Denetci.org

REFERANS İLKELERİ

Yapay Zekâ Güvence İlkeleri

Rİ-YAZ-5

Sürüm 1.0 · 11.08.2026

www.denetci.org

Yapay Zekâ Güvence İlkeleri

Yapay zekâ güvencesi, yapay zekâ kullanan sistemlerin hangi amaçla, hangi yetki ve veri sınırları içinde ve hangi belirsizlikler kabul edilerek çalıştığını ele alır. Amaç yapay zekâyı hatasız, tarafsız veya bütünüyle öngörülebilir hale getirmek değil; modelin sınırlılıklarının oluşturabileceği etkinin tasarımla sınırlandırıldığı ve sisteme ilişkin güven iddialarının gerçek davranış üzerinden doğrulanabildiği bir yapı oluşturmaktır.

İlke 1: Yapay zekâ kullanımı gerçek bir ihtiyaca dayanmalıdır.

Bir işlevin yapay zekâ ile gerçekleştirilebilmesi, yapay zekânın o işlev için uygun olduğu anlamına gelmez. Daha basit, daha öngörülebilir veya daha kolay doğrulanabilir bir yaklaşım ihtiyacı yeterince karşılıyorsa yapay zekânın oluşturduğu ek belirsizlik ve karmaşıklık gerekçelendirilebilir olmalıdır.

İlke 2: Yapay zekânın amacı ve kullanım sınırları açık olmalıdır.

Bir modelin hangi problemi çözmek için kullanıldığı ve hangi kararların onun çıktısından etkilenebileceği anlaşılabilir olmalıdır. Belirli bir kullanım için değerlendirilmiş bir sistemin farklı amaçlara taşınması, önceki güvence sonuçlarının yeni bağlamda da geçerli olduğu varsayımına dayanmamalıdır.

İlke 3: Yapay zekâya duyulan güven oluşturabileceği etkiyle orantılı olmalıdır.

Düşük etkili bir öneride kabul edilebilir olan hata ve belirsizlik, insanların haklarını, mali durumlarını, güvenliğini veya kritik sistemleri etkileyen bir kullanımda kabul edilebilir olmayabilir. Beklenen güvence düzeyi modelin türünden çok hatasının doğurabileceği sonuçlara dayanmalıdır.

İlke 4: Yapay zekâ çıktısı doğrulanmış gerçek olarak kabul edilmemelidir.

Bir model akıcı, ikna edici veya yüksek güven görünümü taşıyan bir sonuç üretirken yine de yanlış, eksik veya uydurma bilgi verebilir. Sistemin sonraki kararları, model çıktısının yapısı gereği doğru olduğu varsayımına dayanmamalıdır.

İlke 5: Ortalama başarı önemli hata alanlarını görünmez hale getirmemelidir.

Yüksek genel performans belirli görevlerde, kullanıcı gruplarında veya nadir fakat kritik koşullarda ciddi başarısızlıkları gizleyebilir. Güvence, yalnızca ortalama başarıya değil kullanım açısından önemli hata biçimlerine ve bunların dağılımına dayanmalıdır.

İlke 6: Yapay zekânın belirsizliği sistemin tasarımında kabul edilmelidir.

Model eğitim veya değerlendirme kapsamının dışında kalan durumlarla karşılaşabilir ve neyi bilmediğini güvenilir biçimde her zaman belirleyemeyebilir. Sistem, belirsiz bir çıktının gereksiz biçimde kesin ve yüksek etkili bir sonuca dönüşmesini sınırlandırmalıdır.

İlke 7: Kesin olmayan bilgi karar vericiden gereksiz biçimde gizlenmemelidir.

Bir tahminin, sınıflandırmanın veya model sonucunun belirsizlik düzeyi karar açısından önemliyse sistem bunu olduğundan daha kesin göstermemelidir. Kullanıcı arayüzü ve süreç, olasılıksal bir sonucu kesin hüküm görüntüsüne dönüştürmemelidir.

İlke 8: İnsan gözetimi gerçek karar kapasitesi taşımalıdır.

Bir insanın model çıktısını görmesi veya biçimsel olarak onaylaması tek başına insan gözetimi oluşturmaz. İnsan müdahalesine güveniliyorsa ilgili kişi sonucu anlayabilecek bilgiye, değerlendirecek zamana ve gerektiğinde modele karşı çıkabilecek gerçek yetkiye sahip olmalıdır.

İlke 9: İnsan gözetimi model zayıflıklarının otomatik telafisi olarak görülmemelidir.

