HANDBOOK FOR AI VALUES ALIGNMENT

CONTENTS

INTRODUCTION

With Mayor's Order 2024-028, Mayor Muriel Bowser firmly committed DC agencies to explore and deploy AI tools in careful alignment with DC's **AI Values**. To help DC hold itself accountable for the important work of AI Values Alignment, Mayor Bowser established a transparent, public **AI Values Alignment Advisory Group**. Led by the Assistant City Administrator and a public co-chair, the Advisory Group is empowered to review agencies' AI Values Alignment efforts, and to share its assessment of such work directly with Mayor Bowser.

DC government's internal **AI Taskforce** is charged by the Order to support agencies in their AI Values Alignment. As part of that duty, the AI Taskforce and the Office of the Chief Technology Officer provide this **Handbook for AI Values Alignment**. This Handbook will guide your agency in preparing a strong, carefully considered **AI Values Alignment Report** for each AI tool you hope to deploy in support of your agency mission.

As elaborated upon more fully in this Handbook, your agency's AI Values Alignment Report will contain the following sections, drawn from Mayor's Order 2024-028's articulated AI Values:

- Clear Benefit to Residents
- Safety and Equity
- Accountability
- Transparency
- Sustainability
- Privacy and Cybersecurity

This Handbook discusses each of these AI Values in granular detail, breaking each into constituent, **Key Concepts** to help your agency structure its consideration of these overarching AI Values. Organized hierarchically under these Key Concepts, this Handbook identifies specific questions your agency should answer as you evaluate whether a given AI tool is right for your agency mission. Your answers to the **Privacy and Cybersecurity** questions, in particular, will be used by OCTO in providing the privacy and cybersecurity review and approval required by Mayor's Order 2024-028.

With this Handbook as your guide, your team will produce an AI Values Alignment Report that documents your agency's meaningful AI Values Alignment work for each AI tool you deploy. A complete AI Values Alignment Report—including OCTO Privacy and Cybersecurity approval—will be a necessary supporting document for any AI tool your agency deploys. Additionally, should the AI Values Alignment Advisory Group review your agency's AI deployment, your AI Values Alignment Report will serve as the most important record of your agency's efforts to ensure the AI deployment aligns with DC's AI Values.

As you work through this Handbook, please reach out to OCTO at ai@dc.gov with any questions that may arise.

GENERAL GUIDELINES

Getting your AI Values Alignment Report right will have a lot to do with the success or failure of a given AI deployment. AI Values Alignment can seem daunting. To get it right, your agency must incorporate diverse voices from within your team, analyze complex technological, human, and operational considerations, and make important decisions on the use of emerging technologies in the delivery of government services. As DC government's domain-specific experts on your agency's mission, your team is in the best position not only to conduct this analysis, but to document it as well. As you prepare your AI Values Alignment Report, keep the following Key Concepts in mind.

KEY CONCEPTS

- Use plain language to say what the tool will do!
- Be specific!
- Meet hard questions head-on!
- Show your work!

Use plain language to say what the tool will do! Your AI Values Alignment Report will be the best tool the AI Values Alignment Advisory Group, and DC residents, will have available to help them understand how your AI deployment will impact life within DC. An AI Values Alignment Report not comprehensible to anyone but computer scientists is not likely to be a very effective safeguard against AI deployments out of synch with our AI Values. As you prepare your AI Values Alignment Report, be sure that your level of usage matches the breadth of your intended audience.

Be specific! People sometimes get too excited when they talk about AI tools. Much of the chatter in and around this emerging technology market is broad and vague. As you prepare your AI Values Alignment Report, place a premium on specificity. Describe specific functionalities, specific use cases, and to the extent domain-appropriate try to describe specific deployment strategies. Big, categorical statements will be less valuable to the AI Values Alignment Advisory Group than specific, concrete statements.

Meet hard questions head-on! An AI Values Alignment Report that skirts all the hardest questions isn't very useful. To do its work, the AI Values Alignment Advisory Group will need to ask those questions the public will ask about how well a given deployment aligns with DC's AI Values. Relying on your agency's domain expertise, anticipate the hard questions, think carefully about them, and answer them head-on in your AI Values Alignment Report.

Show your work! One major benefit of going through a formal AI Values Alignment process is that it allows your agency to demonstrate the depth and breadth of your efforts up-front. In answering this Handbook's questions, your team will work through a variety of source documentation and evidence. Show that work in your AI Values Alignment Report. Attach documents, quote proposed contract terms, quote privacy policies.

DC AI VALUE
CLEAR BENEFIT TO RESIDENTS

Why should an agency spend money and take on risk to deploy AI tools? For that matter, why should an agency ever spend money or take on risk at all? The answer to these questions is straightforward: for the benefit of residents.

The DC government was created with the express purpose of granting "to the inhabitants of the District of Columbia powers of local self-government." Accordingly, the high-level goal of maximizing benefit to resident should guide all agency action—including the planning, deployment, and use of AI tools.

How should you assess your planned AI deployment's benefit to residents? As you complete the section of your AI Values Alignment Report focusing addressing **Clear Benefit to Residents**, structure your discussion around these Key Concepts.

KEY CONCEPTS

- What is your **purpose** in deploying this AI tool?
- Who will **benefit** from the AI tool?
- How have you weighed this tool's benefit to residents against reasonable **alternatives**?

PURPOSE

What is your purpose in deploying the AI tool? AI technologies are versatile, and as a result, there is a wide variety of legitimate purposes your agency might have in mind when deploying an AI tool. You might deploy a machine translation tool on your website to help get meaningful translations to members of the public more quickly, more efficiently, or more flexibly than you otherwise could. You might deploy an AI tool to help process a large volume of handwritten material from the public so that those without access to a computer can still have a meaningful, real-time voice in a given process or program. You might use a variety of AI tools to transform teaching materials across a variety of media to better accommodate a diverse set of learners' preferred mode of instruction. As you describe your AI deployment's purpose in your AI Values Alignment Report, be sure that you are answering these questions.

Identify your specific use case. What specific problems or challenges are you seeking to address through this deployment? For example, is the AI tool intended to help increase the uptake of available benefits by qualified applicants? Is the AI tool intended to smooth traffic congestion before and after big events? Is the AI tool intended to minimize waste in perishable inventory distributed among multiple resident-facing locations across DC? **Explain** which of your proposed AI tool's functionalities you intend to use to address these identified problems or challenges.

What use cases did the provider of the AI tool design or market it to address? **Identify** all of the vendor's intended use cases. **Explain** what functionalities the vendor claims the AI tool is capable of providing. Your answer to this question may include more general uses and functionalities than your answer to the previous question. If the vendor's marketing material specific to the AI tool at issue is available, you should attach it to your AI Values Alignment Report.

If any of your intended use cases are **not** included within the vendor's own proffered use cases and functionalities, **identify** these "off-label" uses, and **explain** fully your basis for determining that the AI tool is an appropriate tool for these uses.

