PALANTIR RESPONSE TO TRADEWIND RESPONSIBLE AI RFI

Engendering Warfighter Trust through Operationally Responsible AI

Introduction

Palantir USG, Inc., a wholly owned subsidiary of Palantir Technologies Inc. (collectively, "Palantir USG", "we", or "our") presents this response to the Tradewind OTA RFI on Responsible AI ("RAI") Expertise, Products, Services, Solutions, and Best Practices. Palantir USG is a software company that develops, manages, and sustains highly configurable commercial software solutions. This includes the implementation of software factories via our commercial software solutions that provide end-to-end AI/ML model development, validation, and deployment in operational contexts. Our customers can build new models within our platform or bring in third party models into the solutions we provide. Our solutions have been used to support mission planning and battlefield analysis in a number of operational environments.

We have extensive experience addressing critical enterprise-scale challenges across Government and Industry, modernizing their businesses, processes, and technologies, and enhancing the quality of their decision making using AI/ML. As with all of our product development and deployment, our approach to Responsible AI is grounded in a corporate tradition of treating the societal and ethical implications of our work as a first-order concern, on par with the challenges and importance of building world-class technologies. We identify that the most critical challenges of RAI implementation are rooted in the full systems and operational contexts in which AI serves *not* as a comprehensive panacea, but rather as a narrow-purpose tool directed at supporting and enabling specific, typically human-centric, outcomes.

With respect to the RAI tenets, our experiences are therefore most heavily weighted towards providing AI solutions that integrate data, analytics, and AI into decision-making and operations that enshrine the AI Ethics Principles into operator workflows, and therefore are directed at underscoring the tenet of *Warfighter Trust*. Throughout our response below we will highlight critical approaches spanning the full AI lifecycle — from program conception to development, deployment, and management — that place the trust of the operators interacting with and using AI as central considerations in ensuring the responsible, ethically defensible use of AI in Defense contexts.

Beyond this response, we are happy to act as a thought partner to the Government and share best practices on the operationally-relevant deployment of RAI/ML, some of which are outlined below in response to specific RFI categories and key questions.

Requested Information – Categories and Key Questions

Our response is intended to address facets of Category 1 (Assessments and Governance), Category 2 (Workforce training and management), Category 3 (Products, tools, and services across development and/or acquisition lifecycles), Category 4 (Research and development), Category 7 (Risk management), and Category 8 (Military ethics and safety).

Category 1 (Assessments and Governance)

In order to mitigate or address the risk of unintended or potential harms of AI systems, we recommend a program assessment framework that commences with a clear-eyed recognition that all phases of AI conception, development, and deployment are subject to inherent limitations (e.g., personal bias, limited understanding) and the extrinsic failings (e.g., poor engineering practices) of their designers and operators. When employing AI/ML tools, it is essential to examine not only the viability and efficacy of the technologies themselves, but also the validity and merits of the purposes for which those models are being deployed. There are critical components in this recommended assessment framework that successively build upon each and collectively work to constitute AI systems that are operationally viable and support trustworthy warfighter workflows:

1. **Don't solve problems that shouldn't be solved.** Problem Selection should serve as an essential threshold consideration in any AI system undertaking. It begins with an acknowledgement that there are certain problems that simply should not avail themselves to AI interventions, for a number of reasons. Some problem types are ill-suited for AI/ML approaches because they violate legal and community norms. For example, research applying facial recognition technology to predict sexual orientation from facial features implicates sensitive category inferences that are morally, scientifically, and legally questionable. Where proposed technology interventions involve complex social behavior, potentially negative feedback loops, and subtly compounding disparate impacts, researchers and technologists should give pause to consider whether adequate mitigation measures can be instituted to rein in risks. Furthermore, AI program administrators should view as their responsibility the need to evaluate, as a threshold matter, whether mathematical optimization and other synthetic machine-driven AI methods are simply ill-equipped to capture fundamentally irreducible qualitative phenomena carrying significant impacts.

2. **Select Features Responsibly.** Predicated on the assumption of a valid and defensible AI application, the second operational step in the assessment framework relates to the selection of model features that may be used to develop AI/ML algorithms. We believe that context matters and is essential to determining whether specific types of information should be used in AI. For example, it is one thing to seek to analyze whether sensitive categories such as race or gender linked to genetic attributes correlate with certain disease prevalences. It is quite another thing to seek an association between those same immutable personal attributes and one’s likelihood to perform with valor on the battlefield. Where AI seeks to model (and predict) outcomes with real impact on peoples’ lives but with no ostensible or causally justifiable association with their sensitive personal attributes (e.g., race, gender, ethnicity, religion, sexual orientation, age, family status, etc.), an active effort may be required to minimize model reliance on

related features, in the form of explicit sensitive data categories, as well as implicit or proxy features. Where such sensitive categories are implicated in AI programs either by necessity or choice, their inclusion should be identified and, where appropriate, contextually justified.

