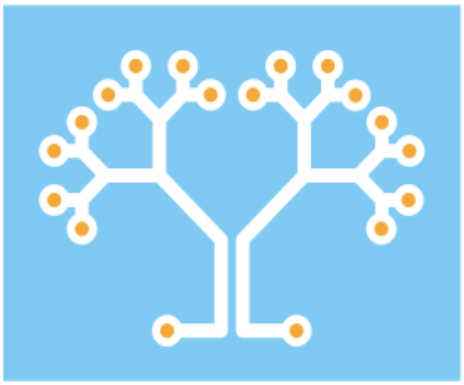

The Enterprise Neurosystem

RFI Response:

Update of the National Artificial Intelligence Research and Development Strategic Plan

Enterprise
Neurosystem

Introduction

The Enterprise Neurosystem is an open source community of Fortune 500 companies, enterprise technology vendors, and academia. It is ultimately designed to address the fundamental challenges of large-scale AI infrastructure to protect our global ecosphere. The founding academic institutions include Stanford SLAC, Harvard Analytics and UC Berkeley Data-X. Participating firms include America Movil, Equinix, Fiducia | AI, IBM Research, Intel, Kove, PerceptiLabs, Verizon Media/Yahoo!, Red Hat, Reliance Jio and others.

The Challenge

Climate change, energy security, agricultural challenges, political unrest, and mass migration pose immediate threats to global stability and the planet's health as a whole. The exponential growth of technology provides us with the means to address these challenges, but only if we, as a nation, can build an enabling infrastructure to leverage that technology effectively. The National Artificial Intelligence Research and Development Strategic Plan guides our opportunities to unlock innovations that the culmination of emerging heterogeneous systems will merge to reveal cross-domain patterns in data, algorithms, technology, and discoveries that lead to robust solutions for the longevity of our public, private, and commercial way of life.

The Enterprise Neurosystem community sees the tailoring of the OSTP Strategic Plan as a means to more optimally guide the construction of a wide-ranging AI infrastructure, one with broad public and private buy in, that will enable rapid application of US creativity to solve the challenges that lie ahead – identifying and circumventing environmental threats, providing advanced insight into human migration, integrating climate-adaptive agriculture from local to global scales, dynamically accommodating supply chain and energy grid disruptions, and ultimately advancing secure fusion and nuclear energy in a heterogeneous supply producer ecosystem. The mission here is to promote sustainably accelerated advancement that is consistent with our hypothesis that diversity and heterogeneity in technology as in thought will yield resilient solutions in a changing world, a hypothesis support by US history and by Nature's own demonstration of using diversity as a tool for bending but never breaking under stress.

Planetary scale objectives begin with smaller developmental steps at the intersection of the scientific community and the Fortune 500 enterprise market. For lasting and meaningful

change, government sponsored scientific advancement must quickly integrate into the commercial market and there grow legs of its own to live and breathe beyond the initial research seed funding. By nurturing a mid-tier environment for the national AI technology resource, it extends far beyond academic research and yields a continuous stream of commercially viable architectures that can self-coordinate as an autonomous and self-aware AI ecosystem. This cannot be overstated, the rapid spin-off productization enhances the next technological development, and this compounded multiplicative effect is exactly the root of the Kurzweilian exponential technological growth law.

A mere handful of major companies cannot harvest the rich intellectual capital in the US research and technology sectors. This Intellectual capital is best served by a quintessentially US culture that encourages and rewards ambitious entrepreneurialism and creative expression. In this way, our national culture provides ideal raw creativity to incubate the solutions needed for our–and the world's–challenges. Honing the forward-minded National AI Research Initiative will provide the tools and the connectivity to exponentially accelerate US innovation in AI to solve the globe's largest problems.

In the field of AI, innovation always begins from a very data and compute intensive position. New ideas require enormous high-performance computing (HPC) and data resources unavailable to garage-level startups. The strategy of the National AI Initiative ideally creates a national infrastructure that more than enables, it encourages exploratory development and provides the tools and platforms that foster discoveries and faster time to market. Having the US government serve this role facilitates a fair entry to the field for newcomers, accelerating novel hardware and algorithms, all while centralizing our ability to forecast the importance of emerging solutions and use that value forecasting to unlock computing resources that otherwise are not available to rising researchers and entrepreneurs.