Describe the impacts you expect to see post-deployment. For example, if the AI tool is intended to improve uptake in available benefits by qualified applicants, do you expect overall increases in uptake? Targeted increases in subsets of qualified applicants where uptake is currently lagging? Reductions in denial rates?

BENEFIT

Who will benefit from the AI tool? While your discussions of the previous Key Concept will help your agency to document your purpose in deploying the proposed AI tool, your answers to this question will help document how you expect the benefits of the AI deployment to be distributed among relevant stakeholders. By thinking purposefully about the distribution of the proposed AI tool's benefits, your agency can help to ensure all stakeholders share equitably in all that these powerful new tools have to offer.

Identify all communities, groups, or populations you intend to benefit through your planned AI deployment. For example, an AI tool intended to promote more efficient traffic patterns might benefit commuters experiencing reduced commute times, motorists, pedestrians, and cyclists enjoying reduced accident rates, and local businesses enjoying increased in-store traffic as their location becomes more easily accessible.

Identify any communities, groups, or populations you do not intend to benefit directly through your planned AI deployment, but whom you are aware are likely to benefit from your use of the AI tool. In answering this question, pay special attention to the incidental benefits a vendor might obtain from the deployment. For example, will use of the AI tool provide valuable, crowd-sourced model training for the vendor? Will deployment lead to the creation of a new and valuable data set at DC's expense?

Identify any impacts your planned deployment will have on oversight, inspection, or enforcement activities. Will the AI tool help your agency identify fraudulent claims, code violations, misapplications of policy?

ALTERNATIVES

How have you weighed this tool's benefit to residents against reasonable alternatives? Bigger, newer, and more expensive do not always mean *better*. Unnecessary complications are rarely

efficient. As you prepare your AI Values Alignment Report, ask yourself the following questions to help document the thoroughness of your consideration of less complicated alternatives to an AI deployment.

In an earlier question, you identified your proposed use case, including the specific problems or challenges you hoped to address through your deployment. **Identify** your current solutions for addressing these problems or challenges. If the problem or challenge is new or anticipated and no status quo measures are in place, say so.

Identify any other technological solutions you considered for this use case. **Explain** how you selected the current AI tool or type of tool as the most appropriate. Please note that this question is not concerned with how you chose between two competing vendors of text-to-image generative tools, for example. It is concerned with how you determined text-to-image generation was a better fit for your use case than obtaining a license to use a stock image library with a text-based search function, or hiring digital artists, for example.

DC AI VALUE
SAFETY AND EQUITY

Residents and visitors trust DC government with their lives. They trust the design and construction of our buildings, roads, and infrastructure. They trust our regulation of the food they eat, and of the licensed professionals who care for them when they're sick. When emergencies arise, they trust our management of those emergencies. We rescue them from fires and rush them to hospitals. They trust our world-class public safety professionals to safeguard their rights, and trust Mayor Bowser's commitments to promote equity and ensure everyone gets a fair shot.

AI tools provide powerful means for improving the quality and efficiency of government services, but with that power come risks of potential harm. As AI tools interact directly with the public, as they mediate relationships among residents, as they are deployed in oversight, adjudicatory, or evaluative postures, those risks of potential harm must be carefully governed, mapped, measured, managed, and documented over time. This section focuses on **Safety and Equity**, and asks you explicitly to consider four categories of harm.

KEY CONCEPTS

- What risks of **direct physical harm** might flow from your proposed tool?
- What risks of **indirect physical harm** might flow from your proposed tool?
- What risks of the **deprivation of fundamental rights** might flow from your proposed tool?
- What risks of **exacerbating inequity** might flow from your proposed tool?

DIRECT PHYSICAL HARM

What risks of direct physical harm might flow from your proposed tool? This section focuses on the most immediate risks of concrete physical harms a given deployment might entail. For illustrative purposes, consider a hypothetical government deployment of self-driving car technology. At some point in the future, self-driving cars may reduce congestion, minimize traffic fatalities, and free up driver attention for other tasks. However, even the safest self-driving government vehicles we can imagine would entail a significant risk of causing or at least failing to avoid a harmful collision. The existence of such risks does not mean that self-driving cars might not one day be beneficial, but it does mean that at all stages in such an AI tool's lifecycle, the deploying agency would need to carefully measure and mitigate the risk that self-driving cars might directly physically harm residents or visitors to DC. As you address the risk of direct physical harm in your AI Values Alignment Report, be sure that you are answering these questions.

Identify all potential direct physical harms your deployment might be capable of causing. Your answer to this question should take broad account of every kind of interaction your proposed tool may have with residents, visitors, and workers. As part of its own product development, your vendor may have done some assessment of these risks. If so, **identify** all steps taken by the

vendor to identify or mitigate the risk that your proposed AI tool might cause direct physical harms.

While your vendor likely has reviewed the harms its products might cause, and likely has taken steps to mitigate those harms, as you plan your specific deployment you may be in a better position to identify risks of that deployment with better granularity than a vendor operating at scale. In your AI Values Alignment Report, **explain** what steps you have taken in designing your planned deployment to identify and mitigate risks of direct physical harm.

Depending on the nature of a planned deployment, the risks of direct physical harm may not be obvious to the people interacting with your tool. As a result, under some circumstances it may be appropriate for your planned deployment to include processes for affirmatively providing those who interact with your AI tool some notice of the identified risks of direct physical harm. **Explain** any plan your agency has to provide such notice as part of your deployment. If you do not believe it is appropriate to include such affirmative notice of risks of direct physical harm as part of your deployment, **explain** why.

Whatever the benefits of an AI deployment might be, some people may still want to avoid interacting with it. **Explain** whether and how those placed at risk of direct physical harm by the deployment will be able to opt out of interaction with the AI tool. If provision of such opt-out opportunities is untenable, **explain** why.

INDIRECT PHYSICAL HARM

What risks of indirect physical harm might flow from your proposed tool? This section focuses on the risks of concrete physical harms owing their direct cause to some source other than your planned deployment, but which might be meaningfully exacerbated by your planned deployment. This question may be particularly important for deployments that involve the incorporation of AI into safety systems meant to mediate risks of harm among third parties. For example, consider an image classifying AI tool deployed to analyze x-ray images collected at a building's security perimeter. Depending on the nature of that tool's accuracy and speed relative to whatever the status-quo solution might be, deployment of the tool may entail some change in the risk profile for the use of weapons within that building. As you describe the risk of indirect physical harm in your AI Values Alignment Report, be sure that you are answering these questions.

Identify all risks of indirect physical harms potentially implicated by your deployment. **Identify** all steps taken by your vendor to identify or mitigate the risk that your proposed AI tool might lead to indirect physical harms. **Explain** what steps you have taken in designing your planned deployment to identify and mitigate risks of indirect physical harm.

Explain your plan for making those at risk of indirect physical harm aware of that risk prior to interacting with the system. If such notice is not appropriate under the circumstances under which your tool will be deployed, **explain** why.

