Yapay Zekâ:
Etik, Sürdürülebilirlik ve Açıklanabilirlik

Editör
Vesile Gül BAŞER

BIDGE

BİDGE Yayınları

Yapay Zekâ: Etik, Sürdürülebilirlik ve Açıklanabilirlik

Editör: Dr. Öğr. Üyesi Vesile Gül BAŞER

ISBN: 978-625-8673-36-4

1. Baskı

Sayfa Düzeni: Gözde YÜCEL

Yayınlama Tarihi: 20.12.2025

BİDGE Yayınları

Sertifika No: 71374
Yayın hakları © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

İÇİNDEKİLER

YAPAY ZEKÂ ETİĞİ: İLKELER, RİSKLER VE YÖNETİŞİM YAKLAŞIMLARI

VESİLE GÜL BAŞER[1]

1. Giriş

Yapay zekâ (YZ), insan bilişsel süreçlerini taklit edebilen, veriden öğrenerek karar verebilen ve zamanla performansını iyileştirebilen sistemleri tanımlayan disiplinler arası bir araştırma ve uygulama alanıdır. Son yıllarda, makine öğrenmesi, derin öğrenme ve büyük veri analitiği alanlarında kaydedilen hızlı ilerlemeler, yapay zekâ sistemlerinin sadece teknik araçlar olmaktan öteye geçerek, sosyal yapıları, kurumsal karar alma mekanizmalarını ve bireylerin günlük yaşamlarını doğrudan etkileyen aktörler haline gelmesine yol açmıştır (Russell ve Norvig, 2021). Bu dönüşüm, yapay zekânın ne kadar etkili çalıştığına dair soruların yanı sıra, nasıl, kimler için ve hangi değerlere göre çalıştığına dair soruların da ortaya çıkmasına neden olmuştur.

YZ teknolojilerinin kritik birçok alanda (sağlık, eğitim, finans, güvenlik ve kamu yönetimi, vb) yaygın biçimde kullanılmaya

[1] Dr. Öğrt. Üyesi , Burdur Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi Orcid: 0000-0002-0752-9498

başlanması, etik sorunların görünürlüğünü belirgin biçimde artırmıştır. Özellikle otomatik karar verme sistemlerinin bireylerin yaşamını doğrudan etkileyen sonuçlar üretmesi (işe alım süreçleri, tıbbi tanılar ve risk analizleri) ve algoritmik çıktılara artan bağımlılık, etik değerlendirmeleri kaçınılmaz hâle getirmiştir (Floridi vd., 2018). Bu çalışmada yapay zekâ etiği, yalnızca mühendislik veya bilgisayar bilimi perspektifiyle değil; felsefe, hukuk, sosyoloji ve eğitim bilimleri gibi alanların katkısını gerektiren çok boyutlu bir olgu olarak değerlendirilmektedir.

Yapay zekâ sistemlerine yönelik etik tartışmaların temelinde, bu sistemlerin karar mekanizmalarındaki şeffaflık eksikliği ve algoritmik süreçlerin insan idraki açısından taşıdığı anlaşılmazlık önemli bir rol oynamaktadır. Özellikle derin öğrenme tabanlı modellerin "kara kutu" olarak tanımlanan yapısı, sistemin hangi girdileri ne şekilde değerlendirdiğinin açıklanmasını zorlaştırmakta ve bu durum hesap verebilirlik sorunlarını beraberinde getirmektedir (Burrell, 2016). Karar süreçlerindeki açıklanabilirlik sorunu, algoritmik hataların ve yanlılıkların teşhis edilmesini zorlaştırırken, kullanıcı nezdindeki güven ortamını da tahrip etmektedir.

Bir diğer önemli unsur ise YZ sistemlerinin büyük ölçüde tarihsel verilere dayalı öğrenme süreçleri, mevcut toplumsal eşitsizliklerin ve önyargıların algoritmik yapılara taşıma riskini ortaya çıkarabilmektedir. Literatürde, cinsiyet, etnik köken veya sosyoekonomik durum gibi değişkenler üzerinden ortaya çıkan algoritmik ayrımcılık örnekleri, YZ'nin tarafsız olduğu yönündeki varsayımları temelden sorgulatmıştır (O'Neil, 2016; Noble, 2018). Bu durum, etik tartışmaları yalnızca teknik doğruluk veya performans ölçütleriyle sınırlı olmaktan çıkararak adalet, eşitlik ve insan hakları eksenine taşımıştır.

Yapay zekânın etiğine yönelik akademik ilginin artışı, uluslararası kuruluşların ve politika yapıcıların da bu alana

yönelmesine zemin hazırlamıştır. OECD, UNESCO ve Avrupa Birliği gibi kurumlar tarafından yayımlanan etik ilke belgeleri, yapay zekânın insan merkezli, güvenilir ve sorumlu bir biçimde geliştirilmesi gerektiğini vurgulamaktadır (OECD, 2019; UNESCO, 2021). Bu belgelerle birlikte yapay zekâ etiği, soyut bir söylem olmaktan çıkmış; yerini kurumsal bir yönetişim mimarisine ve denetlenebilir bir düzenleme zeminine bırakmıştır.

2. Etik Kavramı ve Teknoloji Etiği

Etik; insan eylemlerini doğru-yanlış, iyi-kötü ve adil-adaletsiz gibi normatif kategoriler ışığında inceleyen köklü bir felsefi disiplindir. Güncel etik tartışmalar, bireysel ahlaki yargıların sınırlarını aşarak; toplumsal değerler, kurumsal sorumluluklar ve kamusal yarar gibi makro düzeydeki dinamikleri odağına alır. Bu yönüyle etik; yalnızca öznel eylemlerin değil, aynı zamanda sistemlerin, kurumların ve kompleks teknolojik yapıların meşruiyetini sorgulamaya olanak tanıyan analitik bir zemin inşa etmektedir (Beauchamp ve Childress, 2019).

Teknolojik gelişmelerin hız kazanmasıyla birlikte etik tartışmalar da yeni boyutlar kazanmıştır. Özellikle dijital teknolojiler, insan eylemlerinin sınırlarını genişletirken, bu eylemlerin sonuçlarını öngörmeyi zorlaştırmış ve klasik etik yaklaşımların yeniden yorumlanmasını gerekli kılmıştır. Bu bağlamda ortaya çıkan teknoloji etiği, teknolojik yapıların tasarımı, geliştirilmesi ve kullanımının ahlaki sonuçlarını inceleyen uygulamalı etik alanı olarak tanımlanmaktadır (Mitcham, 2014). Teknoloji etiği, teknolojinin değerlerden bağımsız olmadığı varsayımını temel almakta ve teknik kararların aynı zamanda etik tercihler içerdiğini savunmaktadır.

Teknoloji etiği literatüründe sıkça vurgulanan temel yaklaşımlardan biri, teknolojik determinizm eleştirisidir. Teknolojik

determinizm, teknolojinin toplumsal yapıları kaçınılmaz biçimde şekillendirdiğini öne sürerken, etik perspektif bu yaklaşımı sorgulayarak insan iradesi, tasarım tercihleri ve yönetişim mekanizmalarının belirleyici rolüne dikkat çeker (Verbeek, 2011). Bu çerçevede teknoloji, insan değerlerinden bağımsız, nötr bir araç olarak değil; belirli normları, öncelikleri ve güç ilişkilerini yansıtan sosyo-teknik bir yapı olarak ele alınmaktadır.

Yapay zekâ teknolojileri, teknoloji etiği tartışmalarını daha da karmaşık hâle getiren özgün özellikler taşımaktadır. Öğrenen sistemlerin özerk kararlar üretebilmesi, insan müdahalesinin azalması ve sonuçların geniş kitleleri etkileyebilmesi, etik değerlendirmelerin yalnızca kullanım aşamasında değil, tasarım ve geliştirme süreçlerinde de yapılmasını zorunlu kılmaktadır. Bu durum, etik sorumluluğun yalnızca son kullanıcıya değil; veri sağlayıcılar, yazılım geliştiriciler, kurumlar ve politika yapıcılar arasında paylaşıldığını göstermektedir (Floridi, 2019). Bu çok aktörlü yapı, yapay zekâ etiğini klasik teknoloji etiğinden ayıran temel özelliklerden biridir.

Etik kuramlar açısından bakıldığında, YZ ve teknoloji etiği tartışmaları genellikle üç ana yaklaşım çerçevesinde ele alınmaktadır: deontolojik etik, sonuçsalcı (faydacı) etik ve erdem etiği. Deontolojik yaklaşım, teknolojik uygulamaların evrensel ilkelere ve haklara uygunluğunu vurgularken; sonuçsalcı yaklaşım, ortaya çıkan fayda ve zararların dengelenmesine odaklanmaktadır. Erdem etiği ise teknolojiyi tasarlayan ve kullanan aktörlerin niyetleri, sorumluluk bilinci ve mesleki erdemleri üzerinde durmaktadır (Johnson, 2015). YZ etiği literatüründe bu yaklaşımların çoğu zaman birlikte ele alındığı ve tek başına yeterli olmadığı ve bütüncül bir etik değerlendirme gerektirdiği yönünde güçlü bir uzlaşı bulunmaktadır.

Bahsedilen etik yaklaşımların her biri, yapay zekâ sistemlerinin ortaya çıkardığı karmaşık ve çok boyutlu sorunları tek başına ele almakta yetersiz kalabilmektedir. Bu nedenle YZ etiği literatüründe, farklı etik kuramların tamamlayıcı biçimde değerlendirilmesini mümkün kılan karşılaştırmalı ve bütüncül çerçeveler ön plana çıkmaktadır. Bu bağlamda, başlıca etik kuramların yapay zekâ uygulamalarına yaklaşımları Tablo 1'de karşılaştırmalı olarak özetlenmiştir.

