Cal Gov Ops
Re: Executive Order on Generative AI: Engagement Survey for E01
Oct. 16, 2023

Dear Matthew Tabarangao:

Please find responses collected from affiliates of the UC Berkeley CITRIS Policy Lab and Berkeley AI Research Lab below. We commend the Governor's efforts to better ensure the state of California harnesses AI, especially generative AI, in ways that maximize benefits for all Californians.

We welcome the opportunity to further discuss our comments below and to continue to provide feedback and guidance as the state progresses on this initiative. If the team should have any questions, please contact Brandie Nonnecke.

Thank you,

Brandie Nonnecke, PhD
Director, CITRIS Policy Lab, CITRIS and the Banatao Institute
Co-Director, AI Policy Hub
Co-Director, Artificial Intelligence, Platforms, and Society, Berkeley Center for Law and Technology, Berkeley Law
Assoc. Research Professor, Goldman School of Public Policy
citrispolicylab.org | @BNonnecke | nonnecke.com

Executive Order on Generative AI: Engagement Survey for EO1 September 27, 2023
Governor Newsom's executive order on artificial intelligence seeks to better understand the risks and issues associated with this generative artificial intelligence (GenAI). Your expertise is requested in providing input, insights, and sharing references for the following questions.
Your responses will be shared with staff at the Government Operations Agency, Department of Technology, and Office of Data and Innovation and will contribute to the development of the report as required in section 1 of the executive order.
The Governor's Executive Order has issued a deadline of Monday, November 5th to submit a report on beneficial use cases and risks. In order to incorporate your comments, please submit your responses by Friday, October 13th by emailing Matthew Tabarangao (Matthew.Tabarangao@govops.ca.gov).
We deeply appreciate your willingness to share your expertise! Thank you! Name and Contact Information

1. What are the most potentially beneficial use cases of GenAI tools by the State of California?

1. California has a diverse population who speak 3 major languages: English, Spanish, and Mandarin. Large language models can assist in translating text into different languages and to synthesize and simplify the communication of complex topics for California's diverse population.
2. California educates students with very diverse abilities and backgrounds. Generative AI could be used to provide personalized tutors for a broad variety of subjects.
3. Generative AI has the potential to save the state millions of dollars by providing services in more efficient, effective, and equitable ways, for example providing detailed customer service for all state services.
4. Generative AI can be used to assist in training and offloading simple tasks for state personnel.
5. Generative AI can semi-automate the writing of very structured documents by drafting repetitive and boilerplate parts.

2. Are there any generative AI use cases that government should not pursue? Please provide details behind your recommendations.

- To avoid confusion about authenticity, the state of California should not use generative AI to generate images, videos, and voices of elected or appointed officials or to translate voices into other languages. For example, an Indian politician used generative AI to translate his speech into several languages to communicate with his diverse

constituents(https://www.theverge.com/2020/2/18/21142782/india-politician-deepfakes-ai-elections). While promising in its ability to enable elected leaders and officials to communicate with their constituents, creation and use of synthetic media (images, video, and voice) can inadvertently raise confusion around authenticity, a challenge that will only increase as nefarious actors use these techniques to manipulate and influence constituents and elections.
- The government should absolutely not pursue the use of generative AI to elicit or infer an individual's private or sensitive information.

3. Are there any frameworks or organizing typologies for GenAI, especially as they relate to use cases?
—

4. What would be the consequences if government did not adopt any GenAI technologies at all?
- See answer to 1 above. California could lose out on millions of dollars in cost savings and missed opportunities for residents if these opportunities are not pursued.
- The state should leverage generative AI, but should also leverage other forms of AI, such as machine learning (linear regression, logistic regression, etc.). While discussions on how to harness generative AI are other forms of AI are more relevant and useful for the state.

5. What considerations or criteria should we incorporate in order to determine high risk use cases?

- The state has indicated its intention to implement algorithmic impact/risk assessments, such as the NIST AI risk management framework and will draw upon the EU conformity assessment process. This is promising, but the state should carefully consider
- The EU AI Act and IEEE 1012 provide strong typologies of tiered risk levels (minimal, limited, high, and unacceptable risk and in the case of IEEE, risk as it relates to probability of severity of these risks). The risk typologies must have corresponding requirements for assessment, mitigation, and transparency of risks. With this tiered risk model in mind, the state should consult with internal and external stakeholders to determine risk levels and probability of these risks for particular domains/use cases.

6. Our teams are leveraging risk frameworks from the National Institute of Standards and Technology (NIST) and the European Union. What issues do you anticipate with leveraging these resources for generative AI risk evaluation? Are there any other risk frameworks that should inform our thinking on GenAI?

- The NIST AI Risk Management Framework (RMF) is nascent. As such, it is still unclear how the RMF will effectively be applied to the evaluation of generative AI technologies. Yet work is progressing in this space. The Center for Long-Term Cybersecurity and CITRIS and the Banatao Institute at UC Berkeley have developed a NIST AI risk-management standards profile for general-purpose AI systems (GPAIS), foundation models, and

generative AI, such as large language models. Our profile provides guidance on how to identify, evaluate, and address risks posed by generative AI, including those faced by the public sector. You can view the current draft of the profile at https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/.