Explain whether and how those placed at risk of indirect physical harm by the deployment will be able to opt out of interaction with the system. If provision of such opt-out opportunities is untenable, **explain** why.

DEPRIVATION OF FUNDAMENTAL RIGHTS

What risks of the deprivation of fundamental rights might flow from your proposed tool? "The AI made me do it," is never an acceptable excuse for having infringed upon people's fundamental rights. Federal regulators and other oversight authorities have expressed special attention to the impacts the deployment of AI tools may have upon the fundamental rights of people. Might an AI-powered identity verification tool turn eligible voters away from the polls? Might an AI-powered claim classification tool deployed to increase the efficiency or consistency of adjudications deprive benefits applicants of their due process rights? This section of your AI Values Alignment Report is where you will address such risks and your planned mitigation strategies for your proposed deployment.

Identify all potential deprivations of fundamental rights deployment of your tool might entail. **Identify** all steps taken by your vendor to identify or mitigate the risk of such deprivations. **Explain** what steps you have taken in designing your planned deployment to identify and mitigate the risk of such deprivations. **Explain** how you plan to communicate these risks to affected people. If such communication is inappropriate, **explain** why. **Explain** whether and how those at risk of such deprivations of fundamental rights will be able to opt out of interaction with the system. If provision of such opt-out opportunities is untenable, **explain** why.

EXACERBATING INEQUITY

What risks of exacerbating inequity might flow from your proposed tool? Many aspects of life in America generally, and DC specifically, involve the resolution of competing interests—who among competing bidders gets to buy a house or win a government contract, who among competing applicants receives a job offer or admission to a school. DC government does not choose winners and losers in these competitions. However, to help make sure everyone receives their fair shot, DC may create a first-time homebuyer credit to help make the housing market more fair than it would be if left entirely to market forces. Or, DC may create a CBE contractor preference to make the government contracting market more fair, or targeted recruiting programs to make school and work more equitably accessible. As government programming and market forces evolve over time, the balance between competing interests of these kinds find various resting **equilibria** of varying levels of equity.

Major technological developments routinely affect important equilibria. For example, assistive technologies soften barriers and help people with disabilities to compete for school and job opportunities on fairer ground. Large language modeling tools can help job applicants generate or review resumes before submission, to help people with varying levels of professional writing skill compete more fairly for positions. How might deployment of your proposed tool tip the balance at various important equilibria? Will DC be a more equitable state after deployment than it had been? As you describe your proposed deployment's impacts on equity in DC, be sure that you are answering these questions.

Identify any important equilibria your planned deployment might impact, and **identify** those you expect to impact most heavily. **Describe** the status quo within DC regarding these equilibria, **explain** the impacts you expect your deployment will have on each, and **explain** how you performed this calculation.

To help identify those areas where impacts to equity—both positive and negative—might appropriately bear the greatest scrutiny, **identify** any impacts you anticipate having on equilibria involving classes protected by DC and federal antidiscrimination laws.

One of the most heavily researched and reported aspects of human/AI alignment is the impact of various AI functionalities on fairness, equity, and discrimination. As a result, your vendor likely has done significant testing, and includes within its marketing material detailed discussions of their own equity considerations in design and development. **Identify** any equity or discrimination risks associated with the AI tool identified or disclaimed by your vendor, and **explain** any steps your tool's vendor has taken to address your proposed tool's negative impacts on equity.

Explain how the anticipated impacts you have identified align with DC government's equity planning generally, and any agency-specific equity plans implicated by your planned deployment. **Explain** how you plan to mitigate any misalignments between DC government's equity values and the likely impacts of your planned deployment.

DC AI VALUE ACCOUNTABILITY

DC government's duty to govern cannot be delegated to an AI tool. As powerful as existing tools may be, the highest performing models are—at their core—very large, automated, computational graphs optimized to learn and process hierarchically those meaningful features present in input data. These tools are not "intelligent" in the general sense of the word, they have no conceptual understanding of a knowledge base as a human would, they have no sense of self, and they should always be thought of as tools to assist humans in making decisions. They are **not** adequate replacements for human decision making.

AI tools can be a powerful means of improving government services, but it is important to structure their deployment so that human responsibility for the impacts those tools have remains at all times clear both to those using the tool to serve the public, and to the public being served. This section focuses on **Accountability**, and focuses your attention on two key concepts.

KEY CONCEPTS

- How will you ensure **responsibility** for all government action flows clearly to an appropriate DC government official?
- How will you **measure performance** of the AI tool throughout its lifecycle?

RESPONSIBILITY

How will you ensure responsibility for all government action flows clearly to an appropriate DC government official? AI tools do not make decisions, they process inputs through to outputs. Humans may choose to rely on their outputs. Irresponsible humans may even adopt a policy of categorically deferring to these outputs. But in every case, it is humans—at deployment of an AI tool, in operation of a tool, or in choosing to defer to the outputs of a tool—who retain responsibility for decisions and actions associated with the tool. The following questions are designed to help you assess how well you have structured your proposed deployment to preserve human accountability in a clear, understandable way.

Identify every decision or action your proposed AI tool will be involved in. For example, will your AI tool be involved in hiring, managing high-risk driving of fleet vehicles, strategically replacing equipment before failure? For each decision or action, **explain** what role your AI tool will have in the relevant workflow. Returning to our previous examples, will your tool perform an initial screening of resumes, monitor and score driver performance, predict equipment failures?

While full human review of every task performed by AI would destroy the economies created by deployment of an AI tool, an important aspect of responsible deployment is a considered, human-in-the-loop structure. Respecting this fact, **identify** where in the above workflows human review

or oversight will take place, and **explain** the substance and depth of the review that will be performed.

Explain how you will identify individuals responsible for AI-supported decisions and actions within your agency, how you will notify them of their responsibility, and how you will document their acceptance of that responsibility.

Explain whether and to what extent you intend to provide members of the public interacting with your proposed tool with an opportunity to seek entirely human review of an AI-supported outcome. If such an opportunity would not be cost feasible, or would be inappropriate for some other reason, **please explain**.

MEASURE PERFORMANCE

How will you measure performance of the AI tool throughout its lifecycle? The values alignment process is not complete at deployment. It persists for the entire lifecycle of the tool through to decommissioning and closeout of the relevant program. These questions are intended to help you plan to maintain accountability both for the tool's performance and your use and monitoring.

Identify the performance metrics you will use to track performance of the AI tools involved in your planned deployment, and **describe** how you expect the tools will perform against these identified metrics over their lifecycle.

Apart from the raw performance of those AI tools involved in your deployment, how will you monitor your continued values alignment to detect and remedy **values alignment drift**? For example, consider how a tool that demonstrated a given level of safety and equity at deployment may present a very different safety and equity profile a year into operation. Given the sensitivity of the domains in which you will be deploying these tools, what is an appropriate cadence for reassessment of your alignment with one or all of DC's AI Values? **Explain** your strategy for monitoring and defending against the risk of values alignment drift throughout the full lifecycle of your deployment.