3. **Methodically Assess and Address Sample Bias.** Any credible AI effort must grapple with the "garbage-in-garbage-out" problem. Stated differently, no matter how rigorous the design of an AI application is, if the data used for training and development is irredeemably flawed, the system output will almost certainly be compromised as well. This issue is even more pronounced where training data reflects systemic or other institutionalized forms of bias. The implementation of any AI system requires a clear assessment of the fidelity, quality, and representativeness of the data upon which its models are built and trained. These evaluations must seek to understand how biases in data collection and aggregation may encode and perpetuate disproportionate practices in the real world. If, for example, population groups are systemically under- or over-represented by virtue of skewed data gathering practices, suitable techniques must be employed to re-balance the data through means such as re-sampling, collecting new data, applying adjusted weight classes, or modifying the algorithm itself to explicitly nullify or minimize the known bias. At minimum, where such biases exist, they should be studied, understood, and appropriately documented so that decision makers — including warfighters — can, with full awareness, make competent evaluations of the tradeoffs incurred by accepting some measure of bias (which may be legitimate or acceptable in specific contexts) in exchange for optimizing on other evaluation parameters.

4. **Carefully Monitor Outcomes, Understand Equitability Assessment Tradeoffs.** Beyond addressing feature selection and sample bias, which may contribute to disproportionate or skewed applications in the mechanics of a given AI system, a responsible approach to AI must also deal with the risks associated with disproportionate impacts in that system's empirical outcomes. Evaluating how AI outcomes may be skewed across different population segments of concern, including (for impacted human populations) groups representing sensitive or vulnerable categories (e.g., age, gender, race, ethnicity), can assist in further algorithmic refinements aimed at optimizing outcomes against appropriate evaluation parameters and other intentionally desirable measures (e.g., equitability, fairness, accuracy, etc.).

5. **Ensure Auditability, Explainability, and Interpretability.** We advise that responsible AI uses must enforce accountability through a number of discrete measures and techniques. Auditing in the form of in-process system records and audit logs should be employed to ensure that processing events contributing to specific outcomes can, at later stages, be interrogated and documented as

needed. AI models must also be explainable to various relevant stakeholders, subject to legal, doctrinal, and normative expectations. For all users and stakeholders, including program supervisors, commanders, decisionmakers, government oversight authorities, and warfighters, explanations of the underlying mechanics and decision-making logic of the system must often be presented in reasonably accessible terms, which may include documentation of the features a model uses, how those features are weighted, how it was trained, and how the algorithm decides what is important and what is not. These explanations should ideally be provided in terms that are accessible to a reasonably competent lay audience (i.e., in terms that do not require a graduate degree in artificial intelligence, machine learning, and adjacent fields to understand). Finally, interpretations of models should, in alignment with the preceding “Carefully Monitor Outcomes, Understand Equitability Assessment Tradeoffs” principle, provide a clear assessment and tradeoffs made and consequences of choosing certain optimization parameters over others. (See Category 4 comments for additional details.)

6. **Wherever Possible, Keep Humans in the Loop.** Generally speaking, AI is both most effective and most defensible when employed to assist human execution and decision making rather than replacing it. Particularly in applications that carry significant impacts on individuals' livelihoods and well-being, the limits of AI must be acknowledged as a heuristic for determining the right level of human intervention to ensure moral agency and culpability. In reality, no model, no matter how well-engineered, is going to be 100% accurate. All applications are likely to produce some measure of false assessments. Keeping humans in the loop in processing recommendations or outcomes (as per the example above) helps to provide a critical check on the applications of AI technology. The decisions to remove human oversight or augmentation for specific AI applications should be approached with the utmost gravity and clear acknowledgement of the implied tradeoffs and accountability chains of command to be examined if and when the system fails.

7. **Promote Multi-Stakeholder Engagement.** Finally, we submit as a critical operational principle for responsible AI systems that programs should endeavor to incorporate broad stakeholder involvement in building, deploying, and overseeing applications. While consistent adherence to the above specified principles is essential to the success of an AI program, if the measures taken to comply with these principles are not (to the degree permissible in any given context) legible to a broader community of interested and affected parties, the AI program may be viewed as operating under a veil of secrecy. The corresponding black-box opacity can contribute to suspicion and distrust of the instituting organizations and its practitioners, especially warfighter. In this regard, attempts to engage with institutional stakeholders spanning the domains of policy,

compliance, legal, technical, administrative, and operational competencies are critical from the outset.

Each of the above phases of the AI systems assessment framework requires ownership by appropriate stakeholders. Any break in the chain of AI systems development, deployment, and management can lead to compounded failures and adverse consequences, including fundamental erosion of warfighter trust. When trust in AI systems has been degraded, it is difficult to be reestablished and ultimately can lead to the complete breakdown of an AI application.