The Solution

There is a need not only for the curation of datasets and published "data challenges" but also the access to the appropriate computing infrastructure that only the US Federal Government can support. There is an emerging ecosystem of heterogeneous AI accelerating hardware; the hobbyist scale of compute is no longer a viable option to make full use of the current state of the art in e.g. transformer models like BERT and GPT. The core of such models are trained using a vast set of exemplar training samples without the need of computationally expensive round truth labels. This modality is well positioned for so-called federated machine learning whereby participants can effectively share their data without exposing that data explicitly to others. The core transformer benefits tremendously from generalization across data sources while the final training for niche tasks relies on a vastly smaller sub-set of well labeled data with only the peripheral neurons as "last mile" trainable parameters. This tailoring phase is therefore

compatible with the smaller scale computing resources available to the individual niche stakeholders. Since the High Performance Computing resources like clusters of GPUs, TPUs, or IPUs and such provide the rapid training of the core transformers, allowing access to AI testbeds that are managed by the federal government is in perfect alignment with using public funding for private and public explorative computing that leverages the latest in emerging AI acceleration hardware.

With truly world leading AI infrastructure resources as federal government managed HPC infrastructure, a quantifiable metric system with a standardized use-intention classification system can be developed that allows broad public exploration during low-load periods. As is common in Department of Energy cluster management, exploratory jobs would be preempted when high peak load needs arise. The social/scientific impact metrics and data-producer intention identifiers can be used to encourage participants to follow best practices to promote high value and ethical outcomes in the form of carrots rather than sticks; high value and high ethics models and data will be the last to be preempted from compute resources. Being managed by the federal government and luring participation by the broader AI development community, the US could model a system of autonomous bias monitoring that could effectively warn users when their models might grow corrupted due to either accidental or even nefarious dataset poisoning. Such an agent would be a digital analog of biological T-cells, autonomously scanning the central models and distributed datasets for indications of mis-use or unintentional outcomes. All together, the federal government solves multiple challenges simultaneously; it exposes the next generation of AI developers to the latest high performance technology, it attracts commercial financial and hardware support for community shared resources, and it holds the world's most advanced AI infrastructure at the ready so that when catastrophe does strike, all resources can be immediately turned to the singular emergency response.

In the Enterprise Neurosystem community, there is a study group dedicated to potential ethical and safety issues in using AI technologies that influence or govern humanity. Our community also recognizes the underlying often accidental societal and cultural bias in curated data sets, affecting even the most granular AI determinations. Advanced feature engineering techniques can be broadly encouraged to help mitigate latent human bias, manage preference and develop ancillary AI models that assist the core intelligence and overall system toward driving greater accuracy and equity. Such an ethics system would reside at the kernel level within the primary guidance model to encourage beneficial outcomes for humanity.

Our community has noted the preponderance of AI models being deployed in the enterprise. They currently exist as bespoke solutions that lack deep integration with other models or domains, thus precluding the innovation that only arises from the cross-fertilization of collective findings. A connective data fabric and a centralized cross-correlation model are required for a national AI infrastructure that aligns with the long-term objective to monitor the planet and

take real-time actions as needed. This connectivity will involve the emerging swarm of Edge IoT devices, data sources, and AI models, with a core interpretive model and recommendation engine.

A distributed infrastructure paradigm shift needs to take place. Standardized and composable feature extraction methods should be applied *in situ* to improve efficiencies in distributed model training with enforced confidentiality. Data should only be moved on demand or by way of interpretive metadata or interface layer, one that both increases security and reduces expense and extraneous network traffic. Metadata can be delivered in a tiered fashion to reduce latency and shorten the time to action. Data generalities will only give way to finer granularity based on user requirements, permissions, and authentication. For instance, data sampling techniques that include anonymization through hashing can lead to smaller data sources maintained in their respective silos in a federated model. Data and related features are shared only on an as-needed and as-granted basis.