İnsanlar modele gereğinden fazla güvenebilir, yoğun işlem altında kararları yüzeysel biçimde onaylayabilir veya modelin hatasını fark edecek bilgiye sahip olmayabilir. Güvence, "sonunda bir insan bakıyor" varsayımına gereksiz biçimde bağlanmamalıdır.

İlke 10: Yapay zekânın eylem yetkisi gerçek görev ihtiyacıyla sınırlı olmalıdır.

Bilgi üretmek ile para göndermek, veri silmek, hesap oluşturmak veya bir üretim sistemini değiştirmek aynı güvenlik etkisini taşımaz. Bir yapay zekâ sistemi veya ajanı yalnızca görevin gerektirdiği araçlara, kaynaklara ve işlem yetkilerine sahip olmalıdır.

İlke 11: Modelin önerisi ile işlemin yetkili olması birbirine eşit kabul edilmemelidir.

Bir modelin belirli bir aracın kullanılmasını veya işlemin yapılmasını önermesi o işlemin güvenlik ve iş kurallarını karşılamış olduğu anlamına gelmez. Kritik sistemler kendi kimlik, yetki ve işlem sınırlarını modelden bağımsız olarak korumalıdır.

İlke 12: Yapay zekâ mevcut erişim sınırlarını çevresinden dolaşan yeni bir yol oluşturmamalıdır.

Bir kullanıcı doğal dil, model sorgusu veya ajan aracılığıyla normalde erişemeyeceği veri ve işlemlere ulaşamamalıdır. Yapay zekânın araya girmesi alttaki kimlik ve yetki modelini genişletmemelidir.

İlke 13: Talimatlar güvenlik sınırlarının yerine geçmemelidir.

Modelin belirli bir davranışı yapmaması için verilen doğal dil talimatları yararlı olabilir; ancak değiştirilebilir, yanlış yorumlanabilir veya başka girdiler tarafından etkilenebilir. Kritik güvenlik sınırları yalnızca modelin verilen talimata uyacağı varsayımına bırakılmamalıdır.

İlke 14: Modelin işlediği dış içerik güvenilir talimat olarak kabul edilmemelidir.

Belgeler, web içerikleri, mesajlar, veritabanı kayıtları ve diğer dış girdiler modele yönelik yönlendirici içerik taşıyabilir. Sistemin veri ile yetkili talimat arasındaki ayrımı koruması, belirli bir saldırı tekniğinin adından bağımsız temel bir güven sınırıdır.

İlke 15: Modele sunulan veri görev için gerekli kapsamla sınırlı olmalıdır.

Daha fazla bağlam model performansını artırabilir; ancak aynı zamanda daha geniş veri maruziyeti oluşturur. Modele, bağlamına veya bağlı veri erişim katmanlarına yalnızca görevin gerektirdiği bilgiler taşınmalıdır.

İlke 16: Modelin erişebildiği veri kullanıcının erişim sınırlarını genişletmemelidir.

Merkezi bir yapay zekâ sistemi birçok veri kaynağına erişebilir. Verinin sorguyla ilgili veya anlamsal olarak benzer bulunması, sorguyu yapan kişinin o veriyi görmeye yetkili olduğunu göstermez.

İlke 17: Türetilmiş veri kaynak veriden bağımsız olarak güvenlik ve mahremiyet etkisi taşıyabilir.

Özetler, vektörel temsiller, profiller, puanlar, tahminler ve diğer model çıktıları kaynak veride doğrudan bulunmayan yeni bilgi ortaya çıkarabilir. Bir verinin türetilmiş veya doğrudan okunamaz biçimde olması onun önemsiz ya da hassasiyetsiz olduğu anlamına gelmemelidir.

İlke 18: Modeli geliştiren veri kullanım amacına uygun ve yeterince güvenilir olmalıdır.

Eğitim veya uyarlama verisinin kaynağı, doğruluğu, temsil gücü ve geçmişte taşıdığı önyargılar model davranışını etkileyebilir. Büyük miktarda veri bulunması, bu verinin kullanım bağlamını doğru temsil ettiğini veya güvenilir sonuç üreteceğini garanti etmez.

İlke 19: Hizmet sunumu için verilen veri başka amaçlara kendiliğinden taşınmamalıdır.

Kullanıcıların modele verdiği içerik, yalnızca sistem tarafından işlenmiş olması nedeniyle model geliştirme, profil oluşturma veya başka ikincil kullanımlar için otomatik olarak serbest hale gelmemelidir. Farklı veri kullanım amaçları gerektiğinde birbirinden ayrıştırılabilmelidir.

