The science of defenses for machine learning are somewhat less well developed. Here we consider several defensive goals. First, we consider methods at training and inference time that are robust to distribution drifts–the property that ensures that the model performs adequately when the training and runtime input distributions differ. Second, we explore models that provide formal privacy preserving guarantees–the property that the amount of data exposed by the model is bounded by a privacy budget (expressed in terms of differential privacy). Lastly, we explore defenses that provide fairness (preventing biased outputs) and accountability (explanations of why particular outputs were generated, also known as transparency).

In exploring these facets of machine learning attacks and defense, we make the following contributions:

- We introduce a unifying threat model to allow structured reasoning about the security and privacy of systems that incorporate machine learning. This model, presented in Section III, departs from previous efforts by considering the entire data pipeline, of which ML is a component, instead of ML algorithms in isolation.
- We taxonomize attacks and defenses identified by the various technical communities as informed elements of PAC learning theory. Section IV details the challenges of in adversarial settings and Section V considers trained and deployed systems. Section VI presents desirable properties to improve the security and privacy of ML.
- In Section VII, we introduce a no free lunch theorem for adversarial machine learning. It characterizes the tradeoff between accuracy and robustness to adversarial efforts, when learning from limited data.