Opportunity Title: Understanding the Interaction of Multiple Optimizers in Artificial Intelligence Models
Opportunity Reference Code: ICPD-2023-14

Organization Office of the Director of National Intelligence (ODNI)

Reference Code ICPD-2023-14

How to Apply **Create and release your Profile on Zintellect** – Postdoctoral applicants must create an account and complete a profile in the on-line application system. **Please note: your resume/CV may not exceed 2 pages.**

Complete your application – Enter the rest of the information required for the IC Postdoc Program Research Opportunity. The application itself contains detailed instructions for each one of these components: availability, citizenship, transcripts, dissertation abstract, publication and presentation plan, and information about your Research Advisor co-applicant.

Additional information about the IC Postdoctoral Research Fellowship Program is available on the program website located at: https://orise.orau.gov/icpostdoc/index.html.

If you have questions, send an email to ICPostdoc@orau.org. Please include the reference code for this opportunity in your email.

Application Deadline 2/28/2023 6:00:00 PM Eastern Time Zone

Description **Research Topic Description, including Problem Statement:**

Artificial Intelligence (AI) represents a new paradigm for the Intelligence Community. While AI promises to automate many collection and analytic processes, it does not come without risk, particularly as AI systems develop layers of optimizers to improve results and/or automate solutions. As this occurs there will be a misalignment from the developer's perspective and the AI model. Specifically, AI systems are mesa optimizers, meaning that their internal optimization functions (e.g., reward functions) will not necessarily converge with the learned model (generally speaking, the programmer's goal for the AI system).*1* This leads to misalignment between the objectives of the AI system and those of their human programmers. Furthermore, as more capable AI systems are given meaningful control over critical systems, the ability to control or influence AI decision making diminishes (termed the "Control Problem"). As the IC continues to operationalize AI, this understanding will improve models in everything from enterprise image detection and natural language processing to better understanding and interdiction of control risks before they manifest in critical national security systems

1see Hubinger et al.'s "Risks from Learned Optimization in Advanced Machine Learning Systems."

Example Approaches:

Develop mathematical models of decision-making in deep learning neural networks; Explore statistical methods of minimizing bias in training data sets; Examine potential game-theoretic solutions to the Control Problem; Develop mathematical models of reinforcement learning systems; Examine

Opportunity Title: Understanding the Interaction of Multiple Optimizers in Artificial Intelligence Models
Opportunity Reference Code: ICPD-2023-14

mathematical approaches to understanding mesa optimization.

Relevance to the Intelligence Community (IC):

As the IC continues to operationalize AI, this understanding will improve models in everything from enterprise image detection and natural language processing to better understanding and interdiction of control risks before they manifest in critical national security systems.

Qualifications

Postdoc Eligibility

- U.S. citizens only
- Ph.D. in a relevant field must be completed before beginning the appointment and within five years of the application deadline
- Proposal must be associated with an accredited U.S. university, college, or U.S. government laboratory
- Eligible candidates may only receive one award from the IC Postdoctoral Research Fellowship Program

Research Advisor Eligibility

- Must be an employee of an accredited U.S. university, college or U.S. government laboratory
- Are not required to be U.S. citizens

Key Words: #Artificial Intelligence, #Control Problem, #Machine Learning, #Deep Learning, #Explainability, #Alignment, #Critical Systems, #Mesa Optimization, #Complexity, #Game Theory

Eligibility Requirements

- **Citizenship:** U.S. Citizen Only
- **Degree:** Doctoral Degree.
- **Discipline(s):**
  - **Chemistry and Materials Sciences** (12)
  - **Communications and Graphics Design** (6)
  - **Computer, Information, and Data Sciences** (17)
  - **Earth and Geosciences** (21)
  - **Engineering** (27)
  - **Environmental and Marine Sciences** (14)
  - **Life Health and Medical Sciences** (48)
  - **Mathematics and Statistics** (11)
  - **Other Non-Science & Engineering** (2)
  - **Physics** (16)
  - **Science & Engineering-related** (1)
  - **Social and Behavioral Sciences** (29)