Tablo 1. Etik kuramlar ve yapay zekâya ilişkin temel yaklaşımlar

Etik Kuram	Temel İlke	YZ Bağlamındaki Odak Noktası
Deontoloji	Ödev, haklar	Gizlilik, insan onuru
Faydacılık	Sonuçların faydası	Performans, toplam yarar
Erdem Etiği	Karakter, niyet	Sorumlu tasarım
Bakım Etiği	İlişkiler, bağlam	Savunmasız gruplar
Adalet Yaklaşımları	Eşitlik, hakkaniyet	Ayrımcılık, önyargı

Teknoloji etiği kapsamında öne çıkan bir diğer önemli kavram, etik tasarım anlayışıdır. Bu yaklaşım, etik ilkelerin sonradan eklenen kısıtlar olarak değil, teknolojik sistemlerin en başından itibaren tasarım sürecine entegre edilmesi gerektiğini savunur. Özellikle YZ sistemlerinde etik tasarım, veri seçimi, model geliştirme, test süreçleri ve dağıtım aşamalarının her birinde normatif ve bağlama duyarlı değerlendirmelerin yapılmasını gerektirmektedir (van den Hoven vd., 2015). Böylece etik, teknik inovasyonu sınırlayan bir unsur olmaktan ziyade, sürdürülebilir ve toplumsal olarak kabul edilebilir yeniliklerin önünü açan bir rehber işlevi görmektedir.

3. Yapay Zekâ Sistemlerinin Temel Özellikleri ve Etik Riskler

Yapay zekâ sistemleri, geleneksel yazılımların aksine önceden tanımlanmış sabit kurallar yerine, devasa veri setlerinden örüntüler türeterek karar üretir. Bu öğrenme merkezli mimari, sisteme yüksek performans ve esneklik kazandırsa da beraberinde

özgün etik risk alanları getirmektedir. Bilhassa özerklik düzeyi, ölçeklenebilirlik, karar süreçlerindeki opaklık ve veri bağımlılığı gibi karakteristikler; yapay zeka sistemlerinin etik değerlendirme düzleminde klasik yazılımlardan niteliksel olarak ayrışmasına neden olmaktadır (Floridi vd., 2018).

YZ sistemlerinin en belirgin özelliklerinden biri, otomatik karar verme kapasitesidir. Bu sistemler, insan müdahalesi olmaksızın ya da sınırlı denetim altında, bireylerin yaşamını etkileyen sonuçlar üretebilmektedir. Otomatik kararlar hız ve verimlilik avantajı sağlasa da, hatalı veya adaletsiz kararların çok kısa sürede geniş kitlelere yayılabilmesi ciddi etik riskler doğurmaktadır. Özellikle insan yargısının tamamen devre dışı bırakıldığı senaryolarda, sorumluluğun kime ait olduğu belirsizleşmektedir (Citron ve Pasquale, 2014).

Otomatik karar verme ile doğrudan ilişkili bir diğer temel özellik, YZ sistemlerinin veri bağımlı yapısıdır. YZ modelleri, eğitildikleri verilerin niteliğini ve sınırlılıklarını doğrudan yansıtmaktadır. Tarihsel verilerde yer alan toplumsal önyargılar, eşitsizlikler ve temsil sorunları, YZ sistemleri aracılığıyla yeniden üretilebilmekte ve pekiştirilebilmektedir. Literatürde bu durum, "algoritmik önyargı" olarak tanımlanmakta ve YZ'nin nesnel olduğu yönündeki yaygın algıyı sorgulamaktadır (Barocas ve Selbst, 2016).

YZ sistemlerinin etik açıdan sorunlu hâle gelmesine yol açan bir diğer önemli unsur, karar süreçlerinin açıklanabilirliğinin sınırlı olmasıdır. Özellikle derin öğrenme modelleri, yüksek doğruluk oranlarına rağmen, hangi girdilerin hangi gerekçelerle belirli çıktılara yol açtığını açık biçimde ortaya koymakta zorlanmaktadır. Bu durum, hem kullanıcılar hem de düzenleyici kurumlar açısından temel bir etik ve yönetişim sorunu oluşturmaktadır. Açıklanabilirliğin sınırlı olması, hataların tespit edilmesini

zorlaştırmakta ve hesap verebilirlik ilkesini zayıflatmaktadır (Doshi-Velez ve Kim, 2017).

YZ sistemlerinin ölçeklenebilirliği, etik risklerin etkisini artıran bir başka kritik faktördür. Bir YZ modeli, çok kısa sürede milyonlarca kullanıcıya uygulanabilmekte ve tekil bir tasarım hatası veya etik ihlal, geniş çaplı toplumsal sonuçlar doğurabilmektedir. Bu durum, bireysel etik hataların hızla sistemik sorunlara dönüşmesine yol açmaktadır. Özellikle platform temelli YZ uygulamalarında, etik sorunların fark edilmesi ve düzeltilmesi zaman alabilmekte, bu süre zarfında önemli mağduriyetler ortaya çıkabilmektedir (Zuboff, 2019).

Ölçeklenebilirlik ile birlikte özerklik düzeyi artan YZ sistemleri, insan–makine ilişkisini de yeniden tanımlamaktadır. Karar verme yetkisinin kısmen ya da tamamen makinelere devredilmesi, insan kontrolünün sınırlarını belirsizleştirmekte ve etik sorumluluğun dağılımını karmaşık hâle getirmektedir. Bu bağlamda literatürde sıklıkla vurgulanan "insan denetimi" (human-in-the-loop) yaklaşımı, YZ sistemlerinin tamamen bağımsız aktörler hâline gelmesini önlemeye yönelik bir etik bir güvence mekanizması olarak değerlendirilmektedir (Amershi vd., 2019).

YZ sistemlerinin güvenlik açıkları ve kötüye kullanım potansiyeli de etik riskler arasında önemli bir yer tutmaktadır. Yanlış yönlendirilmiş veya kasıtlı olarak manipüle edilmiş YZ sistemleri, yanlış bilgi üretimi, gözetim, mahremiyet ihlalleri ve toplumsal güvenin zedelenmesi gibi sonuçlar doğurabilmektedir. Özellikle üretici YZ sistemlerinin yaygınlaşması, etik risklerin yalnızca hatalı kararlarla sınırlı olmadığını, aynı zamanda bilgi ekosistemini dönüştüren yapısal sorunlara işaret ettiğini göstermektedir (Bender vd., 2021). Yapay zekâ sistemlerinin temel teknik özellikleri ile bu özelliklerden kaynaklanan etik riskler arasındaki ilişki, Tablo 2'te özetlenmiştir.

Tablo 2. Yapay zekâ sistemlerinin temel teknik özellikleri ve ilişkili etik riskler

YZ'nin Temel Özelliği	Açıklama	Öne Çıkan Etik Riskler
Veri odaklı öğrenme	Büyük veri kümelerinden örüntü çıkarma	Önyargı, ayrımcılık, temsil sorunu
Otomatik karar verme	Sınırlı ya da hiç insan müdahalesi olmadan karar üretme	Sorumluluk belirsizliği, adaletsizlik
Kara kutu yapısı	Karar süreçlerinin opak olması	Şeffaflık eksikliği, hesap verebilirlik
Ölçeklenebilirlik	Kararların kısa sürede geniş kitleleri etkilemesi	Sistemik zarar, toplumsal etki
Artan özerklik	İnsan denetiminin azalması	Kontrol kaybı, etik sorumluluk karmaşası
Kötüye kullanım potansiyeli	Manipülasyon ve yanlış yönlendirme riski	Güven erozyonu, bilgi ekosistemi bozulması

Not. Tablo, yapay zekâ sistemlerinin temel teknik özellikleri ile bu özelliklerden kaynaklanan etik riskler arasındaki ilişkiyi kavramsal düzeyde özetlemektedir (bkz. Floridi vd.., 2018; Barocas & Selbst, 2016; Doshi-Velez & Kim, 2017; Amershi vd.., 2019; Bender vd.., 2021).

Yapay zekâ sistemlerinin teknik üstünlükleri, beraberinde önemli etik riskler getirmektedir. Bu riskler, yapay zekânın tamamen reddedilmesini değil; tasarım, geliştirme ve kullanım süreçlerinde etik kaygıların sistematik ve tutarlı biçimde ele alınmasını zorunlu kılmaktadır. Yapay zekâ sistemlerinin temel teknik özellikleri ile ortaya çıkan etik sorunlar arasındaki ilişkinin anlaşılması, bu teknolojilerin hangi normatif çerçeveler doğrultusunda yönlendirilmesi gerektiğine dair temel bir zemin sunmaktadır. Bu zemin, etik değerlendirmelerin soyut tartışmaların ötesine geçerek, yapay zekâ uygulamalarını yönlendiren ilke ve değerler etrafında yapılandırılmasını gerekli kılmaktadır.

4. Yapay Zekâda Temel Etik İlkeler

Etik ilkeler, YZ uygulamalarını sınırlayan dışsal kurallar olarak değil; tasarım, geliştirme ve dağıtım süreçlerine yön veren normatif çerçeveler olarak değerlendirilmektedir. Literatürde, farklı kurum ve araştırmacılar tarafından önerilen etik ilke setleri arasında terminolojik farklılıklar bulunsa da, büyük ölçüde örtüşen temel ilkelerin öne çıktığı görülmektedir (Floridi vd., 2018; OECD, 2019). Bu ilkeler, YZ sistemlerinin yalnızca teknik olarak değil, normatif olarak da değerlendirilmesini mümkün kılan referans çerçeveler sunmaktadır.

4.1. Adalet ve Ayrımcılıktan Kaçınma

Adalet ilkesi, YZ sistemlerinin bireyler ve gruplar arasında haksız, sistematik veya ayrımcı sonuçlar üretmemesini ifade etmektedir. YZ uygulamalarında adalet, yalnızca çıktılardaki eşitliği değil; veri toplama, modelleme ve karar verme süreçlerinin tamamını kapsayan çok katmanlı ve süreç temelli bir kavramdır. Özellikle eğitim, istihdam, kredi değerlendirme ve ceza adaleti gibi alanlarda kullanılan YZ sistemlerinin, tarihsel eşitsizlikleri yeniden üretme potansiyeli literatürde kapsamlı biçimde tartışılmaktadır (Barocas, Hardt, ve Narayanan, 2019).

