Google

Google'ın Güvenli Yapay Zeka Çerçevesi

Yapay zeka hızla gelişiyor ve onunla birlikte etkili risk yönetim stratejilerinin de değişmesi önemli. Bu değişimin gerçekleşmesine yardımcı olmak amacıyla Güvenli Yapay Zeka sistemleri için kavramsal bir çerçeve olan Güvenli Yapay Zeka Çerçevesi'ni (SAIF) hayata geçiriyoruz. SAIF altı temel ilkeden oluşuyor:

1. Güçlü güvenlik altyapılarını yapay zeka ekosistemine yaymak

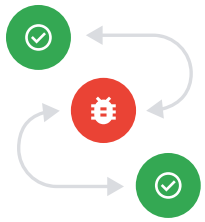
Yapay zeka sistemlerini, uygulamaları ve kullanıcıları korumak için geçtiğimiz yirmi yıl boyunca inşa edilmiş, varsayılan olarak güvenli altyapı korumalarından ve uzmanlıktan yararlanılması. Aynı zamanda, yapay zeka alanındaki gelişmelere ayak uydurmak için organizasyonel uzmanlık geliştirilmesinin yanı sıra yapay zeka ve değişen tehdit modelleri bağlamında altyapı korumalarının ölçeklendirilmeye ve uyumlu hale getirilmeye başlanması. Örneğin, SQL enjeksiyon gibi enjeksiyon teknikleri bir süredir biliniyor ve kuruluşlar istem enjeksiyonu tarzı saldırılara karşı daha iyi savunmaya yardımcı olması için girdi sanitasyonu ve kısıtlama gibi risk azaltma önlemlerini uyumlu hale getirebilirler.

2. Kuruluş için olabilecek tehditlerin daha etkili tespit edilmesi ve tehditlere reaksiyon verilmesi için yapay zekadan yararlanmak

Tehdit istihbaratı ve diğer kabiliyetleri genişleterek yapay zeka bağlantılı siber vakaların zamanında tespit edilmesi ve bu vakalara reaksiyon verilmesi. Kuruluşlar için bu, anormallikleri tespit etmek amacıyla üretken yapay zeka sistemlerinin girdi ve çıktılarının izlenmesini ve saldırıları önceden tahmin etmek için tehdit istihbaratının kullanılmasını da içerir. Bu çaba genellikle güvenlik, tehdit istihbaratı ve kötüye kullanımla mücadele ekipleriyle işbirliği yapılmasını gerektirir.

3. Mevcut ve yeni tehditlerin hızına yetişmek için savunmaları otomatikleştirmek

Güvenlik olaylarına verilen yanıtların ölçeğini ve hızını iyileştirmek için en son yapay zeka inovasyonlarından yararlanılması. Kötü amaçlı aktörlerin daha büyük etki yaratmak için yapay zekayı kullanması olasıdır. Bu durumlara karşı korunmada tetikte kalmak ve bunu uygun maliyet ile yapabilmek için yapay zekadan ve onun mevcut ve yeni ortaya çıkan kabiliyetlerinden faydalanmak önemlidir.



4. Kuruluş genelinde tutarlı güvenlik sağlamak için platform seviyesi kontrollerini uyumlu hale getirmek

En iyi korumaların tüm yapay zeka uygulamalarında ölçeklenebilir ve uygun maliyetli bir şekilde mevcut olmasını sağlamak için yapay zeka riskinin azaltılmasını ve korumaların ölçeklendirilmesini farklı platformlar ve araçlarda destekleyecek şekilde kontrol çerçevelerinin uyumlu hale getirilmesi. Google'da bu, varsayılan olarak güvenli korumaların Vertex AI ve Security AI Workbench gibi yapay zeka platformlarına genişletilmesini ve kontroller ile korumaların yazılım yaşam döngüsünün içine inşa edilmesini de içerir. Perspective API gibi genel kullanım durumlarını ele alan kabiliyetler kuruluşun tamamının en gelişmiş teknolojiyle donatılmış korumalardan faydalanmasını sağlayabilir.

5. Risk azaltma önlemlerini düzenlemek için kontrolleri uyumlu hale getirmek ve yapay zekanın dağıtımı için daha hızlı geri bildirim döngüleri oluşturmak

Uygulamaların devam eden öğrenme aracılığıyla sürekli test edilmesi ve değişen tehdit ortamına yanıt vermek için saptamanın ve korumaların geliştirilmesi. Bu; olaylara ve kullanıcı geri bildirimlerine dayalı pekiştirmeli öğrenme gibi teknikleri içerir ve eğitim veri kümelerinin güncellenmesi, saldırılara stratejik olarak yanıt vermek için modellere yönelik ince ayar yapılması ve model oluşturmak için kullanılan yazılımın bağlama daha fazla güvenlik yerleştirmesine izin verilmesi gibi adımları içerir (örneğin, anormal davranışların tespit edilmesi). Kuruluşlar aynı zamanda yapay zeka destekli ürünler ve kabiliyetler için güvenlik garantisini geliştirmek amacıyla düzenli Kırmızı Takım egzersizleri de gerçekleştirebilir.

6. İş süreçlerinin etrafındaki yapay zeka sistemi risklerini birlikte ele almak

Kuruluşların yapay zekayı nasıl işlerinde uygulayacaklarına dair uçtan uca risk değerlendirmeleri gerçekleştirilmesi. Bu; veri kökeni, belirli uygulamalar için doğrulama ve operasyonel davranış izleme gibi uçtan uca işletme riskinin bir değerlendirmesini içerir. Bunlara ek olarak, kuruluşlar yapay zeka performansını doğrulamak için otomatikleştirilmiş kontroller oluşturmalıdır.