Google

Güvenli Yapay Zeka Framework Yaklaşımı

Güvenli Yapay Zeka Framework'ü
oluşturmak için Hızlı Bir Rehber
(SAIF – Secure AI Framework)

Safer with Google

İçindekiler

Giriş

Güvenli Yapay Zeka Frameworku (*Secure AI Framework (SAIF)*), güvenli yapay zeka (AI) sistemleri için kavramsal bir frameworktur. Google'ın yazılım geliştirmeye uyguladığı tedarik zincirini gözden geçirme, test etme ve kontrol etme gibi en iyi güvenlik uygulamalarından esinlenirken, yapay zeka sistemlerine özgü güvenlik mega trendleri ve riskleri hakkındaki anlayışımızı da dahil etmektedir. SAIF, güvenlik ve risk profesyonelleri için aşağıda belirtilenler gibi en akılda kalan endişeleri ele almak için pratik bir yaklaşım sunar:

- *Güvenlik*
  a) Erişim yönetimi
  b) Ağ / uç nokta güvenliği
  c) Uygulama/ürün güvenliği
  d) Tedarik zinciri saldırıları
  e) Veri güvenliği
  f) Yapay zekaya özgü tehditler
  g) Tehdit algılama ve müdahale

- *Yapay Zeka (AI)/ Makine Öğrenmesi (ML) modelleri risk yönetimi*
  a) Model şeffaflığı ve sorumluluğu
  b) Anomalileri tespit etmek için hataya eğimli manuel incelemeler
  c) Veri zehirlenmesi
  d) Veri kökeni, saklama ve yönetişim kontrolleri

- *Gizlilik ve uyumluluk*
  a) Veri gizliliği ve hassas verilerin kullanımı
  b) Gelişmekte olan düzenlemeler

- *İnsanlar ve organizasyon*
  a) Yetenek açığı
  b) Yönetmelik / Yönetim Kurulu raporlaması

Bu hızlı kılavuz, kuruluşların SAIF yaklaşımını mevcut veya yeni yapay zeka uygulamalarına nasıl dahil edebilecekleri konusunda üst düzey pratik hususlar sağlamayı amaçlamaktadır. İlerideki içeriklerde konular daha derinlemesine incelenecektir - şimdilik *SAIF*'in altı temel unsurunun her biri altında ele alınması gereken öncelikli unsurlara odaklanıyoruz.

- Güçlü güvenlik temellerini yapay zeka ekosistemine taşıma
- Yapay zekayı bir kurumun tehdit evrenine dahil etmek için tespit ve müdahaleyi genişletme
- Mevcut ve yeni tehditlere ayak uydurmak için savunmaları otomatikleştirme
- Kurum genelinde tutarlı güvenlik sağlamak için platform düzeyinde kontrolleri uyumlu hale getirme
- Hafifletmeleri ayarlamak ve yapay zeka dağıtımı için daha hızlı geri bildirim döngüleri oluşturmak için kontrolleri uyarlama

- Yapay Zeka sistemlerinin iş süreçleri çevresindeki risklerini bağlamsallaştırın (ya da bir zemine oturtun).

Güvenli Yapay Zeka Ortamını Uygulamaya Koyma

Adım 1 – Kullanımı Anlama

Birçok kuruluş yapay zekayı ilk kez kullanmayı veya yeni Üretken Yapay Zeka (*Generative AI*) yeteneklerinden yararlanmak için sahip oldukları yapay zeka çözümlerini genişletmeyi düşünüyor. Her durumda, yapay zekanın çözeceği belirli iş sorununu ve modeli eğitmek için gereken verileri anlamak, SAIF'in bir parçası olarak uygulanması gereken politika, protokol ve kontrolleri yönlendirmeye yardımcı olacaktır.

Örneğin, bir analist raporunu özetleyen veya dolandırıcılığı tespit eden modeller gibi mevcut verileri analiz etmek veya bunlara dayalı olarak hareket etmek için tasarlanan modeller, mesela tüketiciler ve geçerli tüketici koruma yükümlülükleri üzerindeki potansiyel etki nedeniyle ek zorluklar ortaya çıkaracak olan tüketici finansmanı için tahminler yapmak için kullanılan modellere (örneğin kredi riski modelleri) kıyasla daha az karmaşık sorunu içerebilir.

