The proposed **AI Vulnerabilities Database (AVID, working title/acronym)** is a open-source knowledge base of vulnerabilities for large-scale, pretrained AI/ML models. Inspired by the MITRE ATT&CK framework for cybersecurity attacks and the MITRE ATLAS database for adversarial ML techniques, AVID will encompass coordinates of trustworthy ML such as fairness, robustness, privacy, reliability, and alignment. AVID will propose a taxonomy of potential harms across these coordinates, and will contain full-fidelity information (model metadata, harm metrics, measurements, benchmarks, and mitigation techniques if any) on the evaluation of these harms for specific versions of specific models that are either open-source, or accessible through APIs.

With the availability of large-scale pretrained models—specifically in NLP—becoming ubiquitous, it is now common practice among ML practitioners to fine-tune such models for specific applications. AVID will enable practitioners to take a proactive approach in debugging their ML workflows built on top of such models. AVID will be expandable through open-source contributions from practitioners to account for novel and hitherto unknown vulnerabilities. AVID will mandate code and evaluation data submission for each and every vulnerability. As the trustworthy ML community works towards setting common, open standards for responsible ML systems, this will add transparency and reproducibility to model evaluation practices.

In the broader space of AI Ethics, the only previous line of relevant work is the Twitter algorithmic bias bounty challenge, which we plan to learn from while building AVID. While the AVID harm taxonomy is also relevant for generic ML models, to begin with we shall focus on pretrained models due to their availability and widespread use among ML practitioners.