İlke 20: Yapay zekâ arayüzü mevcut veri güvenliği ve mahremiyet sınırlarının dışında kalmamalıdır.

Promptlar, konuşmalar, dosya ekleri, getirilen belgeler ve model çıktıları sıradan uygulama verilerinden farklı görünse de aynı veya daha yüksek hassasiyet taşıyabilir. Yapay zekâ kullanımı mevcut veri koruma ilkelerinden muaf yeni bir veri kanalı oluşturmamalıdır.

İlke 21: Modelin davranışını yönlendiren içerik gerçek sırların saklandığı yer olmamalıdır.

Sistem talimatları ve benzeri model bağlamları çeşitli yollarla açığa çıkabilir veya davranıştan çıkarılabilir. Parola, anahtar, erişim belirteci ve benzeri gerçek güven unsurlarının korunması modelin bunları paylaşmaması talimatına bırakılmamalıdır.

İlke 22: Modelin kendisi korunması gereken bir varlık olarak ele alınmalıdır.

Model ağırlıkları, yapılandırmaları ve ilgili bileşenler kurumsal bilgi, yatırım veya özel yetenek taşıyabilir; değiştirilmesi ise sistem davranışını değiştirebilir. Modelin kopyalanması,

değiştirilmesi veya başka sürümle değiştirilmesi yalnızca yazılım operasyonu olarak görülmemelidir.

İlke 23: Üretimde hangi model ve yapılandırmanın kullanıldığı anlaşılabilir olmalıdır.

Aynı model adı altında farklı sürümler, sağlayıcı değişiklikleri veya sistem talimatları önemli ölçüde farklı davranış oluşturabilir. Güvence, değerlendirilen sistem ile gerçekte çalışan sistemin aynı olduğu varsayımına kör biçimde dayanmamalıdır.

İlke 24: Model ve yapay zekâ sistemi değişiklikleri güven durumunu değiştiren olaylar olarak görülmelidir.

Model sürümü, veri kaynağı, talimatlar, bağlı araçlar veya güvenlik katmanları değiştiğinde sistemin davranışı kod değişmeden önemli ölçüde değişebilir. Geçmiş değerlendirme yeni sistem durumuna otomatik olarak taşınmamalıdır.

İlke 25: Üçüncü taraf yapay zekâ hizmetleri kuruluşun güven sorumluluğunu ortadan kaldırmamalıdır.

Modelin başka kuruluş tarafından geliştirilmesi veya işletilmesi onun kuruluş içindeki kullanım sonucunu dışarıya taşımaz. Sağlayıcının model, veri kullanımı veya güvenlik davranışındaki değişiklikleri kuruluşun güven varsayımlarını etkileyebilmelidir.

İlke 26: Model değerlendirmesi gerçek kullanım ve kötüye kullanım koşullarını temsil etmelidir.

Genel testler ve karşılaştırmalar model hakkında yararlı bilgi sağlayabilir; ancak kuruluşun dili, verisi, kullanıcıları, bağlı araçları ve saldırı ortamı farklı olabilir. Güvence, modelin yalnızca normal kullanımda değil makul hata ve kötüye kullanım koşullarında da nasıl davrandığını değerlendirmelidir.

İlke 27: Tek bir koruma katmanı yapay zekâ davranışının nihai güvencesi olmamalıdır.

Filtreler, başka modeller ve benzeri koruma mekanizmaları önemli fayda sağlayabilir; ancak kendileri de hata yapabilir veya aynı saldırı ve belirsizliklerden etkilenebilir. Yüksek etkili güvenlik sınırları tek bir yapay zekâ tabanlı denetleyicinin kusursuzluğuna bağlı olmamalıdır.

İlke 28: Açıklama, sistem hakkında desteklemediği bir kesinlik üretmemelidir.

Bir modelin sonucu için üretilen açıklama kullanıcıya faydalı olabilir; ancak sonradan oluşturulan gerekçe modelin gerçek karar mekanizmasının eksiksiz temsili olmayabilir. Açıklanabilirlik, modelin gerçekte olduğundan daha anlaşılır veya nedensel görünmesine yol açmamalıdır.

İlke 29: Yapay zekâ kaynaklı önemli kararlar gerektiğinde sorgulanabilir olmalıdır.

