Mathematica
Progress Together

Response to

Senator Cassidy's White Paper on Artificial Intelligence

September 27, 2023

Submitted to:

The Honorable Bill Cassidy, MD
Ranking Member
Senate Committee on Health, Education, Labor and Pensions
428 Senate Dirksen Office Building
Washington, DC 20510

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
Phone: (202) 484-9220
Fax: (609) 228-4958

September 27, 2023

RE: AI Frameworks

Dear Ranking Member Cassidy,

Thank you for engaging a wide stakeholder community in your effort to improve the framework in which artificial intelligence (AI) is developed, reviewed, and deployed. As a nonpartisan research and data analytics organization with a mission to improve public well-being, Mathematica recognizes the importance of ensuring the use of AI is centered around the rights and safety of the American people. The rapid development and deployment of AI creates both enormous opportunities and risks to public well-being. Mathematica's deep, interdisciplinary expertise—including our experience using AI tools—positions us to anticipate opportunities and risks alike, and to consider the ways policy might support those opportunities and mitigate those risks.

We appreciate your interest in having this conversation in this moment. the pace of AI development and deployment is so rapid that favorable and unfavorable world-changing consequences are possible in a brief time. In the following pages, we draw on the broad and deep expertise of Mathematica staff across disciplines and policy domains to address many of the critical questions posed in your white paper on AI. Our own disciplined use of AI to inform public decision making across contexts—from public health and education to employment and labor—informs our responses. We have employed our AI expertise in work on the following initiatives:

- We have designed and trained AI solutions for a challenge sponsored by the Centers for Medicare & Medicaid Services (CMS).
- We predicted fatal opioid overdoses using geospatial analytics.
- We employed agent-based models to predict the spread of COVID-19 in schools operating in various operational conditions.
- We explored the value of wastewater data in predicting surges in communicative diseases.
- We used machine-learning techniques to identify academically at-risk students.
- We employed machine learning and natural language processing to assign occupation codes for doctorate recipients.

In these and other projects, we have shown how AI can be a powerful tool to support decisions and augment services that promote public well-being in general and equity in particular. Our experience with AI tools has also informed us about many of the risks they can present, including how poorly designed or implemented AI systems might increase inequities. We are aware that, as with any other decision-making tool, the way AI is designed and deployed can have a profound effect on outcomes.

Our response is informed by more than a half century of rigorous analysis of organizational systems for implementing policy and delivering services in domains such as health, education, employment, and international development. The effects of AI on these systems are likely to vary substantially, depending not only on the particular AI tools in question but institutional structures and features. Mathematica knows these institutional features well; therefore, we can anticipate some of the ways AI's impact might differ across sectors.

For any questions about our response, please contact Mike Burns, Mathematica's senior director for communications and public affairs, at MBurns@mathematica-mpr.com.

Sincerely,

Christopher Trenholm
Senior Vice President; General Manager, Health

Item-by-Item Response

A. What practices are in place to mitigate bias in AI decision-making?

The Blueprint for an AI Bill of Rights and the AI Risk Management Framework play key roles in codifying guiding principles for responsible AI, including addressing issues such as bias. Several responsible AI tools already exist that apply the Algorithmic Discrimination Protection principle of the blueprint at each stage of the data science project life cycle (that is, designing and developing models, identifying and quantifying bias, mitigating bias, and explaining bias). Datasheets for Datasets and Carnegie Mellon University's Data Science Project Scoping Guide are useful during the model design phase to address potential ethical issues before they occur and ensure that the project is designed and implemented with equity as a priority. Once a model is built, various tools are available to assess fairness and quantify harms the AI system might pose. Fairlearn and Aequitas are open-source toolkits that explain how a machine learning model can treat certain groups or individuals differently, especially vulnerable populations. Whereas fairness metrics identify bias, bias mitigation algorithms—such as those available in AI Fairness 360's open-source package—help mitigate sources of bias and increase equity. Lastly, because it might not be possible to remove bias entirely, transparency and explainability tools such as SHAP and Model Cards are useful for disclosing and describing any bias that remains in the model. These tools can help put the blueprint's principles into practice if they are used at each stage of the model development process.