Algoritmik ayrımcılık, çoğu zaman bilinçli bir niyetten ziyade, eksik veya dengesiz verilerden kaynaklanmaktadır. Bu durum, Yapay zekânın tarafsız olduğu yönündeki yaygın varsayımı geçersiz kılmakta ve etik değerlendirmenin teknik performans ölçütlerinin ötesine geçmesini zorunlu kılmaktadır (Noble, 2018). Bu tür veri temelli önyargıların pratikte nasıl somutlaştığı, Amazon'un işe alım algoritması vakasında açık biçimde görülmektedir.

Amazon'un İşe Alım Aracında Cinsiyet Yanlılığı 2014 yılında Amazon, özgeçmişleri (CV) tarayarak en iyi adayları puanlayan bir yapay zekâ aracı geliştirmeye başladı. Amaç işe alım süreçlerini otomatize etmekti. Ancak 2015 yılında sistemin kadın adaylara karşı önyargılı olduğu fark edildi.
Sorun: Model, şirkete son 10 yılda yapılan başvurularla eğitilmişti. Teknoloji sektöründeki erkek egemenliği nedeniyle, veri setinin çoğu erkek adaylardan oluşuyordu.
Sonuç: Yapay zekâ, "erkek olmayı" bir başarı kriteri olarak öğrendi. İçinde "kadın" geçen ifadeleri (örn. "Kadın Satranç Kulübü Kaptanı") içeren CV'lerin puanını düşürdü. Amazon, sistemi "tarafsızlaştıramadığı" için projeyi 2018'de iptal etti. Bu vaka, tarihsel verilerdeki önyargıların algoritmalar tarafından nasıl "öğrenildiğini" ve yeniden üretildiğini gösteren klasik bir örnektir.

4.2. Şeffaflık ve Açıklanabilirlik

Şeffaflık ilkesi, YZ sistemlerinin nasıl çalıştığının, hangi verileri kullandığının ve hangi ölçütlere göre karar verdiğinin anlaşılabilir olmasını ifade etmektedir. Açıklanabilirlik ise bu bilgilerin farklı paydaşlar tarafından yorumlanabilir ve denetlenebilir hâle getirilmesini amaçlamaktadır. Özellikle derin öğrenme temelli modellerde karar süreçlerinin opaklığı, kullanıcı güvenini ve kurumsal hesap verebilirliği zayıflatan temel faktörlerden biri olarak öne çıkmaktadır (Burrell, 2016).

Açıklanabilir YZ yaklaşımları, yalnızca teknik bir gereklilik değil; aynı zamanda etik ve hukuki bir zorunluluk olarak değerlendirilmektedir. Kararların gerekçelendirilememesi, hatalı sonuçların düzeltilmesini zorlaştırmakta ve bireylerin itiraz hakkını fiilen sınırlamaktadır (Doshi-Velez ve Kim, 2017). Açıklanabilirliğin etik ve hukuki bir gereklilik olarak öne çıkması, uygulamada belirli teknik gerilimleri de beraberinde getirmektedir.

Doğruluk ve Açıklanabilirlik Arasındaki Teknik Ödünleşim (Trade-off)
Yapay zekâ sistemlerinin geliştirilme süreçlerinde sıklıkla karşılaşılan temel teknik gerilimlerden biri, modelin tahmin başarımı (accuracy) ile karar süreçlerinin insan tarafından anlaşılabilirliği (explainability) arasındaki ters yönlü ilişkidir. Literatürde doğruluk–açıklanabilirlik ödünleşimi (accuracy–interpretability trade-off) olarak tanımlanan bu durum, özellikle derin öğrenme tabanlı modellerde belirgin hâle gelmektedir. Çok katmanlı yapay sinir ağları, büyük ve karmaşık veri kümelerinde yüksek doğruluk düzeylerine ulaşabilmekte; ancak model karmaşıklığı arttıkça, sistemin iç işleyişi kullanıcılar ve denetleyici aktörler açısından giderek daha opak bir hâl almaktadır.
Bu opaklık, YZ sistemlerinin ürettiği kararların gerekçelendirilebilmesini zorlaştırmakta ve özellikle yüksek etkili alanlarda etik ve hukuki sorunlara yol açabilmektedir. Buna karşılık karar ağaçları veya doğrusal modeller gibi daha basit ve açıklanabilir yaklaşımlar, görece daha düşük performans sunsalar da, hataların izlenebilmesi ve sorumluluğun belirlenebilmesi açısından önemli avantajlar sağlamaktadır. Bu durum, geliştiricileri ve politika yapıcıları, doğruluk ile açıklanabilirlik arasında bağlama duyarlı etik tercihler yapmaya zorlamakta; tekil bir teknik optimumdan ziyade, kullanım alanına göre gerekçelendirilebilir denge noktalarının belirlenmesini gerekli kılmaktadır.

4.3. Hesap Verebilirlik ve Sorumluluk

Hesap verebilirlik ilkesi, YZ sistemlerinin ürettiği karar ve sonuçlardan kimin sorumlu olduğunun açık biçimde belirlenmesini gerektirmektedir. YZ sistemlerinin çok aktörlü bir ekosistem içinde geliştirilmesi, sorumluluğun dağılımını karmaşık hâle getirmektedir. Veri sağlayıcılar, geliştiriciler, kurumlar ve kullanıcılar arasındaki bu dağılım, etik, hukuki ve kurumsal belirsizliklere yol açabilmektedir. (Citron ve Pasquale, 2014).

Literatürde, insan denetiminin tamamen ortadan kaldırıldığı sistemlerin etik açıdan sorunlu olduğu yönünde güçlü bir uzlaşı bulunmaktadır. Bu bağlamda “insan denetimli YZ” yaklaşımları, hesap verebilirliği güçlendiren önemli bir etik mekanizma olarak değerlendirilmektedir (Amershi vd., 2019).

4.4. Gizlilik ve Veri Koruma

Gizlilik ilkesi, YZ sistemlerinin bireylere ait kişisel verileri korumasını ve bu verilerin yalnızca meşru, sınırlı ve açık amaçlarla kullanılmasını ifade etmektedir. Büyük veri temelli YZ uygulamaları, bireylerin davranışlarının sürekli izlenmesine ve profillenmesine olanak tanımakta; bu durum mahremiyetin yapısal biçimde aşınmasına yol açabilmektedir (Zuboff, 2019).

Veri minimizasyonu, anonimleştirme ve açık rıza gibi ilkeler, YZ etiğinde gizliliği korumaya yönelik temel araçlar olarak öne çıkmaktadır. Özellikle Avrupa Birliği Genel Veri Koruma Tüzüğü (GDPR), YZ sistemlerinin etik, hukuki ve kurumsal sınırlarını belirlemede önemli bir referans noktası sunmaktadır (Voigt ve von dem Bussche, 2017).

4.6. Güvenlik, Dayanıklılık ve Zararsızlık

Güvenlik ilkesi, YZ sistemlerinin teknik hatalara, saldırılara ve kötüye kullanıma karşı dayanıklı olmasını gerektirmektedir. Hatalı çalışan veya manipüle edilen YZ sistemleri, yalnızca bireysel zararlar değil, geniş çaplı toplumsal riskler de doğurabilmektedir. Özellikle otonom sistemlerde güvenlik ve etik arasındaki ilişki daha da kritik hâle gelmektedir (Amodei vd.., 2016).

Zararsızlık ilkesi, YZ'nin insanlara, topluma ve çevreye zarar vermemesini temel bir etik öncelik olarak ele almaktadır. Bu ilke, teknik güvenlik önlemlerinin ötesinde, olası dolaylı, kümülatif ve uzun vadeli etkilerin de dikkate alınmasını gerektirmektedir.

4.6. İnsan Merkezlilik ve Özerklik

İnsan merkezlilik ilkesi, YZ sistemlerinin insan karar verme süreçlerini tamamen ikame etmesini değil, desteklemesini esas almaktadır. Bu yaklaşım, bireylerin özerkliğini korumayı ve YZ ile etkileşimde bilinçli tercih yapabilmesini amaçlamaktadır. İnsan

merkezli YZ anlayışı, etik ilkelere dayalı sürdürülebilir teknolojik gelişimin temel dayanaklarından biri olarak kabul edilmektedir (EU High-Level Expert Group on AI, 2019).

4.7. Etik İkilemler ve Dengeleme: İlkelerin Çatışması

Yapay zekâ etiğinde belirlenen adalet, şeffaflık, gizlilik ve güvenlik gibi temel ilkeler, teorik düzeyde birbirini tamamlayan idealler olarak görünse de, pratik uygulamalarda bu değerler arasında kaçınılmaz gerilimler ve takaslar (trade-offs) ortaya çıkabilmektedir. Etik tasarım süreci, çoğu zaman bu rakip değerler arasında hassas bir denge kurma mücadelesine dönüşmektedir. Bu bağlamda literatürde öne çıkan en belirgin ikilemlerden biri, gizlilik ile kişiselleştirme arasındaki ters orantılı ilişkidir. YZ sistemlerinin kullanıcılara daha isabetli ve kişiselleştirilmiş hizmetler sunabilmesi, daha yoğun ve ayrıntılı veri toplanmasını gerektirmekte; bu durum ise veri minimizasyonu ve mahremiyet ilkeleriyle çatışabilmektedir.

Benzer bir gerilim, güvenlik ile özgürlük ekseninde de görülmektedir. Kamu güvenliğini artırma iddiasıyla kullanılan gözetim ve risk analizi algoritmaları, kolektif güvenliğe katkı sağlasa da, bireysel özgürlükleri kısıtlama ve gözetim toplumunu normalleştirme riski taşımaktadır. Teknik açıdan ise performans (doğruluk) ile açıklanabilirlik arasında bir takas söz konusudur. Karmaşık derin öğrenme modelleri genellikle daha yüksek doğruluk oranlarına ulaşırken, karar süreçlerinin şeffaflığı azalmakta ve "kara kutu" sorunu derinleşmektedir.