Category 2 (Workforce training and management) and Category 8 (Military ethics and safety)

While AI comprises a novel field of advanced technology applications, the critical contexts of military operational workflows ultimately should determine and steer how AI is developed and situated in deployment scenarios. This implies that AI ethics should be similarly situated within well-established ethical, normative, and legal frameworks, including the modern Just War tradition, which provides a comprehensive framework for thinking about issues of war and peace through the prism of established International Humanitarian Law (IHL) and international institutions. This tradition provides critical guiding considerations and principles (e.g., necessity, proportionality, etc.), but also must be interpreted and — to some extent — adapted to deal with the novel aspects of modern warfare. *Jus ad bellum* and *jus in bello* principles should therefore serve as a starting point for applied ethics considerations in the development and use of AI systems.

Warfighters are already steeped in this ethical tradition and it can be used as a steppingstone to address new considerations that arise in AI. Its use carries the benefits of immediate legibility and (perhaps more critically) an appropriate framing of AI as an operationally applied technology that is used not in ‘toy’ or laboratory settings, but in real world environments in which all the standard vulnerabilities, risks, and challenges already exist, and where AI systems as tools supporting warfighters should be further adapted.

We advise, therefore, that AI ethics training begin with the premise that AI constitutes a set of warfighter tools, not entirely dissimilar from other tools (including weapons). Those tools should then be situated in the contexts of their applications in the literal and metaphorical hands of operators and warfighters. All of the traditional systemic constraints, vulnerabilities, risks, and rewards should therefore be assumed as applicable, and therefore the existing military ethics frameworks provide a critical starting point for evaluating the ethical valence of AI systems (rather than presuming AI systems as constituting an entirely novel operating environment in need of an ethics *de novo*). From there, ethics training programs should provide trainees with a clearer understanding of what, if anything, is epistemically and ontologically distinct about so-

called *artificial intelligence*, whether it is at all a form of cognition, judgement, or decision-making in the manner that applies to existing moral theory, or if it functions more as a tool in the hands of operators. The extent to which those questions can be addressed and, correspondingly, trainees can begin to identify the reasonable limits of AI systems as distinct from moral agency in the realm of human intelligence, cognition, judgement, and culpability, the better able they will be to understand the supporting role of AI systems and where those systems fit into the military traditions with which they are already familiar.

Category 3: Products, tools, and services across development and/or acquisition lifecycles

A methodical approach to integrating RAI practices and requirements into system design, development, deployment, and management should build upon the considerations outlined in the previous category's response, specifically a fundamental respect for the principles of the Just War tradition. In practice, this translates into proactively calling out the necessity and proportionality of tradeoffs that might be required in the construction of AI systems that play a role in enabling violent action and sometimes even death.

We advise a comprehensive framework for managing AI models from data integration and management to model management and support for operational user workflows, including embedding RAI considerations into the UI/UX. This may include the following elements:

- **Data Integration and Management features for:**
  - Integration and management of all data relevant to AI workflows.
  - Preparing data in uniform, open formats for use by AI models.
  - Orchestrating third-party services, such as data labeling services.
  - Providing the engine for designing and curating data to train AI models.
  - Enabling interoperability with multiple systems via open APIs.
- **Model Management and Orchestration capabilities to:**
  - Execute the training, validation, demonstration, and scaling of AI models.
  - Support model ensembling to combine multiple AI models depending on specific needs.
  - Support asynchronous inference (e.g., user-, API-, or trigger-initiated) for complex and intermittent models.
  - Enable continued push of AI insights back into the data management foundation for incorporation into training datasets.
- **Operational user workflows to:**
  - Provide mission users a reduced number of interfaces for ease of use.
  - Enable users to leverage AI outputs alongside all-source data in support of diverse workflows.

- Generate continuous feedback on the AI models as users perform their work so that feedback can serve as an additional data source to retrain and improve models.

Additionally, designing responsible, ethically defensible AI systems involves enshrining the assessment framework components we outline in our Category 1 response into all phases of AI systems development, especially into the design and construction of UI/UX that serve as the warfighters' interaction layer with AI systems.

The ability of warfighters to interact effectively and responsibly with AI capabilities — and consequently their capacity to gain trust in these systems as tools that support their work — is critically contingent upon their understanding of the ethical dimensions of technology applied to the military context, but also on the experiential modes through which they directly and indirectly engage with AI systems. In our experience, there are at least several important UI/UX product implications that should be considered in this regard:

- **Data curation:** Labelling AI-derived data and identifying its provenance to support user understanding and trust in the fidelity of the data and the details of the operational world it represents;
- **UI/UX Orientation:** Providing guided tours through the product's AI features along with an introductory training guide is important for situating first-time users of the technology;
- **AI Fundamentals:** Requiring users through general and in-program training to acknowledge the limits of AI-derived data analysis before they are granted access to the specific tools or views enabled by the AI system forces an explicit reckoning with how the AI system should and should not be responsibly used; and
- **Workflow support:** Supporting validation workflows where human action is required in order for an AI processing of data to produce 'promoted' analytical or other operational work products or other outputs that may become long-term artifacts for future use in, for example, targeting, general operations, or intelligence.