Model hatası kişi veya kuruluş üzerinde ağır sonuç oluşturabiliyorsa yanlış sonucun yeniden değerlendirilmesini sağlayabilecek anlamlı yollar bulunmalıdır. İtiraz veya müdahale imkânı,

zayıf tasarımın yerine geçmemeli; kalan hata etkisini sınırlandıran ek bir güvence oluşturmalıdır.

İlke 30: Adalet yalnızca hassas özelliklerin doğrudan kullanılmamasına indirgenmemelidir.

Model başka değişkenler üzerinden aynı özelliklerle ilişkili sonuçlar üretebilir ve aynı genel performans farklı gruplarda farklı hata dağılımlarını gizleyebilir. Adalet ve ayrımcılık değerlendirmesi girdilerin adından çok gerçek çıktı ve etki dağılımına dayanmalıdır.

İlke 31: Yapay zekânın güvenilirliği çevresinin değişmeyeceği varsayımına dayanmamalıdır.

Kullanıcı davranışı, veri dağılımı, tehditler, iş kuralları ve modelin bağlı olduğu sistemler zaman içinde değişebilir. Modelin kendisi değişmese bile daha önce geçerli olan performans ve güvenlik varsayımları geçerliliğini kaybedebilir.

İlke 32: Güven kaybına işaret eden değişimler görünür hale gelebilmelidir.

Modelin girdileri, hata örüntüleri, etkileri veya kullanım biçimi değerlendirildiği koşullardan anlamlı ölçüde uzaklaşıyorsa sistem geçmiş güvenceye sınırsız biçimde dayanmamalıdır. İzleme yalnızca teknik performansı değil değişimin gerçek iş, güvenlik ve insan etkisini de anlamlandırabilmelidir.

İlke 33: Yapay zekânın başarısızlık davranışı bilinçli bir sistem kararı olmalıdır.

Model yanıt veremediğinde, hizmet kullanılamadığında veya yeterli güven sağlanamadığında sistemin nasıl davranacağı tesadüfen oluşmamalıdır. Hata durumu daha zayıf güvenlik sınırlarına veya kontrolsüz bir alternatif işleyişe otomatik geçiş üretmemelidir.

İlke 34: Yapay zekâ olmadan çalışabilecek yollar kendi güven sınırlarını korumalıdır.

Yapay zekânın kullanılamadığı durumda insan veya geleneksel sistemler devreye girebilir. Alternatif çalışma biçimi yalnızca süreklilik sağlamak adına normal kimlik, yetki, veri veya kontrol sınırlarını gereksiz biçimde kaldırmamalıdır.

İlke 35: Yapay zekânın kaynak tüketme gücü güvenlik etkisinin bir parçasıdır.

Bir model çağrısı çok büyük hesaplama, maliyet, veri kullanımı veya ardışık araç işlemleri oluşturabilir. Tek bir kullanıcı veya hatalı model davranışı ortak kaynaklar üzerinde ölçüsüz etki oluşturmamalıdır.

İlke 36: Yapay zekâ gözlemlenebilirliği yeni ve gereksiz hassas veri depoları oluşturmamalıdır.

Promptlar, yanıtlar, kullanılan belgeler ve araç çağrıları olay incelemesi için değerli olabilir; ancak kişisel veri, sır ve hassas kurumsal bilgi de içerebilir. İzlenebilirlik ihtiyacı bütün etkileşimlerin sınırsız biçimde saklanmasını gerektirmemelidir.

İlke 37: Önemli yapay zekâ işlemleri sonradan anlamlı ölçüde açıklanabilir olmalıdır.

Gerektiğinde hangi model ve sistem durumunun, hangi önemli veri ve araçlarla hangi sonucu oluşturduğu anlaşılabilmelidir. Bunun için modelin bütünüyle deterministik olması gerekmez; ancak kritik bir sonucun bağlamı tamamen kaybolmamalıdır.

İlke 38: Yapay zekâ güvencesi beyanlar ve ürün özellikleri yerine gerçek davranış üzerinden doğrulanabilmelidir.

Model kartı, etik beyan, karşılaştırma sonucu, sağlayıcı güvencesi, güvenlik filtresi veya risk değerlendirmesi tek başına güvence oluşturmaz. Gerçek güvence; sistemin amaçlanan ve makul kötüye kullanım koşullarında nasıl davrandığı, hangi veri ve araçlara fiilen erişebildiği, kullanıcıların yetkilerini aşıp aşamadığı, önemli hataların etkisinin sınırlandırılıp sınırlandırılamadığı ve sistem değiştikçe güven varsayımlarının yeniden geçerlilik kazanıp kazanmadığı üzerinden değerlendirilebilmelidir.