The national data fabric infrastructure with a curated and multi-tiered security system to authenticate users and enable targeted data sharing would be a requisite primary focus of this new architecture. Independent AI-powered security instances will travel the network to scan and authenticate new users and data sources. In a non-intrusive manner, round-the-clock penetration testing will be enabled, and remote decryption key monitoring and related pattern analysis will be implemented. Instead of granting access to the entire network, focusing on Layer 7 application connectivity based on mTLS and Zero Trust Network Architecture (ZTNA) will isolate and reduce the impact of intrusions to a single application.

Although Public Cloud Offerings have demonstrated industries' acceptance and hunger for cloud-based HPC services, the newly emerging Edge computational and AI paradigms create novel security challenges as these resources are decentralized. Emerging advanced hardware is not currently available in Public Cloud Infrastructure. At the same time, users can already see tremendous benefit from a Federated model that exposes innovations in a secure sandbox. In such a sandbox, both scientific researchers and industry innovators could explore tailored hardware, even work together to co-design new technology for their specific research or market needs. A federated model would also allow data to remain resident in individual silos, fostering a safe but rich collaborative outcome via its tiered metadata capabilities.

Production Architecture

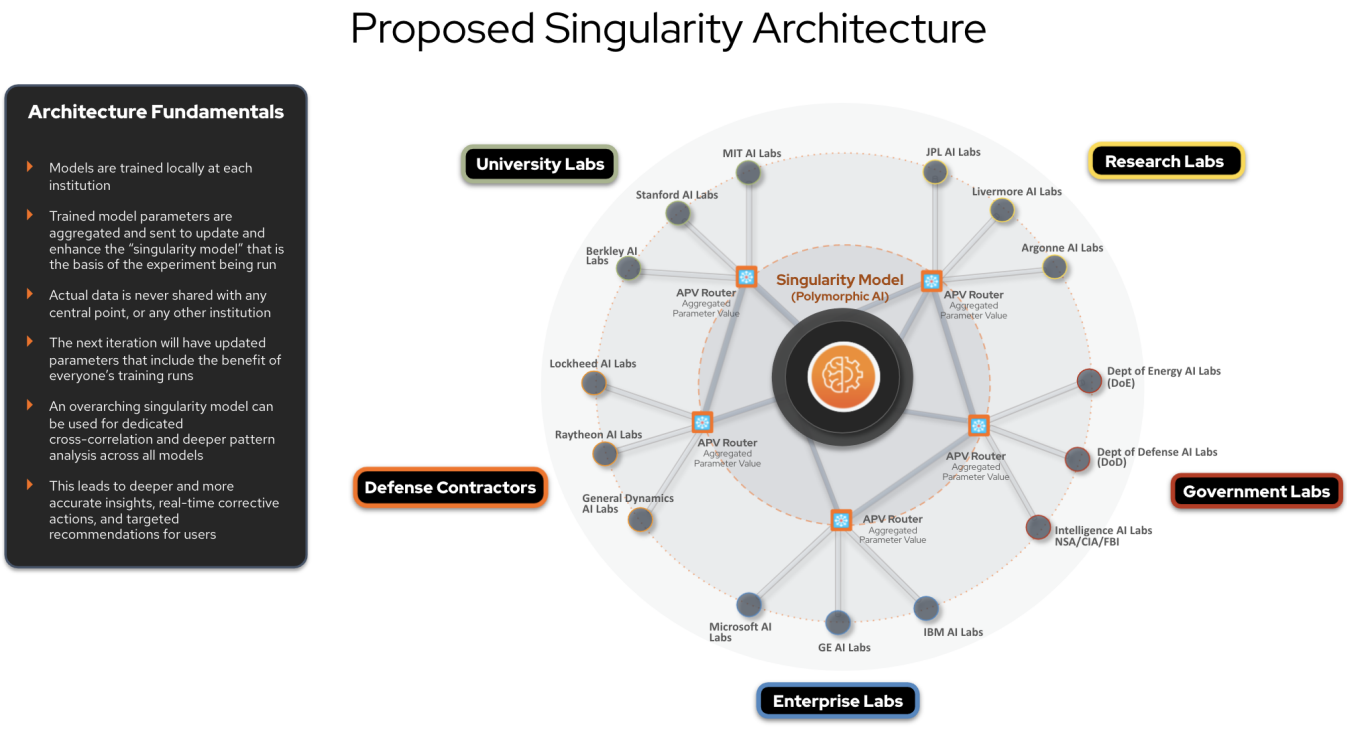

Illustration 4 - Tiered Cross-correlation Framework

Singularity refers to a pervasive synergy between human and artificial intelligence: an analytical engine that conducts ongoing pattern analysis, security reviews, and health checks across all the activities under its purview. As an opt-in model, not every function will require or be granted access to this full suite of resources, but the participating entities, be they private, public, or federal, would leverage a real-time analytic engine to advance science and directly inject technological discoveries into industry development. The complete body of federal research then becomes an immediate national AI resource. Enterprise and AI startups can as well propel intellectual property forward with new feature discovery, digital twin testing, and rapid time to market in an environment with maximal national exposure.

A series of intelligence engines, working together to unify streaming real-time and historical data to develop a more profound and instructive understanding of our environment and all its

possibilities with the forecast possibility to enable dynamic response to challenges as they happen.

An Open Source Community Approach to National AI Infrastructure and Research

In terms of all RFI elements, it seems clear that an open community approach to the ownership and maintenance of this resource would be optimal. Essentially, a federated compute, storage, and AI development environment composed of multiple labs and data centers, connected on a highly secure and lossless framework, with a small group of dedicated paid resources to manage the common infrastructure instance. This funded management team would coordinate architectural upgrades and responses to technical issues.