Modellerin son kullanıcılarla nasıl etkileşime girdiği de önemli bir rol oynamaktadır. Örneğin, son kullanıcı girdisi alan harici bir yapay zeka modeli, hisse senedi alım satımı için kullanılan bir modele kıyasla güvenlik ve veri yönetimi için farklı gereksinimlere sahip olacaktır. Bununla birlikte, üçüncü bir tarafın önceden oluşturulmuş bir modelini kullanmak ile kendi modelinizi geliştirmek ve/veya eğitmek, altyapının ve geliştirme platformunun güvenliğini sağlamak, model davranışını ve sonucunu izlemek, tehdit tespiti ve korumasına ilişkin farklı sonuçlar doğuracaktır.

Bu nedenle, yapay zeka kullanım alanının tam olarak anlaşılması, SAIF uygulamasının karmaşıklığının ve belirli dağıtım risklerinin yakalanmasını sağlayacaktır.

Adım 2 – Ekibi Oluşturma

Yapay zeka sistemlerinin geliştirilmesi ve konuşlandırılması *(deploy)*, tıpkı geleneksel sistemler gibi, çok disiplinli çabalardır ve risk değerlendirmesi, güvenlik / gizlilik / uyumluluk kontrolleri, tehdit modellemesi ve olay müdahalesi gibi benzer unsurları içerir. Ayrıca, yapay zeka sistemleri genellikle karmaşık ve opaktır, çok sayıda hareketli parçaya sahiptir, büyük miktarda veriye dayanır, yoğun kaynak gerektirir, yargıya dayalı kararları uygulamak için kullanılabilir ve saldırgan, zararlı olabilecek veya stereotipleri ve sosyal önyargıları sürdürebilecek yeni içerikler üretebilir. Bu, ekibin yapısını çeşitli kuruluşlardaki paydaşları içerecek şekilde genişletir, örneğin:

- İş kullanım alanı (use case) sahipleri
- Güvenlik
- Cloud Mühendisliği
- Risk ve Denetim ekipleri
- Gizlilik
- Hukuki
- Veri Bilimi ekipleri
- Geliştirme ekipleri
- Sorumlu Yapay Zeka ve Etik

Doğru çapraz fonksiyonlu ekibin kurulması, güvenlik, gizlilik, risk ve uyumluluk hususlarının en başından itibaren dahil edilmesini ve sonradan eklenmemesini sağlar.

Adım 3 – Bir Yapay Zeka Rehberi ile Seviye Belirleme

Yapay zeka, özellikle de Üretken Yapay Zeka (Generative AI), hala yeni ortaya çıkan ve hızla gelişen bir teknolojidir. Ekipler iş kullanımını, geçerli olan çeşitli ve gelişen karmaşıklıkları, riskleri ve güvenlik kontrollerini değerlendirmeye başlarken, ilgili tarafların yapay zeka modeli geliştirme yaşam döngüsünün temellerini, yetenekler, yararlar ve sınırlamalar dahil olmak üzere model metodolojilerinin tasarımını ve mantığını anlamaları çok önemlidir. *Yapay zeka, makine öğrenimi (ML), derin öğrenme (DL), Gen AI, büyük dil modelleri (LLM'ler)* gibi kavramlarla başlamak, teknik olmayan paydaşların *(stakeholder)* yapay zekayı güvenli ve sorumlu bir şekilde yönetmek ve dağıtmak için gereken riskleri ve kontrolleri doğru bir şekilde yakalamasına ve değerlendirmesine olanak tanıyacaktır.