While monitoring performance and values alignment throughout the full lifecycle of the deployment is essential, monitoring means little without remedies available for correction. **Identify** the remedies available to you, should a vendor's performance erode, or prove the cause of values alignment drift. If internal means are available to your agency to course correct on performance or values alignment, **identify** these internal means.

What if something catastrophic happens? While the core technologies behind most AI tools have been around for a very long time, they are being deployed in new spaces and at unprecedented scale. In such situations, even familiar technologies break—dramatically. How will you safeguard against and respond to catastrophic drops in performance or radical values misalignments? Do you have plans to pause use or to decommission the tool? Is it feasible and appropriate to maintain a backup, non-AI system for emergency deployment? **Explain** your plan for catastrophic failure of the planned deployment.

DC AI VALUE TRANSPARENCY

DC government wants residents to be well informed about how their government is working for them, and invites their engagement with agencies on a continuous basis. In recognition of the particular importance of transparency in the deployment of AI tools, Mayor Bowser formed a public AI Values Alignment Advisory Group to review materials like the AI Values Alignment Report you are working on right now. Through your agency's own public engagement efforts, the work of the Advisory Group, and governance materials like this Report, we can all help the public to maintain an active, meaningful understanding of how these new tools are working to further strengthen DC every day. This section of your AI Values Alignment Report focuses on **Transparency**, and focuses your attention on three questions.

KEY CONCEPTS

- What is your plan for **public engagement**?
- How do you plan to **label** AI-generated material?
- How do you plan to **disclose** to residents when they are interacting with non-human agents?

PUBLIC ENGAGEMENT

What is your plan for public engagement? There is no shortage of over-heated rhetoric on AI, both in the popular press and on social media. In any given week you may see it described as a threat on the level of the atomic bomb and a panacea for every inefficiency in human affairs. As a result, the public your agency serves may be laboring under a fundamental misunderstanding of the nature and scope of your deployment, and may come to their first interaction with your planned tools burdened by unrealistic expectations or palpable anxiety. This environment makes a considered, thoughtful public engagement plan indispensable.

Describe your public engagement plan. Have you identified key stakeholders in your planned deployment's success? Have you identified trusted messengers who can help carry your messaging into the communities who will be most affected by the planned deployment? What kind of steps have you taken to identify specific concerns some residents might have about the deployment? How will you plan your messaging to address these concerns. These are some of the considerations that may aid you in planning and revising a fulsome messaging plan.

Explain whether you intend to solicit feedback from residents during the deployment's planned lifecycle. If so, **explain** how you anticipate using or responding to such feedback.

LABELING

How do you plan to label AI-generated material? While generative AI makes up only a subset of current AI tools, some residents may be worried about the provenance of text, images, or video

produced by agencies in the age of AI. And while federal regulators and industry leaders actively discuss what labeling duties the largest producers of synthetic content will carry in the future, as a matter of state-level transparency it is important that DC government agencies properly communicate to residents the provenance of the materials we provide.

Indicate whether your deployment will involve the synthetic generation of text, images, or video. **Indicate** whether this synthetic media will be available to the public. **Explain** any circumstance whereby your deployment may involve releasing synthetic content to the public without its first having been reviewed and approved by agency staff, and **explain** how you plan to ensure all such unreviewed synthetic content will be labeled before release to the public.

DISCLOSURE

How do you plan to disclose to residents when they are interacting with non-human agents? DC government is not in the business of conducting Turing Tests—we have no interest in tricking a person into believing they are talking to another person when in fact they are conversing with a synthetic agent.

Indicate whether your deployment will entail direct communication between members of the public and a chatbot, synthetic agent, or language processing AI tool of any kind. If your deployment does entail such direct communication, **explain** what measures you plan to deploy to ensure the synthetic nature of the agent is meaningfully communicated to the person involved at the start of the interaction. **Explain** how you will monitor the effectiveness of this notice.

DC AI VALUE SUSTAINABILITY

Big, systemic changes in technology have broad and enduring impacts. While AI deployments are likely to have significant positive impacts on the delivery of government services, it is important to be aware of any drawbacks such deployments might entail. This section of the review process focuses on **Sustainability**, and asks you to think about four important questions.

KEY CONCEPTS

- What steps have you taken to ensure this deployment is **cost sustainable** over the long term?
- What consideration have you given to the **environmental impacts** of the deployment?
- How will the deployment impact **job quality** for your existing workforce?
- Have you considered whether the deployment will **displace** existing DC employees?

COST SUSTAINABILITY

What steps have you taken to ensure this deployment is cost sustainable over the long term? Do you remember the first time you used a ride hailing service? How nice the car was, how low its price? Providers of new and disruptive technologies tend to provide excellent services at shockingly low prices as they seek to develop market share. Once the market for a new product has taken shape, providers tend to turn their attention to profitability, and at that stage prices tend to rise. The following questions are meant to help you plan your deployment sustainably with respect to cost.

Explain the billing structure of any AI tool involved in your deployment. Is it token-based? Query based? Do you pay a monthly fee for a specific number of licenses? Are you, instead, buying a product for a flat fee?

Given the billing structure you have identified, **explain** what safeguards, including training, you have in place to help internal or external users avoid accidental or wasteful overuse. **Explain** any enforceable commitments you have in place against unilateral price increases by relevant vendors.

No matter what enforceable commitments you may have in place to help protect against unilateral price increases by relevant vendors, the best long-term protection a consumer of a given product or service may have against unilateral price increase is the portability of their own buying power among competing vendors. Toward this end, **explain** any steps you have taken to safeguard your ability to switch providers in the future without loss in performance. Such safeguards may include strategies to preserve agency queries or inputs to AI tools in a format sufficient to help facilitate transfer learning between competing models in the event an agency elects to change vendors over time.

What will your agency do if the market fails to develop reasonable alternative vendors, or if all vendors of the relevant AI tool raise price beyond a tolerable level? **Explain** any plans you have to continue to satisfy your use case without the use of AI if this class of AI tools is subject to intolerable unilateral price increase.

As the AI industry is still in its relative infancy, the liability implications of an AI deployment may not be as clearly assessable as the risks of more familiar activity. **Indicate** what steps you have taken to assess the financial risk to DC entailed by the proposed deployment. Such steps may include conversation with the Office of Risk Management.

ENVIRONMENTAL SUSTAINABILITY

What consideration have you given to the environmental impacts of the deployment? Specific points in the AI supply chain, including training of large models, entail significant energy costs. These impacts can be significant, and their consideration should form a meaningful part of any decision whether to deploy a given AI tool.

Identify any steps taken by your vendor to mitigate or offset the environmental impact of the development or use of the proposed AI tool. **Describe** any steps your agency intends to take to mitigate or offset unnecessary, resource-intensive use of the proposed tool.