The related hardware and software resources would be donated and shared by the various organizations and participants in a federated network architecture. A governance model for this resource could be based on open community principles via the Linux Foundation or similar community frameworks. Furthermore, given element D, it requires a multi-tiered approach to data access and provenance tracking through Distributed ledger-based technology. All domains will likely benefit significantly from HPC-enabled digital twin simulation and creation. This motivates an infrastructure that enables a unified framework among a diversity of components, from the IoT Edge to the HPC Core. Such heterogeneity will be critical to the success of this endeavor, and several platforms and emerging techniques can be combined to help address this requirement.

Commonly known required components:

- Open and closed data sets that help participants enable model training capability to be identified, built, and curated. Metadata frameworks would be created to increase efficiency and help navigate issues of privacy and bias.
- A generalized digital twin environment that supports discovery and data set generation.
- A user-permission and authentication mechanism.
- A platform to create, distribute and manage AI models, with capabilities including ground-up development, pre-built models, pipeline workflow, lifecycle management, and drift correction.
- A hardware environment including all necessary resources, including storage, processors (GPU/TPU/IPU/FPGA, x86, ARM), networking components, and software infrastructure platforms (Kubernetes, container management, databases, memory optimization, etc.).

Unique elements would include:

- A flexible and dynamic resource federation model that is based on an automated data fabric that extends across all elements. This helps private sector firms gain access to and directly build better hardware and software via a shared space that is co-designed with public sector emerging needs.
- A marketplace that enables both open and private science and industry, with models available to any organization. In essence, a library of templated architectures and base pre-trained AI models for related research purposes that ultimately enable cross-correlation of models, incorporating relevant heterogeneous data sources and sensors that can generate pattern analysis in real-time to enable deeper insights and rapid course correction.
- A library of composable transformation and featurization layers that aid user adoption of the templated architectures and enable security and provenance tracking from the beginning of the development cycle.
- A software-defined memory allocation and virtualization framework (Kove, etc.) that eliminates bottlenecks for large-scale AI workloads and optimally capitalizes on both on-premise and distributed data stores.
- A tiered security architecture that includes:
    - Ongoing 24/7 penetration testing with non-intrusive security scans.
    - Distributed ledger-based user/resource authentication.
    - Layer 7 integration strategy (intrusions relegated to a single application).
    - Biometric authentication and resource isolation within the federated network for sensitive workstreams.
    - Monitor decryption keys for data provenance, audit, and automated discovery of abuse patterns.
    - Independent software instances that autonomously assign themselves to assess and scan newly federated hardware.
- Open source fairness and bias management tools, provided by a dedicated community ethics group. This team will apply operational parameters to data sources, metadata layers, and cross-correlation engines to maintain systemic neutrality. This includes a tight focus on privacy and the protection of individual rights.