5. Yapay Zekâ Etiğinde Uluslararası İlke ve Düzenleyici Çerçeveler

Yapay zekâ teknolojilerinin sınır aşan etkileri, etik sorunların yalnızca ulusal düzeyde ele alınmasını yetersiz kılmakta; küresel ölçekte ortak ilke ve düzenleyici çerçevelere duyulan ihtiyacı

artırmaktadır. YZ sistemlerinin çok uluslu şirketler tarafından geliştirilmesi, farklı hukuk sistemlerinde eş zamanlı olarak kullanılması ve toplumsal etkilerinin geniş kitleleri kapsaması, etik yönetişimin uluslararası bir mesele hâline gelmesine yol açmıştır. Bu bağlamda son yıllarda birçok uluslararası kuruluş, YZ'nin etik, güvenilir ve insan merkezli biçimde geliştirilmesine yönelik ilke setleri ve politika belgeleri yayımlamıştır (Jobin, Ienca, ve Vayena, 2019).

Uluslararası YZ etik belgeleri, bağlayıcılık düzeyleri açısından farklılık göstermekle birlikte, büyük ölçüde ortak normatif değerler etrafında şekillenmektedir. Bu belgelerde öne çıkan temel amaç, YZ inovasyonunu engellemeden, toplumsal zararları sınırlandıran ve insan haklarını koruyan bir yönetişim modeli oluşturmaktır. Böylece etik ilkeler, soyut normlar olmaktan çıkarak politika üretimi ve düzenleyici çerçeveler için referans noktaları hâline gelmektedir.

5.1. OECD Yapay Zekâ İlkeleri

Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından 2019 yılında yayımlanan Yapay Zekâ İlkeleri, YZ etiği alanında kabul gören ilk kapsamlı uluslararası belge olarak değerlendirilmektedir. Bu ilkeler, daha sonra G20 ülkeleri tarafından da benimsenerek küresel etki alanını genişletmiştir (OECD, 2019). OECD ilkeleri, YZ sistemlerinin insan merkezli, adil, şeffaf, güvenli ve hesap verebilir olması gerektiğini vurgulamaktadır.

OECD çerçevesinin ayırt edici yönlerinden biri, etik ilkeleri yalnızca normatif hedefler olarak değil, politika yapıcılar ve geliştiriciler için uygulanabilir rehberler olarak sunmasıdır. Belgede, YZ'nin ekonomik büyümeyi ve toplumsal refahı desteklemesi gerektiği belirtilirken, bu hedeflerin insan hakları ve demokratik

değerlerle uyumlu biçimde gerçekleştirilmesi gerektiği özellikle vurgulanmaktadır.

5.2. UNESCO Yapay Zekâ Etiği Tavsiyeleri

Birleşmiş Milletler Eğitim, Bilim ve Kültür Örgütü (UNESCO) tarafından 2021 yılında kabul edilen Yapay Zekâ Etiği Tavsiyeleri, etik tartışmaları küresel eşitsizlikler ve sürdürülebilir kalkınma perspektifiyle ele alması bakımından özgün bir konuma sahiptir (UNESCO, 2021). UNESCO belgesi, YZ etiğini yalnızca teknoloji odaklı bir mesele olarak değil, insan onuru, kültürel çeşitlilik ve toplumsal adaletle doğrudan ilişkili bir alan olarak tanımlamaktadır.

Bu çerçevede belgede; insan hakları, çevresel sürdürülebilirlik, toplumsal kapsayıcılık ve barış gibi değerler YZ etiğinin ayrılmaz unsurları olarak ele alınmaktadır. UNESCO yaklaşımı, özellikle gelişmekte olan ülkelerin YZ politikalarına etik bir perspektif kazandırmayı hedeflemekte ve YZ'nin küresel ölçekte yeni eşitsizlikler üretmesini önlemeye yönelik önlemler önermektedir.

5.3. Avrupa Birliği Yapay Zekâ Düzenlemeleri

Avrupa Birliği, YZ etiğini somut düzenleyici araçlarla ele alan en kapsamlı bölgesel aktörlerden biridir. Avrupa Komisyonu tarafından yayımlanan Güvenilir Yapay Zekâ Etik Rehberi ve bunu takip eden Yapay Zekâ Tüzüğü (AI Act) taslağı, etik ilkelerin hukuki yükümlülüklere dönüştürülmesi açısından önemli bir örnek sunmaktadır (European Commission, 2019; European Parliament, 2024).

AB yaklaşımının temelini, risk temelli düzenleme modeli oluşturmaktadır. Bu modelde YZ uygulamaları; kabul edilemez risk,

yüksek risk, sınırlı risk ve minimal risk kategorilerine ayrılmakta; özellikle yüksek riskli uygulamalar için sıkı etik ve hukuki yükümlülükler öngörülmektedir. Bu yaklaşım, etik ilkelerin bağlayıcılığını artırarak YZ geliştiricileri ve kullanıcıları için daha net sorumluluk alanları tanımlamaktadır.

5.4. Uluslararası Belgeler Arasında Ortak İlkeler ve Farklılaşmalar

OECD, UNESCO ve Avrupa Birliği belgeleri incelendiğinde, terminoloji ve vurgu farklılıklarına rağmen belirli ortak ilkelerin öne çıktığı görülmektedir. Adalet, şeffaflık, hesap verebilirlik, gizlilik ve insan merkezlilik, hemen tüm belgelerde tekrar eden temel etik değerlerdir (Jobin vd., 2019). Bununla birlikte, belgelerin bağlayıcılık düzeyi, hedef kitlesi ve uygulama mekanizmaları arasında önemli farklılıklar bulunmaktadır.

OECD ve UNESCO belgeleri daha çok rehber niteliği taşırken, Avrupa Birliği düzenlemeleri etik ilkeleri hukuki yaptırımlarla destekleyen bir model sunmaktadır. Bu durum, YZ etiğinin küresel ölçekte tek tip bir düzenleme yerine, çok katmanlı ve bağlama duyarlı bir yönetişim anlayışıyla ele alındığını göstermektedir.

Uluslararası ilke ve düzenleyici çerçeveler, YZ etiğinin teorik tartışmaların ötesine geçerek politika ve uygulama alanına taşındığını göstermektedir (Tablo 3). Bu belgeler, YZ teknolojilerinin insan değerleriyle uyumlu biçimde geliştirilmesi için küresel bir etik zemin oluşturmakta ve ulusal politikalar için referans niteliği taşımaktadır.

Tablo 3. OECD, UNESCO ve AB yaklaşımlarının yapay zekâ etiği ve yönetişimi açısından karşılaştırılması

Boyut	OECD Yaklaşımı	UNESCO Yaklaşımı	AB Yaklaşımı (AI Act)
Yaklaşım türü	İlke temelli politika çerçevesi	Değer ve insan hakları temelli etik çerçeve	Risk temelli bağlayıcı düzenleme
Temel amaç	Güvenilir yapay zekâyı teşvik ederken inovasyonu desteklemek	İnsan onuru, insan hakları ve toplumsal yararı korumak	Temel haklarla uyumlu ve güvenli yapay zekâ kullanımını sağlamak
Normatif dayanak	Demokratik değerler, insan hakları, şeffaflık, hesap verebilirlik	İnsan hakları, sürdürülebilirlik, kapsayıcılık	Risk sınıflandırması ve hukuki yükümlülükler
Uygulama mantığı	Esnek ilke rehberliği ve politika uyumu	Değerlerin yapay zekâ yaşam döngüsüne entegrasyonu	Risk düzeyine göre kademeli yükümlülükler
Şeffaflık ve açıklanabilirlik	İlke düzeyinde teşvik edilir	İnsan haklarının korunması için temel gereklilik	Risk düzeyine bağlı zorunlu yükümlülük
Hesap verebilirlik	Kurumsal sorumluluk ve iyi yönetişim vurgusu	İnsan denetimi ve etik sorumluluk	Hukuki sorumluluk ve yaptırım mekanizmaları
Bağlayıcılık düzeyi	Gönüllü / bağlayıcı olmayan	Gönüllü / küresel etik rehber	Yasal olarak bağlayıcı
Etik–yönetişim ilişkisi	Etik ilkeler üzerinden yönetişim	Değer temelli etik yönetişim	Etik risklerin hukuki düzenleme ile yönetimi

Not. *Tablo, yapay zekâ yönetişimine ilişkin ilke temelli, değer temelli ve risk temelli yaklaşımların temel özelliklerini karşılaştırmalı olarak özetlemektedir.*

Tablo 3'te görüldüğü üzere, OECD ve UNESCO yaklaşımları değer ve ilke temelli yönetişim çerçeveleri sunarken, AB Yapay Zekâ Tüzüğü bu çerçeveleri risk temelli ve bağlayıcı bir

düzenleme modeliyle tamamlamaktadır. Bu farklılıklar, yapay zekâ etiğinin yalnızca normatif ilkelerle değil, aynı zamanda kurumsal ve hukuki mekanizmalarla da desteklenmesi gerektiğini ortaya koymaktadır.

6. Yapay Zekâ Etiğinde Güncel Tartışmalar

Üretici yapay zekâ sistemlerinin yaygınlaşması, YZ'nin yalnızca karar destek aracı değil; içerik üreten, önerilerde bulunan ve insan benzeri etkileşimler kurabilen bir yapı hâline gelmesini sağlamıştır. Bu dönüşüm, mevcut etik ilkelerin yeterliliğini sorgulatan ve yeni normatif soruları gündeme getiren güncel tartışma alanlarını ortaya çıkarmıştır.

6.1. Üretici YZ ve İçerik Üretiminin Etik Sonuçları

Üretici yapay zekâ (generative AI) sistemleri, metin, görsel, ses ve video gibi içerikleri insan girdisine dayalı olarak otomatik biçimde üretebilmektedir. Bu sistemler, yaratıcı süreçlere erişimi genişletme ve üretkenliği artırma potansiyeline sahip olmakla birlikte, özgünlük, doğruluk ve sorumluluk gibi etik sorunları da beraberinde getirmektedir (Floridi ve Chiriatti, 2020).