JOB QUALITY

How will the deployment impact job quality for your existing workforce? AI tools work to improve the lives of humans, both those receiving and those delivering government services. The questions that follow are intended to help you consider the impacts your deployment may have on your existing workforce's job quality.

Identify what subsets of your existing workforce will interact directly with your proposed AI tools, and **identify** any tasks currently performed by staff which would transfer to or be supported by the proposed tools. If the tasks performed by the AI tool are not currently being performed at all, so **indicate**.

AI tools do not "think," and their internal processing is reducible to mathematical operations on distributed feature representations within the hidden layers of neural networks. As a result, it can be a daunting task for employees at every level to properly contextualize AI outputs like they would traditionally contextualize the work product of a fellow staff member. **Explain** your strategy for training those team members who will interact with your proposed tool.

While properly deployed AI tools promise to remove drudgery from the day-to-day work of your staff, an ill-considered deployment may leave the existing workforce feeling as if their professional expertise has been subordinated to the output from a machine. **Explain** whether and how your deployment will allow staff members working alongside an AI tool to override its output with their considered, human judgment.

Staff members will likely experience some degree of trepidation as revolutionary technologies are put into service in your agency. Their feedback—both positive and negative—will be an important tool in finetuning the implementation and operation of your proposed AI deployment. **Describe** your internal communications strategy for this deployment, including whether you will solicit and respond to feedback from your existing workforce.

If your deployment will impact or affect the working conditions of current employees within a recognized collective bargaining unit, **indicate** whether you have engaged the Office of Labor Relations and Collective Bargaining to address any obligations you may have concerning the deployment.

DISPLACEMENT

Have you considered whether the deployment will displace existing DC employees? Contrary to clickbait headlines, the machines are not coming for our jobs. AI tools may be very good at performing one or two of a given employee's dozens of core job functions, but even the most advanced tools are very far away from wholesale worker replacement. This section is designed to help you identify and mitigate any risk of displacement of existing DC employees.

Identify any areas where you anticipate any risk of displacement among your current workforce. If any portion of these areas overlaps with recognized collective bargaining units, **indicate** whether you have engaged the Office of Labor Relations and Collective Bargaining to address any obligations you may have concerning the deployment.

Explain your strategy for minimizing the risk of such displacement, and **explain** your strategy for mitigating any potential negative impacts for any current employees ultimately displaced by the deployment. **Explain** the cost savings, quality improvements, or other benefits you anticipate from the deployment and how you weighed those benefits against the costs of the risk of the displacement of current employees. Any calculation of cost you provide should include all ancillary costs of the deployment, including things like training existing staff to use the AI tools, measurement and oversight of the tool's performance, anticipated price increases, and costs for decommissioning or transitioning among AI tools.

Explain any efforts made by the relevant vendor to mitigate any negative impacts the deployment may make upon DC's labor market. These may include, for example, investments in training opportunities for DC residents, or commitments to hire DC residents.

DC AI VALUE
PRIVACY

AI tools exist to process information. They can offer no benefit without information. In deploying an AI tool, agencies must consider both the planned tool's need for information and the privacy interests of residents, visitors, and DC government employees.

OCTO's privacy review process is designed to help your agency identify the information at issue in a proposed AI tool deployment, document its planned uses in that deployment, and verify that the you have instituted proper safeguards to protect the privacy of residents, visitors, and DC government employees. To do this, think about the following six questions.

KEY CONCEPTS

- **What information** is involved in your planned deployment?
- **Where** will the information reside?
- How will that information be **used**?
- How will the **technical aspects** of your deployment promote privacy?
- How will the **legal terms** of your deployment promote privacy?
- How will you **notify** people whose privacy is impacted by your deployment?

KEY DEFINITIONS

It is important to understand that privacy interests can be more sensitive in some settings than in others. To aid you in preparing this section of your AI Values Alignment Report, consider the following **definitions**, helpful for contextualizing the varying sensitivity level of various kinds of information.

Data: *According to DC's Data Policy, Data means a subset of information, whether quantitative or qualitative, that is regularly maintained by, created by or on behalf of, and owned or licensed by a public body in non-narrative, alphanumeric, or geospatial formats. Data are an asset independent of the systems or formats in which they reside.*

Sensitive Data: *Data and metadata are sensitive if they pertain to an individual in a sensitive domain; are generated by technologies used in a sensitive domain; can be used to infer data from a sensitive domain or sensitive data about an individual (such as a disability-related data, genomic data, biometric data, behavioral data, geolocation data, data related to interaction with the criminal justice system, relationship history and legal status such as custody and divorce information, and home, work, or school environmental data); or have the reasonable potential to be used in ways that are likely to expose individuals to meaningful harm, such as a loss of privacy or financial harm due to identity theft. Data and metadata generated by or about those who are not yet legal adults is also sensitive, even if not related to a sensitive domain. Such data includes, but is not limited to, numerical, text, image, audio, or video data.*

Sensitive Domains: *are those in which activities being conducted can cause material harms, including significant adverse effects on human rights such as autonomy and dignity, as well as civil liberties and civil rights. Domains that have historically been singled out as deserving of enhanced data protections or where such enhanced protections are reasonably expected by the public include, but are not limited to, health, family planning and care, employment, education, criminal justice, and personal finance. In the context of this framework, such domains are considered sensitive whether or not the specifics of a system context would necessitate coverage under existing law, and domains and data that are considered sensitive are understood to change over time based on societal norms and context.*

WHAT INFORMATION IS AT ISSUE

What information is involved in your planned deployment? This section of your AI Values Alignment Report documents your breadth and depth of review of the potential privacy impacts of the AI tool deployment. As you identify various classes of structured data and unstructured information in this section of your AI Values Alignment Report, be sure to note **how it is collected**, **from what sources**, **why** it is being collected, and **what technologies are used** in its collection.

It is important to think broadly not just about what information the agency intends to introduce into a given AI tool, but what information might be inadvertently introduced by users. It is important, to think not only about information introduced into a tool by internal users like DC government employees, but information introduced by external users—members of the public who converse with an internet-facing chatbot, for example—as well. It is important, too, to think not only about information input into an AI tool, but to think about information that might be output by that tool, including any information the AI tool may draw from inputs for the purposes of training itself.

Think about **structured data**—the kind of data that resides in spreadsheets and databases—and how it relates to the privacy interests of stakeholders. **Identify** every data set the agency intends to share with the vendor or input directly into the AI tool as part of this deployment. For every data set, **indicate** whether the data set is currently in DC's Enterprise Data Inventory, and **indicate** its current EDI Classification. **Identify** any personally identifying data fields contained within the data sets, and **explain** what steps the agency will take to protect the privacy interests of those personally identifiable by the data elements at issue. Is an identified data set subject to an existing retention schedule? If so, **explain** how your planned use relates to the relevant retention schedule.