Adım 4 – SAIF'in Altı Temel Öğesini Uygulama

Kullanım alanları ve bağlam bilinerek, ekip bir araya getirilip yapay zeka konusunda hazırlandıktan sonra, daha önce bahsedilen bazı endişeleri gidermek için SAIF'in altı temel öğesini uygulamaya başlayabilirsiniz. Bu öğelerin kronolojik sırayla uygulanmasının amaçlanmadığı, daha ziyade kuruluşlara yapay zeka sistemlerini güvenli ve sağlam bir şekilde inşa etmeleri ve dağıtmaları için toplu olarak rehberlik eden kaldıraçlar olduğu unutulmamalıdır.

Güçlü güvenlik temellerini yapay zeka ekosistemine taşıyın

Güvenlik alanlarındaki mevcut güvenlik kontrollerinin yapay zeka sistemleri için ne anlama geldiğini gözden geçirin

Güvenlik alanlarındaki güvenlik kontrolleri, yapay zeka sistemleri için çeşitli şekillerde geçerlidir. Örneğin, veri güvenliği kontrolleri, yapay zeka sistemlerinin eğitmek ve çalıştırmak için kullandığı verileri korumak için kullanılabilir. Uygulama güvenlik kontrolleri, yapay zeka sistemlerinin uygulandığı yazılımı korumak için kullanılabilir; altyapı güvenlik kontrolleri, yapay zeka sistemlerinin dayandığı temel altyapıyı korumak için kullanılabilir; ve operasyonel güvenlik kontrolleri, yapay zeka sistemlerinin güvenli bir şekilde çalıştırılmasını sağlamak için kullanılabilir.

İhtiyaç duyulan spesifik kontroller, yapay zekanın kullanımının yanı sıra spesifik yapay zeka sistemleri ve ortamlarına bağlı olarak değişecektir.

Mevcut frameworkleri kullanarak geleneksel kontrollerin yapay zeka tehditleri ve riskleriyle uygunluğunu değerlendirin

Geleneksel güvenlik kontrolleri yapay zeka tehditleri ve riskleri ile ilgili olabilir, ancak etkili olmaları için uyarlanmaları veya yapay zekaya özgü riskleri kapsamaya yardımcı olmak için savunma önlemlerine ek katmanlar eklenmesi gerekebilir. Örneğin, veri şifreleme, anahtarların erişimini belirli rollerle sınırlandırarak yapay zeka sistemlerini yetkisiz erişimden korumaya yardımcı olabilir, ancak yapay zeka

modellerini ve bunların temelini oluşturan verileri çalınmalara veya tahrif edilmelere karşı korumak için de kullanılması gerekebilir.

Yapay zekaya özgü tehditler, düzenlemeler vb. gibi nedenlerle hangi güvenlik kontrollerinin eklenmesi gerektiğini belirlemek için bir analiz gerçekleştirin

Oluşturulan ekibi kullanarak, mevcut kontrollerinizin yapay zeka kullanım alanınızla nasıl örtüştüğünü gözden geçirin, bu kontrollerin amaca uygunluk değerlendirmesini yapın ve ardından boşluk alanlarını belirlemek için bir plan oluşturun. Tüm bunlar yapıldıktan sonra, riski azaltıp azaltmadıklarına ve amaçlanan yapay zeka kullanımınızı ne kadar iyi ele aldıklarına bağlı olarak bu kontrollerin etkinliğini de ölçün.

Tedarik zinciri varlıklarını, kod ve eğitim verilerini depolamak ve izlemek için hazırlanın

Yapay zeka sistemleri kullanan kuruluşlar, tedarik zinciri varlıklarını, kodlarını ve eğitim verilerini depolamaya ve izlemeye hazırlanmalıdır. Bu, tüm varlıkların tanımlanması, kategorize edilmesi ve güvenliğinin sağlanmasının yanı sıra yetkisiz erişim veya kullanımın izlenmesini de içerir. Kuruluşlar bu adımları atarak yapay zeka sistemlerini saldırılara karşı korumaya yardımcı olabilirler.

Veri denetiminizin ve yaşam döngüsü yönetiminizin ölçeklenebilir ve yapay zekaya uyarlanmış olduğundan emin olun.