Üretici YZ tarafından oluşturulan içeriklerin doğruluğunun denetlenmesi güçleşmekte, yanlış veya yanıltıcı bilgilerin hızlı biçimde yayılması riski artmaktadır. Bu durum, özellikle eğitim, bilimsel iletişim ve kamusal tartışmalar açısından bilgi güvenilirliğini tehdit eden etik ve epistemik bir sorun olarak değerlendirilmektedir (Bender vd.., 2021).

6.2. Telif Hakkı, Veri Sahipliği ve Emek Sorunu

YZ sistemlerinin büyük ölçekli veri kümeleri üzerinde eğitilmesi, telif hakkı ve veri sahipliği tartışmalarını güncel etik tartışmaların odağına yerleştirmiştir. Özellikle üretici YZ modellerinin, telifli eserlerden öğrenerek yeni içerikler üretmesi, bu

içeriklerin mülkiyetinin kime ait olduğu sorusunu gündeme getirmektedir (Samuelson, 2023).

Bu bağlamda etik tartışmalar, yalnızca hukuki çerçevelerle sınırlı kalmamakta; yaratıcı emeğin değeri, görünmez emek ve adil paylaşım gibi normatif soruları da içermektedir. YZ'nin üretim süreçlerine entegre edilmesi, yaratıcı mesleklerin dönüşümünü hızlandırırken, bu durum, etik yönetişim mekanizmalarının yokluğunda yaratıcı emeğin değersizleşmesi ve adaletsiz sonuçlar üretmesi riskini artırmaktadır.

6.3. Deepfake Teknolojileri ve Gerçeklik Krizi

Deepfake teknolojileri, YZ etiğinde yüksek toplumsal risk potansiyeline sahip güncel uygulamalardan biri olarak değerlendirilmektedir. Gerçekçi sahte video ve ses içeriklerinin üretilebilmesi, bireylerin itibarını zedeleyebilmekte, politik manipülasyonlara zemin hazırlamakta ve toplumsal güveni aşındırmaktadır (Chesney ve Citron, 2019).

Deepfake içeriklerin tespiti teknik olarak mümkün olsa da, bu içeriklerin hızla yayılması ve çoğu zaman geri dönüşü olmayan etkiler yaratması etik müdahalenin yalnızca teknik tespit mekanizmalarına indirgenemeyeceğini göstermektedir. Bu bağlamda etik tartışmalar, sorumluluk paylaşımı, platform yükümlülükleri ve kamusal farkındalık boyutlarını da kapsamaktadır.

6.4. İnsan Benzeri Yapay Zekâ ve Ahlaki Statü Tartışmaları

Gelişmiş YZ sistemlerinin insan benzeri dil kullanımı, duygusal tepkiler üretmesi ve sosyal etkileşim kurabilmesi, bu sistemlerin ahlaki statüsüne ilişkin tartışmaları gündeme getirmiştir. Bu tartışmalar, YZ'nin haklara sahip bir varlık olarak değerlendirilip değerlendirilemeyeceği sorusundan ziyade, insanların YZ'ye nasıl

davrandığının etik sonuçlarına odaklanmakt etik ve psikososyal sonuçlarına odaklanmaktadır. (Coeckelbergh, 2020).

İnsanların YZ sistemlerine aşırı güven geliştirmesi veya onları bilinçli varlıklar olarak algılaması, etik açıdan manipülasyon riskini artırmaktadır. Bu durum, özellikle savunmasız gruplar açısından YZ ile kurulan etkileşimlerin dikkatle tasarlanmasını gerekli kılmaktadır.

6.5. Otonomi, Kontrol ve Uzun Vadeli Riskler

YZ sistemlerinin artan özerkliği, insan kontrolünün sınırlarına ilişkin etik tartışmaları da beraberinde getirmektedir. Kısa vadeli etik sorunların yanı sıra, uzun vadede YZ'nin toplumsal yapılar üzerindeki dönüştürücü etkileri, etik değerlendirmelerin yönetişim ve kontrol perspektifini uzun vadeye taşımaktadır (Bostrom, 2014).

Bu bağlamda güncel etik literatür, YZ'nin kontrol edilebilirliğini, değer uyumunu (value alignment) ve insan refahıyla uyumlu gelişimini güvence altına alacak yönetişim mekanizmalarının önemini vurgulamaktadır. Uzun vadeli riskler, çoğu zaman spekülatif bulunmakla birlikte, etik tartışmaların yalnızca mevcut sorunlarla sınırlı kalmaması gerektiğini göstermektedir.

Üretici YZ, telif hakkı, deepfake teknolojileri ve insan benzeri YZ uygulamaları, etik ilkelerin yeniden yorumlanmasını ve bağlama duyarlı biçimde uygulanmasını zorunlu kılmaktadır. Bu tartışmalar, YZ etiğinin sabit bir normlar bütünü olmaktan ziyade, dinamik ve sürekli güncellenen bir düşünsel çerçeve olduğunu ortaya koymaktadır.

7. Sorumlu ve Etik Yapay Zekâ Geliştirme Yaklaşımları

Yapay zekâ etiği alanında yürütülen tartışmaların olgunlaşmasıyla birlikte, etik ilkelerin soyut normlar olmaktan çıkarılarak somut geliştirme ve yönetişim süreçlerine nasıl entegre edileceği sorusu, bu alandaki temel tartışma konularından biri hâline gelmiştir. Bu bağlamda literatürde sorumlu yapay zekâ (responsible AI) kavramı, YZ sistemlerinin toplumsal yarar, insan hakları ve etik değerlerle uyumlu biçimde tasarlanmasını ve uygulanmasını amaçlayan bütüncül bir yaklaşım olarak öne çıkmaktadır (Dignum, 2019).

Sorumlu YZ yaklaşımları, etik sorumluluğu yalnızca son kullanıcıya yüklemek yerine, YZ'nin tüm yaşam döngüsünü kapsayan çok aktörlü bir sorumluluk anlayışına dayanmaktadır. Veri toplama, model geliştirme, test, dağıtım ve kullanım aşamalarının her biri, etik değerlendirmelerin yapılmasını gerektiren kritik süreçler olarak ele alınmaktadır.

7.1. Etik Tasarım Yaklaşımı

Etik tasarım yaklaşımı (Ethics by Design), etik ilkelerin YZ sistemlerine sonradan dayatılan dışsal kısıtlar olarak değil, tasarım sürecinin ayrılmaz bir parçası olarak ele alınmasını savunmaktadır. Bu yaklaşım, etik risklerin sistem geliştirme sürecinin erken aşamalarında öngörülmesini ve azaltılmasını hedeflemektedir (van den Hoven vd.., 2015).

Etik tasarım, veri seçimi ve temsilinden model mimarisine, kullanıcı arayüzlerinden geri bildirim mekanizmalarına kadar geniş bir alanı kapsamaktadır (Tablo 4). Örneğin, adalet ilkesinin gözetilmesi, yalnızca model çıktılarının değerlendirilmesiyle sınırlı kalmamakta; verilerin hangi toplumsal grupları ne ölçüde temsil ettiğinin de etik bir sorun olarak ele alınmasını gerektirmektedir. Bu

yaklaşım, etik ile teknik kararlar arasındaki yapay ayrımı ortadan kaldırarak, mühendislik pratiğini normatif bir zemine oturtmaktadır.

Tablo 4. Etik tasarım ilkeleri ve yapay zekâ geliştirme sürecindeki karşılıkları

Etik Tasarım İlkesi	İlkenin Kısa Açıklaması	YZ Geliştirme Sürecindeki Karşılığı
İnsan merkezlilik	YZ sistemlerinin insan özerkliğini ve karar yetisini desteklemesi	İnsan denetimli karar mekanizmalarının tasarlanması, kullanıcı kontrol noktalarının eklenmesi
Adalet ve ayrımcılıktan kaçınma	Farklı toplumsal gruplar için eşit ve hakkaniyetli sonuçlar üretme	Veri setlerinde temsil analizi, önyargı testleri, adalet metriklerinin kullanımı
Şeffaflık ve açıklanabilirlik	Karar süreçlerinin anlaşılabilir ve izlenebilir olması	Açıklanabilir model seçimi, model çıktılarının gerekçelendirilmesi
Gizlilik ve veri koruma	Kişisel verilerin korunması ve amaçla sınırlı kullanımı	Veri minimizasyonu, anonimleştirme, gizlilik dostu modelleme
Hesap verebilirlik	Kararların sorumluluğunun belirlenebilir olması	Kayıt tutma, denetlenebilirlik mekanizmaları, sorumluluk ataması
Güvenlik ve sağlamlık	Sistemlerin hatalara ve kötüye kullanıma karşı dayanıklı olması	Test, doğrulama, stres senaryoları ve güvenlik önlemleri
Toplumsal yarar	YZ'nin bireysel ve toplumsal refahı desteklemesi	Kullanım amacının etik değerlendirmesi, etki analizleri

7.2. Etki ve Risk Değerlendirme Mekanizmaları

Sorumlu YZ geliştirme süreçlerinde öne çıkan bir diğer önemli araç, etik etki ve risk değerlendirmeleridir. Bu değerlendirmeler, YZ sistemlerinin potansiyel zararlarını önceden belirlemeyi ve bu zararları en aza indirecek önlemleri planlamayı amaçlamaktadır. Özellikle yüksek riskli YZ uygulamalarında, sistemlerin toplumsal, hukuki ve etik etkilerinin sistematik biçimde analiz edilmesi önerilmektedir (Floridi vd., 2018).

Avrupa Birliği tarafından önerilen risk temelli yaklaşım, bu değerlendirmelerin kurumsal düzeyde standartlaştırılmasına yönelik önemli bir örnek sunmaktadır. Risk değerlendirmeleri, YZ sistemlerinin bağlama özgü etkilerini dikkate alarak tek tip çözümler yerine esnek ve uyarlanabilir etik mekanizmaların geliştirilmesine imkân tanımaktadır.