Identify all data elements subject to any statutory or regulatory privacy protections. These may include, for example, the Health Insurance Portability and Accountability Act, the Family Educational Rights and Privacy Act, the Duncan Ordinance, the Fair Credit Reporting Act, the Fair and Accurate Credit Transactions Act, the Gramm-Leach-Bliley Act, the Rights to Financial Privacy Act of 1978, the Genetic Information Nondiscrimination Act of 2008, the Children's Online Privacy Protection Act and Rule, the Telephone Records and Privacy Protection Act, the Telephone Consumer Protection Act, the Driver's Privacy Protection Act, the Video Privacy

Protection Act of 1988, the National Labor Relations Act, or the Electronic Communications Privacy Act of 1986.

For every data element identified as subject to a statutory or regulatory privacy protection, **list** the relevant statutory or regulatory requirements, and **explain** how the planned deployment will ensure these requirements are met throughout the deployment's full lifecycle. **Identify** any data which appears to meet the definition of **sensitive data**.

Think about **unstructured information**—the kind of information contained in natural language, in pictures, in audio files, in video—and how it relates to the privacy interests of stakeholders. AI tools are capable of meaningfully ingesting a much broader class of information than were earlier tools. Existing AI tools can extract rich information and draw meaningful inferences—including personal identification—from unlabeled photos, videos, natural language text, and more. To ensure adequate pre-deployment privacy review, it is important to account not only for a planned deployment's use of structured data, but unstructured information as well.

Describe the kinds of information that will be shared with the AI tool. This description will likely be less specific and less exhaustive than the description of structured data already provided, as the content of natural language and other such inputs are by their nature less discretely describable than structured data elements, but a meaningful degree of breadth and specificity may be implied by the AI tool's intended use. Is this tool intended to ingest case files, student exam answers, closed-circuit camera footage, pictures of tumors, or resident questions about trash pickup? The answers to these questions should inform your response.

List any of the above-identified kinds of informational inputs which are reasonably likely to include personally identifiable material, and **explain** what steps the agency will take to protect the privacy interests of those personally identifiable by the information reasonably expected to show up in inputs during the life of the tool. Is this information subject to an existing retention schedule? If so, **explain** how your planned use relates to the relevant retention schedule.

List any of the above-identified kinds of informational inputs which are reasonably likely to include material subject to a statutory or regulatory privacy protection, and **explain** how the planned deployment will ensure these requirements are met throughout the deployment's full lifecycle. **Identify** any information which appears relevant to **sensitive domains**.

WHERE WILL INFORMATION RESIDE

Where will the information reside? Depending on the structure and configuration of the specific AI deployment under consideration, the various categories of information above may reside entirely on DC government-owned physical hardware, on vendor-owned hardware configured as virtual machines dedicated to DC government use, in environments directly managed by vendors, and more.

Before OCTO can properly analyze the impacts a given deployment might have on the privacy interests of everyone involved, it will need to understand where the information at issue in the proposed deployment will reside and how it will move. In considering the location and

movement of information, remember to include both an AI tool's inputs and any outputs it derives from those inputs.

For all information—both structured data and unstructured information—**identify** where that information currently resides. For illustrative purposes, a given structured data set may currently reside in OCTO's Data Warehouse or Data Lake, case files may currently reside in existing caseworkers' folders as part of an enterprise cloud instance, closed-circuit video footage may reside on a similar enterprise cloud solution or it may be stored locally on hardware owned by DC government.

How will this change with the planned deployment? **Specify** whether the information at issue in this deployment will remain in its current location to be processed by an instance of an AI tool that will live in the same environment, or whether any information will be copied from its existing environment and introduced into a new environment where the AI tool at issue operates. Will the deployment change the locus of the definitive record copy of any information? If so, please **specify**.

USE RESTRICTION

How will the information be used? Not all uses of information bear identically upon people's privacy. Some different uses of the same piece of information would broadly be considered objectionable. For example, it is appropriate to use a person's credit score in evaluating their loan application, but it would be inappropriate for a court to use a person's credit score to determine their credibility as a witness. Some uses of private information may be considered objectionable because that use might put the confidentiality of private information at risk. For example, if a mental health report evaluating a person were introduced into a large language model, that information may be retained by the vendor for training or evaluating the performance of their tool.

Explain how the agency will use the information identified above as part of its use of the planned tool. **Identify** every use of the relevant information that the vendor will be permitted to make. **Explain** why the vendor will be permitted to make such use.

Many vendors—including those offering AI tools—utilize third parties in their delivery of products and services. For privacy purposes, the introduction of third parties can bring with it unforeseen privacy risks. For example, a provider of AI tools may expressly limit the scope of their use of information to appropriate boundaries, while affirmatively carving out and declining to limit at all the uses third parties might make of that same information. As a result, OCTO cannot make a complete review of the privacy impact of a deployment without verifying what limitations will be placed on third party use of information. In completing this section of your AI Values Alignment Report, **identify** every use of the relevant information that third parties will be permitted to make. **Explain** why third parties will be permitted to make such use.

TECHNICAL PROTECTIONS

How will the technical aspects of your deployment promote privacy? As you've answered these first three questions, you've been compiling and considering those aspects of your planned AI deployment which might give rise to specific privacy risks. Both this question, "how will the technical aspects of your deployment promote privacy," and the one that follows, "how will the legal terms of your deployment promote privacy," examine what tools you have in place to help mitigate those risks.

Privacy considerations should inform AI tool design, build, deployment, and operation through to decommission and retirement. Such considerations can take the form of data minimization, verification of data provenance, accuracy, and currentness, thoughtful data retention scheduling, requiring data processing to occur in agency-controlled environments, role-based access rules, access logging, and the incorporation of **privacy-enhancing technology**. This last term may warrant special attention.

President Biden's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence provides the following definition, which may prove useful as you consider how the technical aspects of your deployment will promote privacy:

Privacy-enhancing technology: *means any software or hardware solution, technical process, technique, or other technological means of mitigating privacy risks arising from data processing, including by enhancing predictability, manageability, disassociability, storage, security, and confidentiality. These technological means may include secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic-data-generation tools. This is also sometimes referred to as privacy-preserving technology.*

It's highly unlikely that the vendors and third parties involved in the development and marketing of the AI tool you propose to deploy are unaware of the concepts of "privacy by design" and "privacy-enhancing technology." Likely, such considerations are a key component of those vendors and third parties' sales strategy. Bearing this in mind, **identify** any steps taken by vendors or relevant third parties to mitigate privacy risks the proposed deployment entails, and **attach** to your AI Values Alignment Report any vendor- or third party-provided explanations or summaries of such steps.

Complementary to or wholly separate from those steps taken by vendors and third parties to mitigate relevant privacy risks, your agency can make strong, positive impacts on a given AI deployment's privacy posture. For example, privacy risk mitigating techniques available to your team may include the application of privacy-enhancing technologies to data or information before its introduction into an AI tool, utilization of AI tools within controlled environments, or limitation on access to AI tool outputs to prevent external serial queries from being used as a means to re-identify de-identified outputs. **Identify** what steps you plan to take in the design, build, deployment, and operation of your proposed AI tool through to decommission and retirement to promote privacy.