Takip ettiğiniz veri denetimi tanımına bağlı olarak, veri denetimi için altı adede kadar karar alanı vardır:

- Veri kalitesi
- Veri güvenliği
- Veri mimarisi
- Meta Veriler
- Veri yaşam döngüsü
- Veri depolama

Yapay zeka veri denetim süreci her zamankinden daha önemli hale gelecektir. Örneğin, yapay zeka modellerinin etkinliğinin temel dayanaklarından biri eğitim veri setleridir. Veri setleri söz konusu olduğunda, yaşam döngüsünün bir parçası olarak güvenliğe güçlü bir vurgu yaparak uygun bir yaşam döngüsü yönetim sistemine sahip olduğunuzdan emin olun (yani, verilerin oluşturulmasından tamamen yok edilmesine kadar yaşam döngüsü boyunca yerleştirilmiş güvenlik önlemlerine sahip olun). Bu noktada veri kökeni de ayrıca önemli bir rol oynayacak ve gizlilik ve fikri hak mülkiyetine ilişkin soruların yanıtlanmasına yardımcı olacaktır. Verileri kimin oluşturduğunu, nereden geldiğini ve veri kümesini neyin oluşturduğunu biliyorsanız, yukarıda belirtilen konulardaki soruları yanıtlamak çok daha kolaylaşır.

Yapay zekanın benimsenmesi arttıkça, kuruluşunuzun başarısı muhtemelen bu karar alanlarını çevik bir şekilde ölçeklendirmeye bağlı olacaktır. Bu çabayı desteklemeye yardımcı olmak için, veri denetimi stratejinizi çapraz fonksiyonlu bir ekiple gözden geçirmek ve yapay zekadaki gelişmeleri yansıtmasını sağlamak için güçlü bir şekilde ayarlamak çok önemlidir.

Devam ettirin ve yeniden eğitin

Yapay zekadan değil, insanlardan bahsediyoruz. Birçok kuruluş için güvenlik, gizlilik ve uyumluluk alanlarında doğru yetenekleri bulmak çok yıllı bir yolculuk olabilir. Bu yetenekleri elinizde tutmak için atacağınız adımlar başarınıza katkı sağlayabilir, zira bu yeteneklerin yapay zeka ile ilgili becerilerle yeniden eğitilmesi, belirli bir yapay zeka bilgisine sahip olup da edinilmesi daha uzun süre alan kurumsal bilgiden yoksun olan yetenekleri dışarıdan işe almaktan daha hızlı olabilir.

Yapay zekayı bir kurumun tehdit ortamına dahil etmek için tespit ve müdahaleyi genişletin

Yapay zekâ kullanım senaryoları, kullanılan yapay zeka türleri vb. için önemli olan tehditler hakkında anlayış geliştirin

Yapay zeka sistemlerini kullanan kuruluşlar, kendi özel yapay zeka kullanım senaryolarıyla ilgili tehditleri anlamalıdır. Bu, kullandıkları yapay zeka türlerini, yapay zeka sistemlerini eğitmek için kullandıkları verileri ve bir güvenlik ihlalinin potansiyel sonuçlarını anlamak demektir. Kuruluşlar bu adımları atarak yapay zeka sistemlerini saldırılardan korumaya odaklanabilirler.

Yapay zekaya yönelik saldırılara ve ayrıca yapay zeka çıktısı nedeniyle ortaya çıkan sorunlara karşılık vermeye hazırlanın

Yapay zeka sistemleri kullanan kuruluşlar, güvenlik olaylarını tespit etmek ve bunlara müdahale etmek için bir plana sahip olmalı ve yapay zeka sistemlerinin zararlı veya taraflı kararlar verme risklerini azaltmalıdır. Kuruluşlar bu adımları atarak yapay zeka sistemlerini ve kullanıcılarını zarardan korumaya katkıda bulunabilirler.