7.3. İnsan Denetimi ve Hibrit Karar Modelleri

Sorumlu YZ yaklaşımlarında sıklıkla vurgulanan bir diğer unsur, insan denetiminin korunmasıdır. İnsan–makine etkileşimine dayalı hibrit karar modelleri, YZ'nin hesaplama gücünü insan yargısı ve etik muhakeme ile birleştirmeyi amaçlamaktadır. Bu modeller, özellikle sağlık, eğitim ve kamu yönetimi gibi yüksek etkili alanlarda etik sorumluluğun insan aktörlerde kalmasını güvence altına almaktadır (Amershi vd.., 2019).

İnsan denetimi, YZ sistemlerinin tamamen özerk hâle gelmesini engelleyen bir etik mekanizma olmasının yanı sıra, hataların erken tespit edilmesine ve kullanıcı güveninin artırılmasına da katkı sağlamaktadır. Bu yaklaşım, YZ'nin insan kararlarını ikame eden değil, destekleyen bir teknoloji olarak konumlandırılmasını mümkün kılmaktadır.

Tablo 5. Tam otomatik, insan denetimli ve hibrit yapay zekâ karar modellerinin etik açıdan karşılaştırılması

Karar Modeli	Temel Özellik	Avantajlar	Başlıca Etik Riskler	Uygun Kullanım Alanları
Tam otomatik YZ karar modeli	Kararlar tamamen YZ sistemi tarafından alınır; insan müdahalesi yoktur	Hız, ölçeklenebilirlik, operasyonel verimlilik	Hesap verebilirlik belirsizliği, şeffaflık eksikliği, hataların sistemik etkisi	Düşük riskli, geri döndürülebilir kararlar
İnsan denetimli YZ karar modeli (human-in-the-loop)	YZ öneri sunar, nihai karar insan tarafından verilir	Etik denetim, hata tespiti, güven artışı	Karar süreçlerinde yavaşlama, insan önyargısının devamı	Sağlık, eğitim, kamu yönetimi gibi yüksek etkili alanlar
Hibrit YZ karar modeli	YZ ve insan karar süreçleri bağlama göre paylaşılır	Esneklik, bağlamsal uyum, dengeleyici sorumluluk	Rol belirsizliği, sorumluluk paylaşımının karmaşıklığı	Orta–yüksek riskli, bağlama duyarlı uygulamalar

7.4. Kurumsal Yönetişim ve Etik Kurullar

Etik ilkelerin sürdürülebilir biçimde uygulanabilmesi, yalnızca bireysel geliştiricilerin etik duyarlılığına bırakılmamalı; kurumsal yönetişim mekanizmalarıyla desteklenmelidir. Bu bağlamda birçok kurum, YZ projelerini değerlendirmek üzere etik kurullar ve danışma komiteleri oluşturmaktadır. Bu kurullar, YZ uygulamalarının etik risklerini değerlendirmekte ve kurumsal politika geliştirme süreçlerine katkı sunmaktadır (Morley vd.., 2020).

Kurumsal etik yönetişim, şeffaflık, hesap verebilirlik ve izlenebilirlik ilkelerinin somutlaşmasını sağlamaktadır. Aynı zamanda kurum içi farkındalığın artmasına ve etik kültürün yaygınlaşmasına katkıda bulunmaktadır.

7.5. Eğitim, Farkındalık ve Disiplinlerarası İş Birliği

Sorumlu YZ geliştirme yaklaşımlarının başarısı, teknik önlemlerin yanı sıra insan faktörüne de bağlıdır. Geliştiricilerin, yöneticilerin ve kullanıcıların etik farkındalık ve sorumluluk bilincinin artırılmas, YZ etiğinin uygulamaya yansımasında kritik bir rol oynamaktadır. Bu nedenle etik eğitimin, bilgisayar bilimi ve mühendislik programlarının ayrılmaz bir parçası hâline getirilmesi önerilmektedir (Fiesler vd.., 2020).

Ayrıca YZ etiği, tek bir disiplinin çözebileceği bir alan değildir. Felsefe, hukuk, sosyal bilimler ve alan uzmanlarının iş birliği, etik risklerin daha bütüncül biçimde değerlendirilmesini mümkün kılmaktadır. Disiplinlerarası yaklaşım, YZ'nin toplumsal etkilerinin daha dengeli ve kapsayıcı biçimde ele alınmasına katkı sağlamaktadır.

Sorumlu ve etik YZ geliştirme yaklaşımları, YZ'nin yalnızca teknik olarak başarılı değil, aynı zamanda toplumsal olarak meşru ve güvenilir bir teknoloji hâline gelmesini amaçlamaktadır. Etik tasarım, risk değerlendirmeleri, insan denetimi ve kurumsal yönetişim mekanizmaları, bu amaca ulaşmada birbirini tamamlayan temel bileşenler olarak öne çıkmaktadır.

8. Gelecek Perspektifi ve Politika Önerileri

Yapay zekâ teknolojilerinin gelişim hızı, etik ve düzenleyici yaklaşımların çoğu zaman bu gelişimi geriden takip etmesine neden olmaktadır. Bu durum, YZ etiğinin yalnızca mevcut risklere odaklanan reaktif bir alan olarak değil, gelecekte ortaya çıkabilecek

toplumsal etkileri öngören proaktif bir politika alanı olarak ele alınmasını gerekli kılmaktadır. Gelecek perspektifi, etik ilkelerin sürdürülebilir biçimde uygulanabilmesi ve YZ'nin insan refahını artıracak şekilde yönlendirilmesi açısından belirleyici bir rol oynamaktadır.

8.1. Yapay Zekâ Etiğinin Gelecekteki Yönelimi

Gelecekte YZ etiğinin, statik ilke setlerinden ziyade uyarlanabilir ve bağlama duyarlı etik çerçeveler üzerinden gelişmesi beklenmektedir. YZ sistemlerinin farklı sektörlerde ve farklı toplumsal bağlamlarda kullanılması, tek tip etik çözümlerin yetersiz kalmasına yol açmaktadır. Bu nedenle literatürde, dinamik etik yönetişim modellerinin ve sürekli izleme mekanizmalarının önemi giderek daha fazla vurgulanmaktadır (Floridi, 2021).

Ayrıca üretici YZ ve otonom sistemlerin yaygınlaşması, etik değerlendirmelerin yalnızca sistem tasarımına değil, kullanım pratiklerine ve ortaya çıkan beklenmedik sonuçlara da odaklanmasını gerektirmektedir. Bu bağlamda YZ etiğinin geleceği, teknik gelişmelerle birlikte sürekli güncellenen, öğrenen ve geri bildirim temelli bir alan olarak şekillenmektedir.

8.2. Politika Yapıcılar İçin Öneriler

Yapay zekâ etiğinin gelecekte etkili olabilmesi, güçlü ve uygulanabilir politika çerçeveleriyle desteklenmesine bağlıdır. Politika yapıcılar için öneriler, genellikle üç ana eksende toplanmaktadır: düzenleme, yönetişim ve katılım. İlk olarak, risk temelli düzenleme yaklaşımlarının benimsenmesi, yüksek etkili YZ uygulamalarının daha sıkı etik ve hukuki denetime tabi tutulmasını mümkün kılmaktadır (European Commission, 2024).

İkinci olarak, çok paydaşlı yönetişim modelleri, YZ politikalarının yalnızca teknik uzmanlar tarafından değil; sivil

toplum, akademi ve kullanıcı temsilcilerinin katılımıyla şekillendirilmesini sağlamaktadır. Bu yaklaşım, etik kararların demokratik meşruiyetini güçlendirmektedir (Jobin vd., 2019).

Üçüncü olarak, etik etki değerlendirmelerinin ve şeffaf raporlama yükümlülüklerinin kurumsal düzeyde teşvik edilmesi, YZ uygulamalarının izlenebilirliğini ve hesap verebilirliğini artırmaktadır.

8.3. Kurumsal ve Sektörel Düzeyde Stratejiler

Gelecekte YZ etiğinin başarısı, yalnızca ulusal veya uluslararası politikalarla değil; kurumsal ve sektörel stratejilerle de doğrudan ilişkilidir. Kurumların etik ilkelere dayalı iç politikalar geliştirmesi, etik riskleri yönetmede önemli bir avantaj sağlamaktadır. Bu bağlamda etik denetim mekanizmaları, iç raporlama sistemleri ve sürekli eğitim programları, kurumsal düzeyde uygulanabilecek temel stratejiler arasında yer almaktadır (Morley vd.., 2020).

Sektörel düzeyde ise ortak etik standartların ve iyi uygulama örneklerinin paylaşılması, YZ'nin farklı alanlarda daha dengeli ve sorumlu biçimde kullanılmasına katkı sağlamaktadır. Bu tür kolektif yaklaşımlar, rekabet baskısının etik ilkeleri zayıflatmasını önlemeye yardımcı olacaktır.

8.4. Uzun Vadeli Toplumsal Etkiler ve Etik Sorumluluk

Yapay zekânın uzun vadeli toplumsal etkileri, etik tartışmaların zaman ufkunu genişletmektedir. İş gücü dönüşümü, karar alma süreçlerinin otomasyonu ve insan–makine ilişkilerinin yeniden tanımlanması, YZ'nin yalnızca bugünü değil, geleceğin toplumsal yapısını da şekillendirdiğini göstermektedir (Bostrom, 2014).

Bu durumda etik sorumluluk, yalnızca mevcut kullanıcıların değil, gelecek kuşakların da haklarını gözeten bir perspektifi gerektirmektedir. YZ'nin sürdürülebilir ve insan merkezli biçimde geliştirilmesi, etik ilkelerin uzun vadeli toplumsal refah hedefleriyle uyumlu hâle getirilmesini zorunlu kılmaktadır.

