

Carl Lochstampfor

CYSE 420 — Module 08 Assignment

October 15, 2026

Understanding Generative AI in Cybersecurity

Core Principles of Generative AI. Generative AI creates new content by learning the statistical structure of existing data and sampling from it, producing examples absent from the training set. Three architectures dominate. A Generative Adversarial Network pits a generator, which turns noise into candidate samples, against a discriminator that judges whether each is real, training as a minimax game until the generator's output is indistinguishable from real data. A Variational Autoencoder encodes input as a distribution over a latent space rather than a single point, and decodes samples from it. A Transformer uses self-attention to weigh every element of a sequence against every other in parallel. All three are pre-trained broadly, then fine-tuned for specific tasks later.

Generative and Discriminative Models. Discriminative models learn the boundary separating classes, optimizing for classification accuracy. A generative model instead learns how the data was produced, so it can create new examples, work where labels are scarce, and describe what normal looks like rather than where the boundary sits.

Applications in Cybersecurity. Security datasets are chronically imbalanced because genuine attacks are rarer than normal network traffic. A GAN trained on captured network flows can generate synthetic traffic, benign and malicious, so a detector sees enough examples of a rare technique to learn it. However, simulated DDoS or SQL injection traffic exposes it to patterns costly or infeasible to collect from production. Synthetic records also replace real logs whose release would expose proprietary data, a common obstacle to sharing incident data.

Threat detection works in the opposite direction: a model learns the distribution of ordinary activity and flags departures from it. Because it weighs context rather than fixed rules, a credential that usually authenticates from one subnet in business hours and then reaches bulk file shares at 3 a.m. registers as anomalous though no single action breaks a rule.

Machine learning models are also attack surfaces. Generative models can produce adversarial examples, inputs perturbed to force misclassification or poisoned records that corrupt training, letting defenders measure how far a model can be pushed before it fails. Testing a malware classifier against obfuscated samples reveals evasion paths before an adversary finds them. The same arrangement hardens the defender: a discriminator trained to separate generated attack traffic from genuine traffic becomes an anomaly detector.

Ethical Implications and Challenges. The capability that generates phishing emails for awareness training generates convincing ones for real campaigns, and the same tools produce deepfakes. Training them requires large corpora that may contain sensitive records, raising obligations under GDPR and CCPA. Bias is a quieter but important risk to consider: a model trained on skewed incident data generates skewed scenarios, so defenses over-prepare for familiar threats. The technology also drives escalation: as defenders adopt generative tools, so do adversaries, each round raising the cost of the next.

Two technical problems compound these implications and challenges. Deep generative models are effectively black boxes, making it hard to justify why a system flagged an event. Evaluation is equally difficult, because conventional accuracy misleads: a generative model can score perfectly against the data it trained on while producing output that is incoherent on inspection. Transparent design, held-out validation, and collaboration between practitioners and developers are the countermeasures.