LEGAL PROTECTIONS

How will the legal terms of your deployment promote privacy? Currently, there is no comprehensive, general data privacy regulation at the federal level. Even in countries with the most comprehensive, general data privacy laws, the public are heavily dependent on vendor privacy policies, privacy and confidentiality terms in agreements, and contractual limitations on permissible uses of data to safeguard people's legitimate privacy interests. It is for these reasons that this section of your AI Values Alignment Report should catalogue—with attachments—the various policies and contractual terms relevant to the information put at issue by your planned deployment. **Please note** it may be especially important for your agency counsel to review this portion of your AI Values Alignment Report before final submission.

Even when the relevant parties are not in privity of contract, **privacy policies** are generally enforceable against those parties issuing them, as federal and state-level regulatory authorities routinely prosecute their violation as deceptive or unfair trade practices. While parties are free to change their privacy policies from time to time, in many cases, courts have found that information gathered at a given point in time remain subject to the privacy policy in effect at the time the information was gathered. As a result, privacy policies are a key consideration in a thorough privacy review.

Does your agency have a general privacy policy? Does it have one or more privacy policies tailored to specific information gathering or processing activities? Please review your agency's existing and former privacy policies and **attach** to your AI Values Alignment Report those policies relevant to the structured data and unstructured information identified above. If specific policies are only applicable to specific subsets of that information, **explain** which privacy policies are applicable to which classes of information.

Attach all vendor privacy policies relevant to your planned deployment. In performing this analysis, consider requiring the vendor to perform this search and to produce all relevant privacy policies, as AI vendors often have different privacy policies for their public tools and their enterprise tools. If the attached privacy policy is not applicable to all information identified thus far in your AI Values Alignment Report, **explain** the applicability of each policy attached. Does your vendor have the right to update the applicable privacy policy during your relationship? If so, **explain** under what circumstances the vendor may update the applicable privacy policy, and **explain** whether your agency is entitled to receive notice or to contest the change.

Many vendors expressly limit the scope of their privacy policies to exclude the activities of third parties, even where third parties are essential to their performance. To ensure OCTO has a complete record from which to assess the relevant privacy policies, **attach** all third-party privacy policies relevant to your planned deployment. In performing this analysis, consider requiring the vendor to perform this search and to produce all relevant third-party privacy policies. If you have attached third-party privacy policies, **explain** the applicability of each policy attached. Does the third party have the right to update the applicable privacy policy during your relationship? If so, **explain** under what circumstances the third party may update the applicable privacy policy, and **explain** whether your agency is entitled to receive notice or to contest the change.

While vendors may rely on general privacy policies to govern the relationship between their AI tool and broad classes of unidentified users, vendors very often directly negotiate **privacy**

contract terms to govern privacy concerns that arise within the scope of a specific agreement. Privacy contract terms are useful both for promoting the regulatory compliance of those persons party to a contract, and for indirectly protecting the privacy interests of people not party to the contract at issue. These privacy contract terms ordinarily supersede the terms of the vendor's general privacy policies anywhere the two conflict. While the privacy contract terms applicable to a specific agreement are generally easier to locate than general privacy policies, it is important to check for terms included directly in contracts, terms implied by statute or procurement regulation, as well as terms incorporated by reference. **Attach** all privacy contract terms applicable to your planned deployment, and **explain** their scope of application.

Closely related to privacy terms, **confidentiality terms** protect the parties' broad interest in maintaining the confidentiality of non-public information which may be exposed or exchanged among the parties. While confidentiality terms often protect more broad interests than just privacy, they often offer additional protection for information bearing on the privacy of interested people. **Attach** all confidentiality terms applicable to your planned deployment, and **explain** their scope of application.

Disclosure of private or confidential information is not the only means by which legitimate privacy interests might be critically undermined. Consider, for example, a hypothetical vendor who obtains incredibly private healthcare information for the purposes of providing medical billing services to a healthcare provider. Even if that vendor never disclosed that information externally, how upset would patients be to learn that the vendor was using that same data to support an entirely separate line of business targeting the patients?

Because such internal use of private information can cause privacy harms to people, it is important that agencies use **permissible use terms** to restrict vendors' and third parties' use of information. Will you permit use of information for training the vendors' or third parties' AI tools? Will you permit use of information for vendors' or third parties' marketing purposes? Will you require notice of such usage to your agency. To the impacted people? To enable OCTO to evaluate this aspect of your proposed deployment, **attach** to your AI Values Alignment Report all permissible use terms applicable to your planned deployment, and **explain** their scope of application.

Information in your possession may already be subject to contractual restrictions. Please perform a thorough internal review of the information identified above for any **pre-existing restrictions on use or sharing**. If any such restrictions exist, **attach** all language restricting your use or sharing of the relevant information, and **explain** the scope of application.

Of course, contractual rights mean little without appropriate **contractual remedies**, and contractual remedies mean little if they are subject to inappropriate limitations. Do the relevant agreements contain damage caps or restrictions on available remedies? Are they carved out for breaches of privacy, confidentiality, or permissible use terms? Carefully review the agreements relevant to your proposed deployment, **attach** any terms bearing on available remedies in the event of breach of privacy, confidentiality, or permissible use terms, and **explain** their scope of application.

NOTIFICATION

How will you notify people whose privacy may be impacted by your deployment? Notification is not always appropriate, especially where too specific or too fulsome a notice would undermine investigations, oversight, enforcement, or permit better financed or more sophisticated parties to reverse-engineer AI tools or otherwise game their operation in an adversarial way to produce results that undermine DC's AI Values. This particular risk may be especially likely where an AI tool is intended to play some role in an adjudicatory process, an enforcement process, an investigatory process, or in any process affecting the rights or responsibilities of the public in zero-sum or competitive contexts. However, in most cases, most of the time, it is an advisable practice to advise people when a new process will impact their legitimate privacy interests. Toward that end, **explain** any privacy notice processes you intend to implement as part of your proposed AI deployment.

DC AI VALUE
CYBERSECURITY

DC's AI tools must be deployed in a way that promotes the confidentiality, integrity, and availability of DC's information technology assets. Effective cybersecurity means robust encryption, access controls, and regular security audits to protect against unauthorized access, data breaches, unintended sharing, and malicious tampering. It means maintaining up-to-date software, conducting continuous monitoring for vulnerabilities, and conducting regular and thorough risk assessments.

There are cybersecurity risks peculiar to some AI tools. Some AI tools are subject to adversarial attacks enabled by the way neural networks learn and classify data. This means we have to pay special attention to the cybersecurity implications of incorporating these emerging technologies into DC's information technology assets.

It is important to note that you may need to redact some portions of this section of your AI Values Alignment Report on recommendation of the Chief Information Security Officer, if such redaction would promote the security of DC's information technology assets. Additionally, given the technical nature of this portion of the review, you may want to be sure your agency's Chief Information Officer is closely involved in developing this portion of your AI Values Alignment Report.

