actions may be impracticable to implement and potentially burdensome to organizations, particularly to SMEs or organizations with limited resources. We also urge NIST to reflect on how the Profile could impact the broader GAI ecosystem, including open-source models. NIST should also take steps to clarify which actions apply to foundation model developers, and which to downstream

[7] CIPL, "Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age", December 2023, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-understanding-pets-and-ppts-dec2023.pdf

[8] For example, research on "model collapse" associated with use of synthetic data in model training, and how to prevent or mitigate it, continues. See, Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget, April 14, 2024, https://arxiv.org/pdf/2305.17493, and Seddik et al., "How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse," April 7, 2024, https://arxiv.org/pdf/2404.05090.

developers. Furthermore, the categorization by AI RMF functions (i.e., govern, map, measure, manage) leads to apparent duplication on some themes (e.g., several actions relating to provenance). CIPL suggests NIST consider alternative methods of breaking down the presented actions that can aid in reducing potential duplication of similar tasks (e.g., according to the set of risks above or the relevant stage of the model lifecycle).

Appendix A – NIST's set of GAI risks

Below are the 12 risks that NIST identified in the Profile as unique to or exacerbated by GAI:

1. **CBRN Information** - Lowered barriers to entry or eased access to materially nefarious information related to chemical, biological, radiological, or nuclear (CBRN) weapons, or other dangerous biological materials.
2. **Confabulation** - The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications").
3. **Dangerous or Violent Recommendations** - Eased production of and access to violent, inciting, radicalizing, or threatening content, as well as recommendations to carry out self-harm or conduct criminal or otherwise illegal activities.
4. **Data Privacy** - Leakage and unauthorized disclosure or de-anonymization of biometric, health, location, personally identifiable, or other sensitive data.
5. **Environmental** - Impacts due to high resource utilization in training GAI models, and related outcomes that may result in damage to ecosystems.
6. **Human-AI Configuration** - Arrangement or interaction of humans and AI systems which can result in algorithmic aversion, automation bias or over-reliance, misalignment or mis-specification of goals and/or desired outcomes, deceptive or obfuscating behaviors by AI systems based on programming or anticipated human validation, anthropomorphization, or emotional entanglement between humans and GAI systems; or abuse, misuse, and unsafe repurposing by humans
7. **Information Integrity** - Lowered barrier to entry to generate and support the exchange and consumption of content which may not be vetted, may not distinguish fact from opinion or acknowledge uncertainties, or could be leveraged for large-scale dis- and mis-information campaigns.
8. **Information Security** - Lowered barriers for offensive cyber capabilities, including ease of security attacks, hacking, malware, phishing, and offensive cyber operations through accelerated automated discovery and exploitation of vulnerabilities; increased available attack surface for targeted cyber attacks, which may compromise the confidentiality and integrity of model weights, code, training data, and outputs.
9. **Intellectual Property** - Eased production of alleged copyrighted, trademarked, or licensed content used without authorization and/or in an infringing manner; eased exposure to trade secrets; or plagiarism or replication with related economic or ethical impacts.
10. **Obscene, Degrading, and/or Abusive Content** - Eased production of and access to obscene, degrading, and/or abusive imagery, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.
11. **Toxicity, Bias, and Homogenization** - Difficulty controlling public exposure to toxic or hate speech, disparaging or stereotyping content; reduced performance for certain sub-groups or languages other than English due to non-representative inputs; undesired homogeneity in data inputs and outputs resulting in degraded quality of outputs.
12. **Value Chain and Component Integration** - Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not

cleaned due to increased automation from GAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users.

Appendix B – Emerging Best Practices in Accountable AI Programs, Mapped to the CIPL Accountability Framework

The following table outlines a sample of emerging best practices and examples from accountable AI programs used by organizations from different sectors, geographies, and sizes. These practices are mapped to the corresponding element of the CIPL Accountability Framework. The practices are not intended to be mandatory industry standards but rather serve as examples of how companies are implementing specific practices to foster accountability in their development, deployment, and use of AI technologies. Each of the following should be calibrated based on risks, industry context, business model, size, and maturity level of the organization.

ACCOUNTABILITY ELEMENT	**RELATED PRACTICES**
Leadership and Oversight	• Establishing "tone from the top" and demonstrating a commitment to advance ethics, values, and specific principles in AI development, deployment, and use • Implementing systematic processes and escalation pathways for AI-related decision making • Establishing AI ethics oversight bodies or committees (internal or external) to review risky AI use cases and promote ongoing improvements to AI practices • Appointing a board member for AI oversight • Appointing a responsible AI lead, AI officer, or AI champion • Setting up an internal interdisciplinary AI board or AI committee • Establishing organization-wide AI ethics principles • Ensuring inclusion and diversity in AI model development and AI product teams • Creating a centralized governance framework with oversight from the top that still provides flexibility within internal teams • Expanding the remit of privacy teams to include AI-related responsibilities • Leveraging the expertise of other relevant teams (e.g., engineering, data science, legal, ethics and compliance, etc.) to ensure multidisciplinary, cross-functional AI teams • Encouraging employee reporting throughout all levels of the organization by offering escalation pathways to resolve potential AI-related issues
Risk Assessment	• Developing algorithmic impact assessments or fairness assessment tools to monitor and continuously test algorithms to avoid human bias, unfair discrimination, and "concept drift" throughout the entirety of the AI lifecycle • Requiring AI risk assessments at multiple points throughout the AI lifecycle, particularly for new or updated use cases or applications • Creating ethics, human rights, and/or data protection impact assessments • Creating a risk taxonomy that categorizes AI-related risks and allows for uniform assessment • Keeping a centralized repository of all risk assessment documentation