AI çıktısına odaklanın, özellikle Gen AI için - içerik güvenliği politikalarını uygulamaya hazırlanın

Gen AI, metinden görsellere ve videolara kadar çeşitli içerikler oluşturmak için güçlü bir araçtır. Ancak, bu güç, kötüye kullanma potansiyelini de beraberinde getirmektedir. Örneğin, Gen AI nefret söylemi veya şiddet içeren görüntüler gibi zararlı içerikler oluşturmak için kullanılabilir. Bu riskleri azaltmak için içerik güvenliği politikalarını uygulamaya hazırlanmak önemlidir.

Kötüye kullanım politikanızı ve olay müdahale süreçlerinizi, kötü amaçlı içerik oluşturma veya yapay zeka gizlilik ihlalleri gibi yapay zekaya özgü olay türlerine göre ayarlayın

Yapay zeka sistemleri daha karmaşık ve yaygın hale geldikçe, kötüye kullanım politikanızı ilgili durumlarla başa çıkacak şekilde ayarlamanız ve ardından olay müdahale süreçlerinizi yapay zekaya özgü olay türlerini hesaba katacak şekilde ayarlamanız önemlidir. Bu tür olaylar arasında kötü niyetli içerik oluşturma, yapay zeka gizlilik ihlalleri, yapay zeka önyargısı ve sistemin genel olarak kötüye kullanılması yer alabilir.

Mevcut ve yeni tehditlere ayak uydurmak için savunmaları otomatize edin

Yapay zeka sistemlerini, eğitim veri hatlarını vb. korumayı hedefleyen yapay zeka güvenlik yeteneklerinin listesini belirleyin.

Yapay zeka güvenlik teknolojileri, yapay zeka sistemlerini veri ihlalleri, kötü niyetli içerik oluşturma ve yapay zeka önyargısı gibi çeşitli tehditlerden koruyabilir. Bu teknolojilerden bazıları geleneksel veri şifreleme, erişim kontrolü, yapay zeka ile artırılabilen denetim ve eğitim verisi koruması ve model koruması gerçekleştirebilen daha yeni teknolojileri içerir.

Yapay zeka tehditlerine karşı koymak için yapay zeka savunmalarını kullanın, ancak gerektiğinde karar vermek için insanları işin içinde tutun

Yapay zeka, veri ihlalleri, kötü niyetli içerik oluşturma ve yapay zeka önyargısı gibi yapay zeka tehditlerini tespit etmek ve bunlara yanıt vermek için kullanılabilir. Bununla birlikte, neyin tehdit oluşturduğunun ve buna nasıl yanıt verileceğinin belirlenmesi gibi önemli kararlar için insanların işin içinde kalması gerekir. Bunun nedeni, yapay zeka sistemlerinin önyargılı olabilmesi veya hata yapabilmesidir ve yapay zeka sistemlerinin etik ve sorumlu bir şekilde kullanılmasını sağlamak için insan gözetimi gereklidir.

Zaman alan görevleri otomatikleştirmek, uğraşları azaltmak ve savunma mekanizmalarını hızlandırmak için yapay zekayı kullanın

Yapay zekanın kullanım alanlarına bakıldığında daha basit bir yaklaşım gibi görünse de, zaman alan görevleri hızlandırmak için yapay zekanın kullanılması nihayetinde daha hızlı sonuçlar elde edilmesini sağlayacaktır. Örneğin, bir kötü amaçlı yazılım dosyasını tersine mühendisliğe tabi tutmak zaman alıcı olabilir. Analist bu bilgiyi kullanarak sistemden bu eylemleri arayan bir *YARA* kuralı oluşturmasını isteyebilir. Bu örnekte, savunma pozisyonu için harcanan emekte doğrudan bir azalma ve daha hızlı bir çıktı söz konusudur.

Kurum genelinde istikrarlı güvenlik sağlamak için platform düzeyinde kontrolleri uyumlu hale getirin

Yapay zeka kullanımını ve yapay zeka tabanlı uygulamaların yaşam döngüsünü gözden geçirin

Yapay zeka kuruluşunuzda daha yaygın olarak kullanılmaya başlandığında, güvenlik risklerini belirlemek ve azaltmak için kullanımın periyodik olarak gözden geçirilmesi yönünde bir süreç uygulamalısınız. Bu, kullanılan yapay zeka modellerini ve uygulamalarının türlerini, yapay zeka modellerini eğitmek ve çalıştırmak için kullanılan verileri, yapay zeka modellerini ve uygulamalarını korumak için uygulanan güvenlik önlemlerini, yapay zeka güvenlik olaylarını izleme ve bunlara yanıt verme prosedürlerini ve tüm çalışanlar için yapay zeka güvenlik riski farkındalığı ve eğitimini değerlendirmeyi içerir.