Capabilities and services for prioritization

- Curated and openly available data sets would be the top priority in our estimation.
- Various users and institutions can access openly available models in a registry.
- Readily available hardware infrastructure for model training, openly available to all member institutions. Beyond training requirements, the use of more expensive chipsets (GPUs, TPUs, IPUs, FPGAs, etc.) should be evaluated from a price/performance perspective against standard chip architectures (x86, etc.) in terms of production lab use. This will lead to a balanced workload approach, cost savings, and co-designed hardware

and algorithm compositions as well as encourage community exploration of emerging new technology.

Current building blocks and resources

The various national lab environments (Argonne, Oak Ridge, Stanford SLAC, etc.) can be unified as a federated infrastructure, using containerized technologies and integration via container platform and data fabric/layer 7 application networks. With a metadata engine and digital twin layer, the proposed federated approach would enable secured data to remain on-premises but would allow each of the partner labs to leverage the computing resources across the DOE complex.

Public-private partnerships

Community-based research organizations like the Enterprise Neurosystem serve as examples of grass-roots efforts that act as a crossroads between academia, government, and the private sector, with a common goal of creating global scale infrastructure for humanity and the environment. The high level of private interest in such an organization shows the appetite for participation in a fair and federally regulated AI infrastructure. By encouraging partnerships between industries, academia, and government, the resulting ecosystem provides an open arena for encouraging equitable access, early experience with developing technologies, and diverse insight into the strengths and weaknesses of participant approaches that ultimately yields a more robust AI infrastructure and industry.

With common overarching objectives, a multi-tiered project approach has been robust and proven in practice. Members and the private sector both support and contribute to the Enterprise Neurosystem proof of concept. The membership of the community provides the related infrastructure as an open-source and shared approach to hardware infrastructure, with set permissions for users mapped to projects. We have security development tracks, application layer connectivity, and a distributed ledger authentication end state. And the ethical guideline development track is a fascinating exercise in canvassing operational philosophies and multicultural guidelines to create an intelligence that will non-intrusively contribute to society in a positive and unbiased manner. As we have a lab and private sector participation with shared hardware and open and available data sets, this navigation has been pain-free to date.

There is strong support for academic, government, and industry projects within this community, and these activities allow the private sector to find benefits through federal research and vice versa. For example, creating a similar neurosystem for the Fortune 500 would support a Business Singularity for each corporation. A real-time intelligence that helps each corporation

would be an evolutionary step in AI development that spans both AIOps and Business Intelligence functionality; creating mutual benefit for all parties is a primary objective.

D**emocratic access to AI R&D**

Data sensitivity to sharing across organizations is a significant challenge for a national AI infrastructure. Federated Machine Learning and confidential computing best practices will certainly help address the challenge by maintaining the integrity of individual data silos and adding the abstraction layers of Federation and metadata generation to provide accurate results without compromising privacy or security.

Contributing Authors and Review Committee:

Ryan Coffee
Senior Staff Scientist
SLAC National Accelerator Laboratory

Ben Cushing
Federal Health and Science Lead
Red Hat Inc.

Erik Erlandson
Senior Principal Software Engineer
Red Hat Inc.

Ganesh Harinath
CEO
Fiducia | AI

Vishnu Hari Kumar
Product Manager
Meta AI

Pierre Mathys
Global Senior Director, Telco Edge Solutions
Red Hat, Inc.

John Overton
CEO
Kove

Audrey Reznik
Senior Software Engineer
Red Hat Inc.

Dinesh Verma
CTO, Edge Computing and IBM Fellow
IBM Research

Bill Wright
Head of AI/ML and Intelligent Edge
Industries and Global Accounts
Red Hat, Inc.