- The state should also consider implementation of human rights impact assessments (HRIAs) to identify and address risks of generative AI. Features of HRIAs can easily be incorporated into the NIST AI RMF and EU conformity assessments. Brandie Nonecke and Philip Dawson published work through the Harvard Kennedy School on the human rights implications of risk/impact assessments (see https://carrcenter.hks.harvard.edu/publications/human-rights-implications-algorithmic-impact-assessments-priority-considerations. While promising, ill-determined scope and methodologies employed in AI risk/impact assessments for generative AI can inadvertently overlook the risks they seek to address. Generative AI can be applied in a variety of domains (e.g., health) and for different use case applications within a domain (e.g., preventative care, medical billing). Thus, the state should not only implement risk/impact assessments at the domain level, but for each use case application and at regular intervals throughout its implementation.

7. What other considerations should we include regarding GenAI risks stemming from bad actors and insufficiently guarded governmental systems, unintended or emergent effects, and potential risks toward democratic and legal processes, public health and safety, and the economy?

- The state should implement red teaming (drawn from the field of cybersecurity) to assess the potential unintended or emergent effects and risks. The state should require any provider/vendor of generative AI to demonstrate it has adequately implemented red teaming to identify and protect against these unintended or emergent effects and risks before procurement and throughout the state's use of the generative AI tool.
- The state should consider providing funding and support to academia and industry for building evaluation tools and frameworks for evaluating the safety, security, and in general trustworthiness of AI models and systems.
- Constant monitoring of generative AI systems—especially after deployment—is needed to guarantee that they are not used for purposes they were not intended to.

8. Do you have suggestions for frameworks to evaluate the efficacy of GenAI use cases within an organization?

All implementations should clearly indicate that Generative AI is biased based on what it was trained on and can be unreliable. All interfaces should include a button for users to report harmful (e.g. biased) and questionable responses so these can be investigated.

DecodingTrust (decodingtrust.github.io) is the first comprehensive evaluation framework for trustworthiness of LLMs. Developed by faculty from UC Berkeley and other collaborators, DecodingTrust has been adopted by a number of leading tech companies.

9. What criteria should we consider in whether to continue or stop GenAI pilots? Please include links to examples where possible.

- Every pilot project should be thoroughly tested by outside experts and random residents to identify bugs, confusing user interfaces, and faulty reasoning before being deployed.
- The state must set risk thresholds for AI, including in the areas of security, safety, rights, and environmental sustainability. During sandboxing (i.e., small-scale testing of a generative AI tool within a protected environment) the state must evaluate whether the rollout and scaling of the AI application will cross these thresholds and whether effective risk mitigation strategies are possible and are in place.
- It's important to establish testing best practice and standardization for evaluating trustworthiness of AI models and systems. Here's an example of recent work in establishing the first comprehensive evaluation framework for trustworthiness of LLMs with faculty from UC Berkeley and other collaborators: decodingtrust.github.io
- After deployment, feedback from users and residents should be constantly analyzed and taken into account to ensure that systems are still working as expected.

10. What barriers do you anticipate impacting state government adoption and deployment of generative AI?

- In order for the state to appropriately use generative AI, especially in its service provision, the state will need to ensure the model was trained with appropriate and robust data. Since generative AI models are built using scraped publicly available data, which represents all of the implicit and explicit biases present in society and may not be directly relevant to the provision of state services, the state must ensure these models operate appropriately for their application areas and provide warnings that responses may be biased or incorrect (see response to item 8).
- While the state has proposed the use of sandboxing, robust testing of generative AI tools will not be simple. As a generative AI tool is applied and scaled "in the wild", spurious or unintended outcomes may be produced. It is imperative that the state establish a process to assess the performance of generative AI tools at regular intervals over time.
- Generative AI systems are often both overperforming and unreliable at the same time. It's important to acknowledge the limitations and have constant human supervision. When supervised, they can be useful tools, otherwise their output will likely lead to unintended consequences.

11. Please provide a list of 3-5 additional resources that might be useful to the Governor in his evaluation of GenAI.

- Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, Krystal Jackson, "AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models: Second Full Draft," 2023, https://drive.google.com/file/d/1NHFJrhgemsMxbEnPngDRgwQNPEBcg1ZQ/view.
- "Decoding Trust: Comprehensive Assessment of Trustworthiness in GPT Models," 2023, Decodingtrust.github.io
- Using Generative AI in Research, USC Libraries, Oct. 2, 2023, https://libguides.usc.edu/generative-AI/limitations.
- The Benefits and Limitations of Generative AI: Harvard Experts Answer Your Questions, Harvard, April 19, 2023, https://www.harvardonline.harvard.edu/blog/benefits-limitations-generative-ai.

12. Any other advice, insights, or comments regarding the potential deployment of GenAI in government?

- Generative AI is a new technology that is not well understood, even by experts. It can perform surprisingly well in image and text generation, but it is prone to bias and errors. It is VERY important that those considering its use learn about the many limitations of generative AI and keep these in mind when making decisions about its use.
- California should develop resources for state employees considering use of generative AI. The government must support the development of training materials for government personnel in generative AI tool capabilities, how to use them appropriately, and how to evaluate their effectiveness, including mechanisms for personnel to report when
- The state's efforts and the executive order should not only focus on generative AI, but on a broad range of AI methodologies, such as machine learning, reinforcement learning, deep learning and more. The state is currently using many of these methodologies and would benefit greatly from the execution of a similar process as outlined in the Governor's executive order on generative AI.