Araçlar ve frameworkler üzerinde standartlaşmaya çalışarak kontrollerin parçalanmasını önleyin

Yukarıdaki süreci uygulayarak mevcut araçları, güvenlik kontrollerini ve halihazırda yürürlükte olan uygulamaları daha iyi anlayabilirsiniz. Aynı zamanda, parçalanmayı azaltmaya yardımcı olmak için kuruluşunuzun güvenlik ve uyumluluk kontrolleri için farklı veya örtüşen frameworklere sahip olup olmadığını incelemek önemlidir. Parçalanma karmaşıklığı artıracak ve önemli ölçüde çakışmalar oluşturarak maliyetleri ve verimsizlikleri artıracaktır. Frameworklerinizi ve kontrollerinizi uyumlu hale getirerek ve bunların yapay zeka kullanım bağlamınıza uygulanabilirliğini anlayarak, parçalanmayı sınırlandıracak ve riski azaltmak için kontrollere *'doğru uyum' ('right fit')* yaklaşımı sağlayacaksınız. Bu kılavuz öncelikle mevcut kontrol frameworklerine ve standartlarına atıfta bulunmaktadır, ancak aynı ilke (örneğin, genel sayıyı mümkün olduğunca küçük tutmaya çalışın) yapay zeka için yeni ve gelişmekte olan frameworkler ve standartlar için de geçerli olacaktır.

Azaltmaları ayarlamak ve yapay zeka dağıtımı için daha hızlı geri bildirim döngüleri oluşturmak için kontrolleri uyarlayın

Yapay zeka destekli ürün ve yeteneklerin korunmasını ve güvenliğini artırmak için Kırmızı Ekip tatbikatları gerçekleştirin

Kırmızı Takım *(red team)* tatbikatları, etik bilgisayar korsanlarından *(ethical hackers)* oluşan bir ekibin bir kuruluşun sistem ve uygulamalarındaki güvenlik açıklarından yararlanmaya çalıştığı bir güvenlik testi yöntemidir. Bu, kuruluşların yapay zeka sistemlerindeki güvenlik risklerini kötü niyetli aktörler tarafından istismar edilmeden önce belirlemelerine ve azaltmalarına yardımcı olabilir.

Komut istemi enjeksiyonu (prompt injection), veri zehirleme ve kaçırma saldırıları gibi yeni saldırıları takip edin

Bu saldırılar, önemli verileri sızdırmak, yanlış tahminlerde bulunmak veya operasyonları aksatmak gibi zararlara neden olmak için yapay zeka sistemlerindeki güvenlik açıklarından faydalanabilir. Kuruluşlar, en son saldırı yöntemleri konusunda güncel kalarak bu riskleri azaltmak için adımlar atabilirler.

Makine öğrenmesi tekniklerini uygulayarak tespit doğruluğunu ve hızını artırma

Yapay zeka kullanımının güvenliğini sağlamaya odaklanmak kritik öneme sahip olsa da yapay zeka, kuruluşların geniş ölçekte daha iyi güvenlik sonuçları elde etmesine de yardımcı olabilir (bkz. 3. Adımda ki referans). Örneğin, yapay zeka destekli tespit ve müdahale yetenekleri, herhangi bir kuruluş için önemli bir unsur olabilir. Aynı zamanda, ilgili yapay zeka sistemlerini, süreçlerini ve kararlarını denetlemek için insanları döngüde tutmak çok önemlidir. Zaman içinde bu çaba, yapay zeka tabanlı korumaları iyileştirmek, esas modeller için veri setlerinin eğitimini ve ince ayarını güncellemek ve yapının korunması için kullanılan makine öğrenimi modellerini geliştirmek için sürekli öğrenmeyi teşvik edebilir. Bu da kurumların tehdit ortamı geliştikçe saldırılara stratejik olarak yanıt vermesini sağlayacaktır. Sürekli

öğrenme, doğruluğu artırmak, gecikmeyi azaltmak ve korumaların verimliliğini artırmak için de kritik öneme sahiptir.