Mathematica has experience with implementing principles into analytic best practices through our work developing decision support algorithms across a variety of sectors, including health, education, and social services. At the start of every new project, Mathematica screens for potential high-risk or high-impact applications of machine learning that require further oversight and potential use of the aforementioned mitigation tools. We mitigated bias in our AI decision-making solution in response to the CMS AI Challenge; specifically, Mathematica used mixed methods, such as codesigning the AI Solution with the Patient Advocate Foundation and adjusting models for the under-representation of Hispanic people in the training data. In addition, for the CMS AI Challenge, Mathematica used SHAP to help create an explainability dashboard that conveys model results to clinicians and patients.

B. Who should be responsible for determining safe and appropriate applications of AI algorithms?

Voluntary oversight of AI systems is not sufficient, given the collective action problem: individual firms know there is substantial private value in being first in developing new AI technologies, and society bears much of the risk. These realities make a prototypical case for government regulation of AI—although exactly how government should regulate is a much harder question.

Some agencies, such as the Food and Drug Administration, have policies and procedures for regulation of certain AI-enabled products. However, to date, most auditing of algorithmic bias is done reactively by academics and activists. These audits focus on widely used algorithms and their results are shared publicly, stimulating bias mitigation via public shame, lawsuits, or subsequent government investigation. A public interest group conducted such an audit in 2016 of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism software, which revealed the tool's predictions were racially biased. This set off public debate regarding (1) the predictive validity of AI-driven tools, especially those with proprietary algorithms

like COMPAS, and (2) whether it is ethical to use these tools in high-risk scenarios in which people's lives are severely and adversely affected if the model makes an incorrect prediction. Although these high-impact audits give government agencies the evidence they need to open an investigation into an algorithm, authorities cannot rely on the altruism of private actors, particularly as use of AI systems increases.

In response to this problem, the U.S. government might conduct oversight by establishing an agency whose mission is to use audits as a means of enforcing legislation that regulates the use of AI. The government could draw upon legislation such as the European Union (EU) AI Act to establish concrete rules regarding the levels of risk posed by various types of AI systems and the allowability of such systems based on their level of risk. In the early stages, the government could contract with a third-party organization to conduct these audits and then build in-house capabilities over time.

In addition to conducting downstream audits of AI systems, the government could focus on upstream risk management, such as the use of responsible AI tools. For example, regulators could push for more public assessments of data quality, such as the work Mathematica conducts on behalf of CMS to assess Transformed Medicaid Statistical Information System (T-MSIS) data. Next, the government could encourage the development and use of algorithmic fairness tools, such as Mathematica's Bayesed and Confused, an open-source application to quantify algorithmic fairness. Furthermore, regulators could encourage the use of Bayesian-backed AI systems to quantify uncertainty when assessing model risk; Mathematica previously demonstrated the approach's utility, using it to stabilize measures of school performance. The tool helps reduce the likelihood that policymakers will make consequential decisions based on random error.

C. What are the best practices currently being used to ensure that AI systems are designed, developed, and deployed in a manner that protects people's rights and safety?

For AI to protect people's rights and safety, AI-driven solutions must effectively address the problems they seek to solve. Right now, few AI solutions are evaluated rigorously to determine effectiveness. One example of a well-evaluated AI platform is Bayesian Health's clinical platform, which is backed by research showing reductions in mortality. Researchers and regulators should promote frameworks on how to evaluate the effectiveness of AI solutions, such as rapid-cycle evaluations, because large experimental designs are not always possible—especially for large-scale, consumer-facing AI tools.