KEY CONCEPTS

- How will the AI tool be **configured** in your deployment?
- With what systems will the tool **interface**?
- What **data** will it touch, and where?
- Will the tool be **public facing**?
- How will the tool be actively **supported** throughout its lifecycle?
- What tools will you use for **risk management** throughout the lifecycle?

CONFIGURATION

Any information technology can become a cybersecurity risk vector if it is not properly configured. As you complete this section of your AI Values Alignment Report, think carefully about and provide detailed information on the proposed AI deployment. **Explain** in technical detail the underlying architecture, scalability, and data requirements of the proposed AI deployment. **Explain** any business process reviews associated with the deployment. If the solution is hosted, **explain** in technical detail the host provider and the responsibilities of all parties involved in the hosting, both in the form of a Responsible, Accountable, Consulted, and Informed (commonly, "RACI") format, and in a lifecycle matrix format. **Explain** in technical detail the AI tool's logging options and monitoring capabilities. **Explain** how the logging capabilities of the AI tool support general troubleshooting, and **explain** in technical detail how those logging capabilities will support any potential security investigations.

INTERFACE

With any technology solution understanding and documenting how a system is built is only half the equation. To finish the job, you must identify, document, and validate the additional systems it will need to interface with to provide the desired outcomes. Toward that end, **identify** all systems and applications that the AI tool will integrate with, and **explain** in technical detail the nature of those integrations. For every integration, expressly **identify** whether these integrated systems or applications are internal or external to DC's information technology environment. **Explain** in technical detail the technical, regulatory, and administrative controls that your team will enforce, and how you will monitor such compliance throughout the full lifecycle of the deployment. For every integration that will provide access to DC government data, **identify** what data sharing agreements you intend to enter into.

DATA

Data is a key information technology asset for every entity, and an indispensable consideration in both privacy and cybersecurity reviews. When bad actors target entities for attack, access to or interference with data is among their most common goals. To ensure the cybersecurity section of your AI Values Alignment Report is properly self-contained, please incorporate in this section of your report the descriptions of relevant data sets and your AI tools' proposed access thereto which you will have already provided in the Privacy section of your report.

PUBLIC-FACING AI TOOLS

Even in the absence of AI, any public-facing, or internet exposed system should be classified high risk from a cybersecurity perspective. This exposure makes some portion of the relevant system accessible to bad actors the world over. When the exposed system incorporates AI functionality, additional security considerations arise. AI-focused adversarial attacks, for example, can make use of paired inputs and outputs to functionally approximate the inner workings of a given neural network. This approximation can then be used as part of a strategy to manipulate outputs from that system. Or, a bad actor might use their public access to the input layer of an AI tool to introduce command language encouraging the AI tool to take further action on behalf of the attacker. Or, an actor might overrun a public-facing AI tool with queries in an effort to multiply costs or interfere with functionality.

These known AI cybersecurity risks, as well as those yet to be discovered, make the decision to configure an AI tool as public facing a significant one. In your AI Values Alignment Report, **identify** whether any AI tool involved in your deployment will be public facing. **Identify** whether the public will interact with the AI tool directly, as with a chatbot or AI-powered dashboard? **Identify** whether the public will obtain outputs directly from the AI tool, or if there will be a DC government human in the loop. If anyone not currently employed by DC government will have access directly to the input layer of an AI tool, **explain** why such access is necessary. **Explain** any technical safeguards you will incorporate into the AI deployment, including hard numerical limits on the number or complexity of queries, and any guardrails like scrubbing, filtering, or pseudonymizing of outputs.

ACTIVE SUPPORT

The modern information technology market relies on active support—usually provided by the original seller of a given software tool—to seek out vulnerabilities, identify solutions, and push out updated versions of software tools which are more resilient from a cybersecurity perspective. Over time, software tools age and are replaced by whole new tools. Support for older tools often continues for a time, but eventually software providers cease providing active support for outdated products, leaving their final version unpatched for any vulnerabilities discovered after that point in time. There is significant cybersecurity vulnerability in continuing to use software tools that are no longer actively supported.

Just as values alignment is subject to drift overtime, the robustness of a given AI deployment's cybersecurity posture changes with its level of active support. **Explain** the relevant AI tool providers' current and future support obligations. **Explain** your plans for managing the AI deployment in the event that the provider ceases to provide adequate support. Will you discontinue use of the relevant AI tools? Will you arrange for your agency or some third party to assume responsibility for active support throughout the remainder of the AI deployment's lifecycle? If you intend the latter, **explain** how you will ensure access to resources like source code, model weights, percept histories, or any other asset necessary to transition this support obligation away from the provider, and **attach** any contract language escrowing such materials, or protecting your agency's legal rights to take such action.

Even during periods of adequate active support, change control from one version release to another is of serious importance to performance and continuity of operations. **Explain** your change control strategy concerning regular general releases and agile releases to address serious vulnerabilities or functional issues. **Explain** the relevant tools' planned standard patching schedules and maintenance windows.

RISK MANAGEMENT

All systems—including those powered by AI—carry risk. The difference between a responsible entity and an irresponsible one isn't the absence of risk, but the presence of and adherence to mature risk management strategies. And while mass-market AI is still an emerging technology sector, there are a number of powerful risk management tools available to you and your agency. Consult OCTO's AI/ML Adoption and Usage Guidelines for DC government. Consult the National Institute of Standards and Technology's cybersecurity risk management framework and AI risk management framework. Consult the Cloud Security Alliance's Cloud Control Matrix. Consult FedRAMP or StateRAMP standards if the relevant AI tools are marketed as FedRAMP or StateRAMP compliant. If your team is unfamiliar with risk management generally, you can consult with DC's Office of Risk Management. Your team will be well-supported in this undertaking.

In this section of your AI Values Alignment Report, you will want to provide a detailed **explanation** of your internal risk management processes, and **explain** how you will apply your general approach to this specific AI deployment. **Explain** how you will map, measure, manage,

and govern risks associated with the AI deployment. **Identify** the controls you will use to identify and mitigate risks arising from the AI deployment. **Explain** how you will document internal temporary acceptance of a given risk, and what remediation steps you will apply in the period following temporary acceptance. **Explain** your proposed incident reporting and response protocols. **Explain** how your mapping, measurement, management, and governance of these risks will account for risks arising outside DC's information technology environment, including, for example, those risks arising within a cloud service provider's environment. **Explain** in technical detail any disaster recovery rating associated with any tool or platform that is a part of your AI deployment.

CONCLUSION

This is an exciting time to be working in government technology. AI tools promise greater efficiency and fairness, and may prove powerful resources in delivering the highest level of government services to DC residents. By undertaking the kind of thorough AI Values Alignment described in this Handbook, your agency can help make sure that DC residents realize the fullest possible return on these emerging technologies, while at the same time promoting those core values that make DC the best place in the world to live and work.