P | I | P | S

In Defense of Data

How the DoD Can Strengthen AI Development

PIPS White Paper 12.5: *Executive Summary*

Clara Waterman, Research Fellow
Selene Swanson, Research Intern

The Department of Defense (DoD) is investing in artificial intelligence (AI) to prepare for future warfare against peer adversaries. However, Washington is devoting insufficient attention and funding to the datasets that underpin AI algorithms. Poor training data can create flawed algorithms that misidentify operating environments, potentially degrading battlefield decision-making and hindering the Pentagon's ability to maintain a strategic advantage over adversaries. Consequently, high-quality datasets are crucial to the DoD mission.

At this early stage of development, the DoD has an opportunity to standardize and improve the quality of its AI training data by creating a data clearinghouse. This clearinghouse would coordinate data collection, establish best practices for both vetting AI data for bias and minimizing human error, and standardize metadata formatting. Increased collaboration and attention to developing datasets today will maximize the efficiency and effectiveness in which the DoD develops AI for the military.

The DoD, Data, and AI

Artificial intelligence without data is like a car without gas: the framework for movement exists, but the vehicle is incapable of movement without the proper inputs. Before data can influence AI, it first must be collected, labeled, and vetted. This data then is used to train AI algorithms. These AI algorithms produce results, which often function as the input data for other algorithms, thus completing the data cycle. Consequently, flawed data produces flawed AI algorithms and results, which can be misleading in subtle ways.

The DoD has already included AI elements in its mission-support and enterprise efforts, but its goal is to incorporate AI into operational tasks as well. The Pentagon plans to use AI to improve battlespace navigation, enhance U.S. threat-assessment capabilities and minimize risks to fielded forces, and to coordinate its decision-making across battle domains and military services. However, the goal of operational AI is undermined if the data that underpin AI development is not adequately collected, vetted, labeled, and shared. For example, a group at Boston University recently created a back door into an AI system by poisoning just 0.025 percent of the training data, which could have gone unnoticed under the DoD's current data practices

Lack of Data Coordination and Standardization

As the largest data producer and consumer in the United States government (USG), the Pentagon faces multiple challenges to its current data practices. Many of the datasets created or used by the DoD lack contextual information because there is no standardized way for dataset creators to record information about their data's origins and the vetting and labeling processes it went through. Lack of information about a dataset's background may lead to confusion about what the data is truly representing and to unintentional misuse of that dataset.

Another problem facing the DoD is that there is little clear enterprise-wide communication about what data is being collected. Interviews with DoD employees indicate that two or more groups within the USG will often curate collections of similar data on the same topic, making their efforts redundant and inefficient. Because the DoD collects so much data and preparing it can be expensive, some projects run out of funding and are shelved before completion. Furthermore, inconsistencies in data coding and format limit the potential for collaboration between offices, which is essential for the efficient comparing and combining of datasets to develop AI algorithms.

Unintended Consequences of Bad Data

If DoD uses AI algorithms trained on flawed data, then it is more likely to misunderstand its operating environment. Failure to understand the operating environment fully can lead to misinformed decision-making, which has historically led to the improper and unintended use of lethal force. Bad data also may leave USG personnel more vulnerable to attack if they are reliant upon an algorithm that fails to alert them to dangerous conditions. Further, if the Pentagon adopts flawed data and algorithms on a wide scale, this misestimation could ultimately lead to a loss of U.S. strategic advantage over adversaries, such as Russia and China, who are making substantial investments in AI-based technologies.

Policy Solution: Data Clearinghouse

The DoD should create a data clearinghouse to increase the efficiency and effectiveness of its AI development. While the clearinghouse would not be a central repository for all DoD data, it would provide employees with a working index of what datasets exist on a given topic, the quality of those datasets, and pertinent information from the originator. The clearinghouse would allow for groups both internal and external to the DoD to submit requests for datasets, reducing costs and redundancy in collection. While this clearinghouse would not be as costly or labor-intensive as other collaborative data projects (like the National Counter Terrorism Center), it could service almost as many members of the USG. Using this clearinghouse, the DoD could introduce a standardized pedigree form for datasets that indicates the data origin, age, method of collection, vetting, and labeling, and the level of confidence the originator has in the dataset.