Bir geri bildirim döngüsü oluşturun

Yukarıdaki üç unsurun etkisini en üst düzeye çıkarmak için bir geri bildirim döngüsü oluşturmak çok önemlidir. Örneğin, Kırmızı Ekibiniz yapay zeka sisteminizi kötüye kullanmanın bir yolunu keşfederse, bu bilgi yalnızca düzeltmeye odaklanmak yerine savunmaları iyileştirmeye yardımcı olmak için kuruluşunuza geri gönderilmelidir.
Benzer şekilde, kuruluşunuz yeni bir saldırı vektörü keşfederse, sürekli öğrenmenin bir parçası olarak eğitim veri setinize geri bildirimde bulunmalıdır. Geri bildirimin iyi bir şekilde kullanılmasını sağlamak için, çeşitli besleme yollarını göz önünde bulundurmak ve geri bildirimin korumalarınıza ne kadar hızlı dahil edilebileceğini iyi anlamak önemlidir.

Yapay Zeka sistemlerinin iş süreçleri çevresindeki risklerini bağlamsallaştırın (ya da bir zemine oturtun)

Model bir risk yönetimi frameworku oluşturun ve yapay zeka ile ilgili riskleri anlayan bir ekip kurun

Kuruluşlar, yapay zeka modelleriyle ilişkili riskleri tanımlamak, değerlendirmek ve azaltmak için bir süreç geliştirmelidir. Ekip; yapay zeka, güvenlik ve risk yönetimi uzmanlarından oluşmalıdır.

Üçüncü taraf çözümlerinden ve hizmetlerinden yararlanırken belirli kullanım durumlarına ve paylaşılan sorumluluğa dayalı olarak yapay zeka modellerinin ve risk profillerinin bir envanterini oluşturun

Kuruluşlar, üçüncü taraf çözüm ve hizmetlerinden yararlanırken kapsamlı bir yapay zeka modelleri envanteri oluşturmalı ve risk profillerini belirli kullanım durumlarına, veri hassasiyetine ve paylaşılan sorumluluğa göre değerlendirmelidir. Bu, kullanılan tüm yapay zeka modellerinin tanımlanması, her modelle ilişkili belirli risklerin anlaşılması ve bu riskleri azaltmak için güvenlik kontrollerinin uygulanması ile birlikte net rol ve sorumluluklara sahip olunması anlamına gelir.

Model geliştirme, uygulama, izleme ve doğrulamaya rehberlik etmek için makine öğrenmesi modeli yaşam döngüsü boyunca veri gizliliği, siber risk ve üçüncü taraf risk politikaları, protokolleri ve kontrolleri uygulayın

Kuruluşlar model geliştirme, uygulama, izleme ve doğrulamaya rehberlik etmek üzere makine öğrenimi modeli yaşam döngüsü boyunca veri gizliliği, siber risk ve üçüncü taraf risk politikaları, protokolleri ve kontrolleri uygulamalıdır. Bu, makine öğrenmesi modeli yaşam döngüsünün her aşamasıyla ilişkili belirli riskleri ele alan politikalar, protokoller ve kontroller geliştirmek ve uygulamak anlamına gelir.
Gereksiz parçalanmaya yol açmadığınızdan emin olmak için yukarıdaki frameworkun dördüncü maddesini aklınızda bulundurun.