	• Developing standardized risk assessment methodologies that consider the benefits of the AI application or use, the likelihood and severity of risk factors on individuals and/or society, the level of human oversight needed for individually automated decisions with significant impact (e.g., legal ramifications), the ability to explain the technology in the appropriate context, and the ability to audit its effectiveness • Documenting considerations (e.g., accuracy, data minimization, security, transparency, scope of impact, benefits to society) for high-risk processing • Assessing data quality against key performance indicators (KPIs) • Evaluating the data vis-à-vis the purpose of its use (i.e., the quality of the data, its provenance, whether it's personal, synthetic, in-house, or externally sourced) • Developing frameworks for data preparation and model assessment – including feature engineering, cross-validation, back-testing, standardized KPIs • Enabling close collaboration between business and data experts (e.g., data analysts, data engineers, IT, and software engineers) on a regular basis to assess accuracy, ensure appropriate outputs, and allow for proper use of the model • Using privacy enhancing technologies (PETs) to preserve the privacy and security of AI systems • Outlining escalation pathways to send AI-related issues to an AI ethics council or other oversight body • Evaluating and testing models in specific application contexts prior to widespread deployment
Policies and Procedures	• Adopting specific AI policies and procedures on how to develop, deploy, or sell AI • Drafting policies on the application of privacy and security by design principles throughout the AI lifecycle • Setting rules on the level of verification for data input and output • Requiring pilot testing of AI models before release • Specifying the use of protected data (e.g., encrypted, pseudonymized, tokenized, or synthetic data) in training AI models • Creating a glossary of AI-related terms for internal use and reference • Promoting the use of smaller, higher quality data sets • Cleaning and curating data sets before model training through automated or manual checks • Considering relevant and appropriate use of PETs and PPTs to integrate privacy and security controls into AI models • Outlining special considerations for organizations creating and selling AI models, software, applications • Developing a fairness or AI impact assessment to analyze and mitigate AI-related risks • Creating due diligence/self-assessment checklists or tools for business partners deploying AI

	• Clearly defining escalation steps for reporting high-risk AI issues • Implementing an ideation phase with all stakeholders (e.g., data scientists, business, final user, control functions) where needs (including explainability), outcomes, validations rules, maintenance, and budget are discussed • Implementing specific policies for internal GenAI use • Requiring consideration for diversity in relevant teams and business functions • Implementing internal policies in parallel with forthcoming AI regulation • Translating internal principles-based policies to third-party vendor agreements, language, and due diligence processes • Creating processes for review of high-risk AI use cases by an AI ethics board or council
Transparency	• Tailoring transparency measures for the different needs of end users, regulators, business partners, and internal stakeholders at all stages of the AI lifecycle • Communicating disclosures in a simple, easy-to-understand manner • Considering how AI disclosures can be inclusive and accessible for those with special needs/disabilities • Establishing a transparency trail to explain automated decision-making and broad workings of algorithms • Providing notice when the system relies on AI/ML • Providing counterfactual information (e.g., how different inputs can affect the output of an AI model) • Understanding customers' expectations and deploying AI technologies based on their readiness to embrace AI • Implementing tiered transparency • Defining criteria for internal deployment of AI technologies based on usage scenarios and communicating them to users • Publishing model or system cards (i.e., short documents accompanying AI models that describe the context in which a given model is intended to be used and how the model performs in a variety of conditions) • Creating a data hub for information regarding data governance, data accessibility, data lineage, data modification, data quality, etc. • Tailoring transparency to the identified risk (e.g., using watermarking for generative AI output) where possible and appropriate • Participating in benchmarking opportunities, public engagement, and regulatory sandboxes • Using visualization tools to depict difficult, technically complex concepts to end users
Training and Awareness	• Providing specific training for data scientists and engineers, including how to address relevant ethical issues (e.g., how to limit and address bias) • Creating opportunities for cross-functional training (e.g., between privacy professionals and AI engineers) • Tailoring trainings regarding ethics and fairness in AI for relevant teams

	• Compiling and making available AI use case information where relevant risks have been mitigated or deployment has been halted • Creating a "translator" role that helps explain the impact and technical capacities and limitations of AI • Sharing case studies to help employees learn how to address potentially complex, ethically challenging AI cases • Incentivizing compliance with completing ethics training by pairing it with eligibility for bonuses, pay raises, and/or promotions, or incorporating it into other mandatory training activities
Monitoring and Verification	• Incorporating "human in the loop" (HITL) in design, oversight, and redress • Identifying and understanding which business functions are using AI • Providing the capability for human audit of input and output • Ensuring human review of individual decisions with legal or similarly significant effects • Monitoring the data ecosystem—from data flow in, through data process, to data flow out • Using different auditing techniques • Deploying counterfactual testing techniques • Pre-defining AI audit controls • Creating an internal audit team with expertise in AI and other emerging technologies • Allowing human control or intervention where technically possible and reasonably necessary • Monitoring AI models (e.g., back-testing and feedback loop) and conducting ongoing maintenance • Red teaming and adversarial testing of AI models
Response and Enforcement	• Enabling redress mechanisms to remedy an AI decision • Permitting redress through a human, not to a bot • Developing communication channels for internal (e.g., for employees) and external (e.g., end users, business customers) to report and address feedback, complaints, requests, etc.