Gelecek perspektifi ve politika önerileri, YZ etiğinin yalnızca mevcut sorunlara yanıt veren bir alan değil; teknolojik gelişimi yönlendiren stratejik bir çerçeve olduğunu ortaya koymaktadır. Eğitim, yönetişim ve katılımcı politika üretimi, YZ'nin etik açıdan sürdürülebilir bir geleceğe taşınmasında temel araçlar olarak öne çıkmaktadır.

9. Sonuç

Yapay zekâ teknolojilerinin hızla yaygınlaşması, etik tartışmaları teknik bir yan mesele olmaktan çıkararak, toplumsal, hukuki ve politik boyutları olan merkezi bir tartışma alanına taşımıştır. YZ sistemleri, karar alma süreçlerinde giderek daha fazla rol üstlendikçe; adalet, şeffaflık, hesap verebilirlik, gizlilik ve insan merkezlilik gibi etik ilkeler, teknolojik gelişimin ayrılmaz bileşenleri hâline gelmiştir. Bu bağlamda YZ etiği, yalnızca riskleri tanımlayan eleştirel bir alan değil, aynı zamanda teknolojik yenilikleri yönlendiren normatif bir çerçeve sunmaktadır.

Bu çalışmada ele alınan kuramsal yaklaşımlar, uygulama alanları ve uluslararası düzenleyici çerçeveler, YZ etiğinin çok katmanlı ve disiplinlerarası yapısını ortaya koymaktadır. Etik sorunların, YZ'nin teknik özelliklerinden bağımsız olmadığı; aksine veri temelli öğrenme, otomatik karar verme ve ölçeklenebilirlik gibi temel özelliklerle doğrudan ilişkili olduğu görülmektedir. Bu durum, etik değerlendirmelerin yalnızca kullanım aşamasında değil, tasarım ve geliştirme süreçlerinin tamamında yapılmasını kaçınılmaz hâle getirmektedir.

Uluslararası ilke ve düzenlemeler incelendiğinde, YZ etiğine yönelik küresel bir farkındalığın oluştuğu açıkça görülmektedir. OECD, UNESCO ve Avrupa Birliği gibi aktörler tarafından geliştirilen çerçeveler, ortak etik değerler etrafında önemli bir yakınsama sergilemekte; ancak uygulama ve bağlayıcılık düzeylerinde farklılaşmaktadır. Bu tablo, YZ etiğinin tek tip bir düzenleme ile ele alınamayacağını, bağlama duyarlı ve çok düzeyli yönetişim modellerine ihtiyaç duyulduğunu göstermektedir.

Uygulama alanlarına göre yapılan değerlendirmeler, etik risklerin soyut değil; sağlık, eğitim ve kamu yönetimi gibi alanlarda doğrudan insan yaşamını etkileyen sonuçlar ürettiğini ortaya koymaktadır. Bu alanlarda ortaya çıkan etik sorunlar, YZ'nin toplumsal kabulü ve meşruiyeti açısından belirleyici bir rol oynamaktadır. Etik ilkelerin göz ardı edildiği YZ uygulamaları, kısa vadeli verimlilik kazanımları sağlasa dahi, uzun vadede toplumsal güvenin zedelenmesine ve teknolojik ilerlemenin sorgulanmasına yol açabilmektedir.

Sorumlu ve etik YZ geliştirme yaklaşımları, bu risklere karşı geliştirilen en önemli yanıtlar arasında yer almaktadır. Etik tasarım, risk ve etki değerlendirmeleri, insan denetimi ve kurumsal yönetişim mekanizmaları, YZ'nin toplumsal yararla uyumlu biçimde geliştirilmesini mümkün kılmaktadır. Bu yaklaşımlar, etik ile inovasyon arasında zorunlu bir karşıtlık olmadığı; aksine etik ilkelerin sürdürülebilir yeniliğin temel koşullarından biri olduğu yönündeki görüşü desteklemektedir.

Gelecek perspektifi açısından bakıldığında, YZ etiğinin statik bir normlar bütünü olarak değil, sürekli güncellenen ve öğrenen bir çerçeve olarak ele alınması gerektiği anlaşılmaktadır. Üretici YZ, otonom sistemler ve insan–makine etkileşimindeki dönüşümler, etik tartışmaların kapsamını genişletmeye devam edecektir.

Yapay zekânın geleceği, yalnızca ne kadar akıllı sistemler geliştirilebildiğiyle değil, bu sistemlerin hangi değerler doğrultusunda ve kimin yararına kullanıldığıyla belirlenecektir. Bu nedenle etik, yapay zekâ tartışmalarının merkezinde yer almaya devam edecektir.

Kaynakça

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint*. https://arxiv.org/abs/1606.06565

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. https://fairmlbook.org

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

Coeckelbergh, M. (2020). AI ethics. *MIT Press*.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer. https://doi.org/10.1007/978-3-030-30371-6

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*. https://arxiv.org/abs/1702.08608

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Brussels.

EU High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

European Commission. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament. (2024). *Artificial Intelligence Act*. https://www.europarl.europa.eu

European Parliament & Council of the European Union. (2024).
Regulation (EU) 2024/... laying down harmonised rules on artificial

intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Fiesler, C., Garrett, N., & Beard, N. (2020). What do we teach when we teach tech ethics? *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–22. https://doi.org/10.1145/3392878

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L. (2021). The ethics of artificial intelligence for the sustainable development goals. *Philosophy & Technology, 34*, 1175–1187. https://doi.org/10.1007/s13347-021-00428-1

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, D. G. (2015). *Technology and society: Social networks, power, and inequality*. MIT Press.

Long, D., & Magerko, B. (2020). What is AI literacy? *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Mershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Mitcham, C. (2014). *Ethics and technology: Controversies, questions, and strategies for ethical computing* (2nd ed.). Cengage Learning.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: Translating AI ethics principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

OECD. (2019). *OECD principles on artificial intelligence*. https://www.oecd.org/ai/principles/

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Samuelson, P. (2023). Generative AI and copyright. *Communications of the ACM, 66*(8), 20–23. https://doi.org/10.1145/3583070

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). *Handbook of ethics, values, and technological design*. Springer. https://doi.org/10.1007/978-94-007-6994-6

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.

Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.

YEŞİL YAPAY ZEKÂ VE SÜRDÜRÜLEBİLİRLİK

ONUR SEVLİ[2]

1. Giriş

Sürdürülebilirlik, günümüzün en kritik küresel sorunlarından biridir ve ekonomik, çevresel, toplumsal boyutlarıyla çok disiplinli bir yaklaşım gerektirmektedir. Hızla artan dijitalleşme ve büyüyen bilgi işlem altyapıları, enerji tüketiminin ve karbon salınımının önemli ölçüde artmasına neden olmaktadır. Yapay zekâ (YZ) sistemleri ise bu büyümenin merkezinde yer almaktadır. Giderek artan veri miktarı, büyük ölçekli yapay zekâ modelleri ve yüksek hesaplama gücü ihtiyacı, YZ'nin çevresel etkilerini bilimsel tartışmaların odağına taşımaktadır (Strubell et al., 2019).

Son zamanlarda "Yeşil Yapay Zekâ (Green AI)" olarak adlandırılan yeni bir yaklaşım literatürde yerini almıştır. Bu yaklaşım, YZ modellerinin yalnızca performans ölçütlerine göre değil; enerji verimliliği, karbon ayak izi, hesaplama maliyeti ve sürdürülebilirlik katkısı açısından da değerlendirilmesi gerektiğini savunmaktadır (Schwartz et al., 2020). Yeşil YZ, çevresel riskleri azaltılmış modellerin geliştirilmesini hedeflerken, aynı zamanda

[2] Doç.Dr., Burdur Mehmet Akif Ersoy Üniversitesi, Bilgisayar Mühendisliği, Orcid: 0000-0002-8933-8395

YZ'nin sürdürülebilir kalkınma alanlarında kullanılmasını da teşvik etmektedir.

Günümüzde YZ'nin çevresel etkileri yalnızca algoritmik tasarım seviyesinde değil; veri merkezlerinin enerji tüketimi, donanım üretiminin doğal kaynaklara etkisi, model eğitiminin karbon maliyeti ve dijital eşitsizlik gibi geniş bir perspektiften ele alınmaktadır (Henderson et al., 2020). Bu nedenle Yeşil YZ hem teknik bir araştırma alanı hem de etik, politik ve ekonomik sonuçları olan çok boyutlu bir kavram hâline gelmiştir.

2. Yeşil Yapay Zekâ: Kavramsal Çerçeve

Yeşil Yapay Zekâ kavramı, yapay zekâ sistemlerinin çevresel etkilerini en aza indirmeyi amaçlayan bütüncül bir yaklaşımı ifade eder. Geleneksel yapay zekâ araştırmalarında performans odaklı değerlendirme ölçütleri (doğruluk, kesinlik, duyarlılık, F1 skor vb.) ön planda yer alırken, Yeşil YZ bu ölçütlere ek olarak enerji tüketimi, hesaplama maliyeti, karbon emisyonu ve donanım verimliliği gibi sürdürülebilirlik parametrelerini de değerlendirme sürecine dahil etmektedir. Bu yönüyle Yeşil YZ hem metodolojik hem de etik bir dönüşümün temelini oluşturmaktadır.

2.1. Kırmızı Yapay Zekâ Paradigması

Yapay zekâ araştırmalarının son on yılına, "State-of-the-Art" (SOTA) (ulaşılabilen en iyi) sonuçlara ulaşma hedefi yön vermiştir. Schwartz ve arkadaşları (2020) tarafından literatüre kazandırılan Kırmızı YZ kavramı, model performansında marjinal iyileştirmeler elde etmek amacıyla hesaplama gücünün ve veri setlerinin devasa boyutlara ulaştırılmasını ifade etmektedir. Bu yaklaşım, özellikle Doğal Dil İşleme alanında parametre sayılarının milyarlardan trilyonlara çıkmasıyla (örneğin GPT-4, Gemini Ultra) somutlaşmıştır (Strubell et al., 2019).