Yapay zekanın kurumsal kullanımını dikkate alan bir risk değerlendirmesi gerçekleştirin

Kuruluşlar, yapay zeka kullanımıyla ilişkili riskleri belirleyip değerlendirmeli ve bu riskleri azaltmak için güvenlik kontrolleri uygulamalıdır. Kuruluşlar ayrıca model çıktısının açıklanabilirliği ve sapmanın izlenmesi dahil olmak üzere kontrol etkinliğini izlemek ve doğrulamak için güvenlik uygulamalarını da dikkate almalıdır. Adım 1 ve 2'de atıfta bulunulduğu gibi, bu çabayı desteklemek için çapraz fonksiyonlu, çok işlevli bir ekip oluşturmak ve ilgili kullanım durumları hakkında daha derin bir anlayış oluşturmak önemlidir. Kuruluşlar, çalışmalarına rehberlik etmek için mevcut risk değerlendirme frameworklerini kullanabilirler, ancak yeni ortaya çıkan yapay zeka risk yönetimi frameworklerini ele almak için muhtemelen yaklaşımlarını artırmaları veya uyarlamaları gerekecektir.

Yapay zeka sistemlerini kimin geliştirdiğine, model sağlayıcı tarafından geliştirilen modelleri kimin dağıttığına, modelleri kimin düzenlediğine veya kullanılmaya hazır çözümleri kimin kullandığına bağlı olarak yapay zekayı güvence altına almak için paylaşılan sorumluluğu birleştirin

Yapay zeka sistemlerinin güvenliği, bu sistemlerin geliştiricileri, dağıtıcıları ve kullanıcıları arasında müşterek bir sorumluluktur. Her bir tarafın spesifik sorumlulukları, yapay zeka sisteminin geliştirilmesi ve dağıtımındaki rollerine bağlı olarak değişecektir. Örneğin, yapay zeka sistemi geliştiricileri, tasarım açısından güvenli yapay zeka sistemleri geliştirmekten sorumludur. Bu, güvenli kodlama uygulamalarının kullanılmasını, yapay zeka modellerinin temiz veriler üzerinde eğitilmesini ve yapay zeka sistemlerini saldırılardan korumak için güvenlik kontrollerinin uygulanmasını içerir.

Yapay zeka kullanım durumlarını risk toleranslarıyla eşleştirin

Bu, her bir yapay zeka kullanım alanınızla ilişkili belirli riskleri anlamak ve bu riskleri azaltmak için gerekli güvenlik önlemlerini uygulamak anlamına gelir. Örneğin, sağlık veya finans gibi insanların hayatlarını önemli ölçüde etkileyebilecek kararların alınmasına yardımcı olmak için kullanılan yapay zeka sistemlerinin, pazarlama veya müşteri hizmetleri gibi daha az acil görevler için kullanılan yapay zeka sistemlerinden daha yoğun bir şekilde güvenliğinin sağlanması gerekecektir.

Sonuç

Yapay zeka dünyanın hayal gücünü ele geçirdi ve birçok kuruluş bu yeni teknolojiden yararlanarak yaratıcılığı artırma ve üretkenliği geliştirme olanakları görüyor. Google olarak, on yılı aşkın bir süredir yapay zekayı ürün ve hizmetlerimize dahil ediyoruz ve bu konuya cesaretli ve sorumlu bir şekilde yaklaşmaya kararlıyız.

SAIF *(Güvenli Yapay Zeka Ortamı)*, yapay zeka sistemlerini geliştirirken ve dağıtırken güvenlik çıtasını yükseltmeye ve genel riski azaltmaya yardımcı olmak için tasarlanmıştır. Yapay zekanın varsayılan olarak güvenli *(secure-by-default)* bir şekilde ilerlemesini sağlamak için iş birliği içinde çalışmak önemlidir. Müşteriler, iş ortakları, endüstri ve hükümetlerin desteğiyle, bu frameworkun temel unsurlarını geliştirmeye ve kuruluşların geniş ölçekte daha iyi güvenlik sonuçları elde etmelerine yardımcı olmak için pratik ve uygulanabilir kaynaklar sunmaya devam edeceğiz.

AISecLab Çeviri Ekibi

Mentor: Cihan Özhan
Furkan Berk Koçoğlu
Şevval Ayşe Kenar