Google

Sorumlu Makine Öğrenimi için Yönergeler

Yönergeler

Kontrol akışları Makine Öğrenimi modellerinin çıktılarına göre değiştirildiğinde tedbirli hareket edin.

Bir makine öğrenimi primitifinin çıktısını kullanan iş akışları, güvenliğinin sağlanması zor olan saldırı yüzeyleri üretecek şekilde değiştirilebilir. Mevcut tavsiyede olduğu gibi, açıkça beklenmeyen bir uygulama iş akışına veri girilmesi tavsiye edilmez.

Modellere ve model meta verilerine doğrudan erişimden kaçının.

Makine öğrenimi primitifleri ile doğrudan etkileşim; sızma, ters çevirme ve diğerleri dahil farklı saldırı yüzeylerine yol açabilir. Bu objeleri kullanıcılardan uzak tutun ve sadece kontrollü ve onaylı etkileşimlere izin verin. Çıktının, bir veri tabanı gibi bilinen bir özelliğe iliştirildiğinden emin olun. Çıktıyı doğrudan kullanıcının kendisine sağlamayın.

Modellerin gerek eğitim safhasında gerekse uygulama sonrası safhada bütünlük kontrolünü kullandığından emin olun.

Yazılımda, geliştirme ve dağıtım ardışık düzeninde birden fazla aşamada değişiklikler meydana gelebilir; aynısı makine öğrenimi için de geçerlidir. Kontrol edin ve tekrar kontrol edin. Sadece beklendiğini bildiğiniz uygulamaları çalıştırın: Çalıştırmadan önce doğrulayın.

Özelliklerin karmaşık olması karmaşık kontrollere yol açabilir.

Etkileşimlerin kısa, yerinde ve özlü olmasını sağlayın. Büyük ve karmaşık iş akışları, başarılı saldırılar açısından çok daha büyük maruz kalma risklerini beraberinde getirir.

Eğitim için, genişletilmiş mantıkla başlayın ve gerektiğinde özellikler ve çıkarım için içlemsel mantığa geçin.

Eğitim için, doğrulanmış, bilinen iyi bir veri kümesi kullanmak veri zehirlenmesi ve diğer saldırıları önlemenin iyi bir yoludur. Ancak çok daha büyük çıkarım kabiliyetlerine sahip özellikler oluşturmak için yeni verilerin kullanılması gerekir; bu, eğitimin sonuçlarının bilinen iyi girdilerden beklenen sonuçlara göre dikkatlice test edilmesi daha dayanıklı sistemlerin ortaya çıkmasını sağlar.

Primitifin amaçlanan kullanım durumu ile birlikte bir entegrasyon testinin oluşturulduğundan emin olun.

Makine öğrenimi iş akışları da başka herhangi bir yazılım gibi entegrasyon ve birim testi yapılmasını gerektirir. Makine Öğrenimi yazılımlarının buluşsal niteliği güvenilir onaylamalar geliştirilmesini zorlaştırabildiği için iş akışının arzu edilen, uzun vadeli sonuçlarına odaklanılması önerilir.

Özellikle “ne olduğunuza” odaklanan kontrollerle ilgili olarak makine sentezi kullanan saldırılara dikkat edin.

Özellikle erişim hakları sağlamak için kullanıldığında ses, video ve metnin gerçekliği konusunda varsayımlarda bulunmadan önce daima “*Bir makine bunu, şimdi yapabilir mi?*” diye sorun.

Ekonomik modelin bir adım ötesinde olmak ve güçlü, düşük maliyetli bir güvenlik modeli sağlamak için ortak politikalar kullanın.

Bir saldırının maliyeti, müteakip düzeltmenin maliyetinden daha düşükse, o zaman kontrol muhtemelen etkisiz olacaktır. Güvenlik maliyetlerini, başarılı saldırıların etkisine göre otomatik olarak ayarlayarak saldırgana yönelik maliyetleri kademeli olarak artıracak şekilde operasyonel hale getirin.

Google

Tescilli veriler üzerinde eğitilmiş modelleri, örneğin bir cihaza yerleştirmek gibi kullanıcının kontrolüne vermekten kaçının.

Makine öğrenimi primitifi oluşturulurken ilgili kullanım durumunu göz önünde bulundurun ve teorik olarak, bunu oluşturan eğitim verisinin geri kurtarılabilir olduğunu varsayın. Ürünleri dağıtırken bunu iş sürekliliği planlaması ve risk değerlendirmeleri için kullanın.

Ekstrapolasyon saldırılarından kaçınmak ve gereksiz kaynak tüketimini önlemek için hız kısıtlamasını kullanın.

Geleneksel kontroller hala büyük ölçüde etkilidir; makine öğrenimi teknolojilerini tasarlarken ve uygularken erişim kontrol listeleri, hız kısıtlama ve kimlik doğrulama gibi katman kontrolleri bunlara örnek olarak gösterilebilir.