These examples are just a handful of ways that user interactions with AI systems can be constituted at the design level to reinforce RAI principles and help underwrite warfighter trust in those systems.

Category 4: Research and development (e.g., explainability, interpretability, privacy, etc.);

We advise an approach to AI explainability and interpretability that spans data-oriented, systems-oriented, development-oriented, and model-oriented measures to holistically

explain model behavior and increase trust by users, stakeholders, and other oversight authorities:

- **Data-level measures** involve understanding the training data atop which models are built, and developing confidence in its origin, relevance, and quality. All training data should retain full provenance to sources, and all intermediate transformations and their logic should be version controlled, enabling visibility into how training examples were derived. Logical steps should also be implemented to include selection and prioritization of data for labeling, feature derivations, image processing (e.g., chipping, orthorectification, format conversions), data collection prioritization, and metadata enrichment. Automated quality control measures should enable auditing of label quality and labeling guidance, and coordination of re-labeling if quality is low or labeling conventions need to be adjusted. Mission-feedback and production monitoring enable alerting should be in place to flag when the production scenario has diverged meaningfully from training data or model success regimes and characterizing the difference.

- **Systems-level measures** involve understanding the impact of different inputs (whether at training time or at inference time) on both specific model outputs as well as model metrics. T&E metrics should be employed to split out by various categories and scenario types to understand where a model excels, and conversely, patterns amongst failures. Both T&E and production model monitoring can also flag anomalous predictions. Black Teams should be employed to examine counterfactuals and branching scenarios that might allow, for example, assessments of degrading or corrupting data, selecting or simulating “stress-test” scenarios, or simulating an intentional adversarial attack, and measure which manipulations affected model outputs. By pairing manipulated or alternative inputs with corresponding outputs, users are in a position to better understand the causes of model breakdown. For any such analysis (T&E-style or Black Team style), full pedigree and lineage of model processes should be provided to facilitate trust in the analyses themselves.

- **Development-oriented** explainability involves understanding what changes are made to modeling logic – either by capturing and versioning modeling code and training parameters, or by capturing and structuring release notes, submission metadata, and execution configuration, to tie changes in model behavior to specific versions. If an instance of unexpected model output is flagged by an end-user or knowledge manager, that instance should be compared across prior model versions to understand if a specific update introduced that behavior, and isolate the change resulting in that behavior.

- **Model-level measures** for specific model types should involve codifying or deploying algorithmic approaches for model introspection – for example, weight

inspection, tree feature importance, SHAP values, LIME, DICE (counterfactual explanations), activation maps (for neural networks), attention maps (for applicable neural networks), and more. Model introspections should be preserved so that they can be used as part of any future activities such as health-checks that flag if a given feature importance diverges from baseline, or, triaging the test set down to examples for which two candidate model versions have significantly different attention maps.

Category 7: Risk management

As noted above, an operationally viable approach to building and deploying RAI should focus foremost on grounding AI systems as tools in their appropriate settings, i.e., as embedded within the full environment in which AI is to be deployed in production, as distinct from the considerations that may play out in laboratory settings. What this means in practice is that RAI should encompass several important features in order to manage associated deployment risks:

- It should treat explainability and traceability as applicable to the full lifecycle of data and model management (not just at the level of algorithmic assessment) and provide tools for continuous testing and evaluation. This means:
    - tracking the full provenance and lineage of all data and model branchings;
    - version controlling changes to data, models, parameters, conditions, etc.;
    - tracking how dynamic environmental factors modify usage and outcomes;
    - performing continuous testing and evaluation, data quality and integrity checks; and
    - creating a persistent and reliable audit trail for all data processing steps for later analysis, troubleshooting, oversight, and accountability
- It should hold model and systems maintenance as a critical and enduring condition for keeping the AI running, not building a model and assuming it will run unaided or somehow learn to keep itself running in good order as the world changes around it;
- It should view end-user interactions as a central feature to the workings of the full AI system, not as an afterthought (see Category 3 response);
- It should bias towards adopting a human-centric approach that keeps humans in the loop in recognition of their role in providing critical contextual grounding, but also in serving as the moral fabric for ethical consequences of AI tool use; and
- It should continuously consider an honest accounting of the trade-offs, limits, and failings of the system as an essential deployment responsibility, not an afterthought.