Measures taken to protect rights and safety must account for the fact that various types of AI systems exist, and AI systems evolve over time. Over the past decade, we have seen an increase in the use of AI systems to guide decision making and resource allocation in sectors where protecting people's rights and safety are of the utmost importance, such as health care, education, transport, employment, and welfare. To date, most of these AI systems have used classification and ranking algorithms for which research and frameworks exist to support responsible design, development, and deployment. In contrast, far less guidance is available on how to responsibly build and deploy AI systems that use performant large language models (LLMs) such as GPT4, which are being deployed rapidly to execute various tasks, including customer support, content creation, and translation. The pervasiveness of AI begs the question: Is AI now a social determinant of health and well-being?

Looking to the future, people's rights and safety must consider even broader applications of AI that do not exist yet, such as artificial general intelligence (AGI) (Russell 2019). Although there is no consensus about when AGI will arrive, the possibility of a misalignment of interests between humanity and a super-intelligent AGI is real and demands the attention of policymakers ahead of its creation and deployment. The possibilities of AGIs demonstrate the need for investing in AI alignment research, with the goal of ensuring safe solutions are built in to future AGIs when they launch.

D. What do policymakers need to know about the development of AI standards?

Proper regulation requires improved understanding of AI's risks and safety on the part of policymakers and the public. Public agencies and philanthropies should therefore invest in translational media to inform decision makers and the public about AI's risks and safety. Increased collaboration between academic researchers, policymakers, and the technology sector could promote translational work. Translational media (such as the kinds of infographics, blog posts, podcasts, webinars, and conferences Mathematica frequently supports) can play a key role in clarifying what AI systems look like in practice, providing policymakers with knowledge that will enable them to competently collaborate with researchers and technologists to develop feasible standards and relevant regulations.

In the case of LLMs, the most pressing need is research that establishes best practices for responsible design, development, and deployment. As shown by OpenAI's grant program for research on best practices, the extent to which AI systems might threaten people's rights and safety remains unclear, and prevention and mitigation measures for these issues are yet to be established. Although the 117th Congress proposed several bills that would regulate generative AI, filling the knowledge gap is key to ensuring these regulatory bills address the varied challenges generative AI systems might present. For example, some methods exist for detecting and measuring bias in generative AI outputs, such as adversarial training (for example, red teaming) and counterfactual methods. However, there is a need for methods that quantify the extent of bias present in these models and the consequences of these biases. These methods could answer questions about issues such as the frequency with which ChatGPT outputs harmful stereotypes. (What are the impacts of those stereotypes if, for example, ChatGPT is used to power a mental health chatbot?) In addition, regulators should consider the quality and representativeness of the data used to train LLMs.

E. What are high-risk use cases of AI with respect to discrimination?

The EU AI Act provides a useful framework for considering AI uses and risk factors, with specific definitions for vulnerable populations. The EU AI Act categorizes three levels of potential risks: minimal or limited risk, high risk, and unacceptable risk.

One high-risk field that currently uses AI involves the provision of reentry services to formerly incarcerated individuals. In a Mathematica survey of U.S. Department of Labor (DOL) Reentry Project grantees, 84 percent reported using participants' risk and needs assessments for screening or service planning, and most grantees used more than one assessment (Stein et al. 2023). Using risk and needs assessments for service planning and program screening can help staff develop focused treatment plans and screen out participants who might not benefit from services (Duran et al. 2013; Taxman and Smith 2021). However, potential issues associated with risk and needs assessments and associated AI tools that support decision making could compromise their usefulness or introduce

harms to participants. Predictive validity statistics alone are insufficient, as they depend on a data sample that might not represent the intended population or how the AI system is used in practice.

Agencies can potentially mitigate some of these issues by using risk and needs assessments that an independent third party has evaluated and tested with the populations they serve (Desmarais et al. 2022; Electronic Privacy Information Center 2020; Larson et al. 2016), as well as fidelity reviews to monitor implementation of risk and needs assessments (Wormith and Bonta 2018).[1] This Mathematica brief outlines these equity-driven challenges to using risk and needs assessments and their potential solutions in more detail.

[1] Wormith, J. S., and J. Bonta. "The Level of Service (LS) Instruments." In *Handbook of Recidivism Risk/Needs Assessment Tools,* edited by J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, and Z. Hamilton. John Wiley & Sons, 2018, pp. 117–145.