Kırmızı YZ'nin temel karakteristiği, enerji verimliliğinin ikincil plana atılmasıdır. Model doğruluğunda lineer bir artış sağlamak için gereken hesaplama gücü genellikle üstel olarak artış göstermektedir. Örneğin, bir modelin hata oranını %1 azaltmak, enerji tüketimini 10 katına çıkarabilmektedir. Bu yaklaşım iki temel soruna yol açmaktadır (Schwartz et al., 2020):

Çevresel maliyet: Hesaplama yoğunluğu, veri merkezlerinde tüketilen elektrik ve soğutma için harcanan su miktarını artırarak doğrudan karbon ayak izini büyütmektedir.

Erişilebilirlik ve Eşitsizlik: Yüksek maliyetli donanım gereksinimleri, YZ araştırmalarını yalnızca büyük teknoloji şirketlerinin tekelinde tutmakta, akademik kurumların ve bağımsız araştırmacıların alana katılımını büyük ölçüde engellemektedir.

2.2. Enerji Verimliliği Odaklı Makine Öğrenimi

Yeşil YZ, enerji verimliliğini makine öğrenimi tasarımının merkezine yerleştirir. Bu bağlamda araştırmalar, algoritmaların enerji tüketimini azaltmak için model mimarilerinin sadeleştirilmesi, hesaplama adımlarının optimize edilmesi ve daha az parametreyle benzer performans elde edilmesi üzerine yoğunlaşmaktadır (Patterson et al., 2021). Enerji verimliliği kavramı yalnızca eğitim sürecini değil; modelin çıkarım aşamasındaki performansını da kapsamaktadır. Özellikle gerçek zamanlı uygulamalarda çıkarım maliyetleri sürdürülebilirlik açısından kritik öneme sahiptir.

2.3. Düşük Karbon Ayak İzi Hedefleri

Makine öğrenimi modellerinin karbon ayak izi, eğitim sürecindeki enerji tüketimi ile bu enerji kaynağının karbon yoğunluğuna bağlı olarak hesaplanmaktadır (Lacoste et al., 2019).

Yeşil YZ, karbon salınımını azaltmak için iki temel strateji önermektedir:

- Enerji tüketimini azaltan modeller kullanmak.
- Karbon yoğunluğu düşük enerji kaynaklarıyla çalışan veri merkezleri tercih etmek.

Son yıllarda Google, Microsoft ve Amazon gibi teknoloji şirketleri veri merkezlerini yenilenebilir enerji kaynaklarıyla çalıştırma konusunda önemli adımlar atmıştır. Akademik çalışmalar da karbon muhasebesi ve yapay zekâ enerji raporlamasının standartlaşması gerekliliğini vurgulamaktadır (Henderson et al., 2020).

2.4. Hesaplama Maliyeti, Model Karmaşıklığı ve Çevresel Etki

YZ modellerinin çevresel maliyetleri, büyük ölçüde parametre sayısı, kullanılan donanım türü ve eğitim süresine bağlıdır. Kaplan, McCandlish ve arkadaşlarının (2020) çalışmaları, model boyutunun performansla doğrusal olmayan bir ilişkiye sahip olduğunu göstererek gereğinden büyük modellerin verimsizliğine dikkat çekmiştir. Bu bulgu, Yeşil YZ yaklaşımının temel dayanaklarından biridir. Daha büyük modeller her zaman daha yararlı değildir; aksine çevresel maliyetler çoğu zaman faydanın çok üzerindedir. Bu bağlamda Yeşil YZ; hesaplama maliyetinin, enerji yoğunluğunun ve donanım kullanımının performansla birlikte değerlendirilmesini önermektedir.

2.5. Yeşil Yapay Zekâ Vizyonu

Yeşil YZ, yapay zekâ teknolojilerinin geliştirme süreçlerinde verimliliği, en az doğruluk kadar kritik bir değerlendirme metriği olarak konumlandıran bütüncül bir yaklaşımdır. Yeşil YZ'nin kapsamı sadece algoritmik optimizasyonla sınırlı kalmayıp; donanım tedarikinden, veri merkezi konumlandırmasına ve kullanım

ömrü sonu yönetimine kadar geniş bir alana yayılmaktadır. Bu paradigma karbon ayak izini azaltmak için dört ana stratejiye odaklanmaktadır:

Sorumlu Donanım Tedariki: Donanımın üretim aşamasındaki "gömülü karbon" miktarının hesaplanması ve azaltılması.

Döngüsel Ekonomi: Donanımların yeniden kullanımı, onarımı ve geri dönüştürülmesi süreçlerinin entegrasyonu.

Altyapı Optimizasyonu: İş yüklerinin, yenilenebilir enerji kaynaklarının yoğun olduğu bölgelere ve zaman dilimlerine kaydırılması.

İnsan Odaklı Tasarım: Gereksiz karmaşıklık yerine, ihtiyaca uygun ölçekte ve enerji verimli modellerin tercih edilmesi.

Tablo 1, Kırmızı ve Yeşil YZ arasındaki farkları özetlemektedir.

Tablo 1 Kırmızı ve Yeşil YZ Arasındaki Temel Farklılıklar

	Kırmızı YZ	Yeşil YZ
Birincil Hedef	En yüksek doğruluk (SOTA) ve performans	Enerji verimliliği, sürdürülebilirlik ve yeterli doğruluk
Kaynak Kullanımı	Maliyet ve enerji kısıtı gözetilmeksizin maksimizasyon	Kaynak kısıtlı optimizasyon, minimum enerji ile maksimum çıktı
Model Mimarisi	Devasa parametre sayıları, yoğun (dense) ağlar	Kompakt modeller, seyrek (sparse) ağlar, Küçük Dil Modelleri (SLM)
Erişilebilirlik	Erişilebilirlik düşüktür. Yüksek donanım ve bütçe gerektirir	Erişilebilirlik yüksektir. Demokratikleştirilmiş, uç cihazlarda çalışabilir
Çevresel Etki	Yüksek karbon ayak izi ve su tüketimi	Minimize edilmiş ekolojik etki, yaşam döngüsü analizi

3. Yapay Zekâ Sistemlerinin Çevresel Etkileri

YZ sistemlerinin kullanım alanları genişledikçe çevresel etkileri de giderek daha görünür hâle gelmektedir. Sürdürülebilirlik odağında yapılan çalışmalar, YZ modellerinin yalnızca eğitim süreçlerinde değil; veri hazırlama, depolama, donanım üretimi ve çıkarım (inference) aşamalarında da önemli enerji ve kaynak tüketimine yol açtığını göstermektedir (Patterson et al., 2021). Bu nedenle YZ'nin çevresel etkileri, bütüncül bir yaşam döngüsü yaklaşımıyla değerlendirilmelidir.

3.1. Donanım Üretimi ve Kaynak Tüketimi

YZ sistemlerinin temel yapı taşlarından biri olan donanım bileşenleri (GPU, TPU, işlemciler, devreler) yüksek miktarda doğal kaynak kullanımına ihtiyaç duymaktadır. Yarı iletken üretimi, su tüketimi, kimyasal atık oluşumu ve madencilik faaliyetleri nedeniyle ciddi çevresel etkiler oluşturmaktadır (Belkhir & Elmeligi, 2018). Özellikle nadir toprak elementlerinin çıkarılması hem ekolojik tahribata hem de yoğun karbon üretimine neden olmaktadır.

3.2. Veri Merkezlerinin Enerji Tüketimi

YZ modellerinin eğitildiği veri merkezleri, küresel enerji tüketiminin önemli bir bölümünü oluşturmaktadır. 2030 yılına kadar veri merkezlerinin küresel elektrik talebinin %8'ine ulaşabileceği öngörülmektedir (Andrae, 2020). YZ eğitimi yoğun hesaplama gerektirdiği için, bu merkezlerde kullanılan soğutma sistemleri ve yüksek performanslı işlemciler toplam enerji tüketimini daha da artırmaktadır.

Özellikle günümüzün popüler teknolojileri olan Büyük Dil Modellerinin (Large Language Model, LLM) eğitimi, binlerce GPU'nun (Grafik İşlem Birimi) aylar süren yoğun hesaplama işlemlerini gerektirmektedir. Bu süreçte tüketilen enerji miktarı,

modelin parametre sayısı, eğitim veri setinin büyüklüğü ve kullanılan veri merkezinin Enerji Kullanım Verimliliği (Power Usage Effectiveness, PUE) oranına bağlı olarak değişmektedir.

Literatürdeki çarpıcı karşılaştırmalardan biri, OpenAI'nin GPT-3 modeli ile BigScience projesi kapsamında geliştirilen BLOOM modeli arasındadır. GPT-3'ün eğitimi sırasında yaklaşık 1.287 MWh enerji tüketildiği ve bunun sonucunda 502 ton CO_2'ye eşdeğer emisyon oluştuğu tahmin edilmektedir. Bu miktar, yüzlerce aracın yıllık emisyonuna veya New York ile San Francisco arasında yüzlerce uçuşa eşdeğerdir (Visual Capitalist, 2024). Buna karşılık, benzer boyuttaki (176 milyar parametre) BLOOM modelinin eğitimi, Fransa'daki Jean Zay süper bilgisayarında gerçekleştirilmiştir ve Fransa'nın nükleer ağırlıklı düşük karbonlu enerji şebekesi sayesinde, BLOOM'un eğitimi sadece 25 ton CO_2'ye eşdeğer (donanım üretimi dahil edildiğinde 50.5 ton) emisyona neden olmuştur (Luccioni et al., 2023). Bu sonuçlar, algoritmanın etkinliği kadar kullanılan enerjinin kaynağının da Yeşil YZ için büyük öneme sahip olduğunu göstermektedir.

3.3. Model Eğitimi ve Karbon Ayak İzi

Model eğitimi, YZ sistemlerinin çevresel etkilerinin en belirgin olduğu aşamadır. Derin öğrenme modellerinin, milyarlarca parametreyi optimize etmek için çok uzun süreli GPU/TPU çalıştırmaları gerektirir.

Lacoste ve arkadaşlarının (2019) geliştirdiği ML CO_2 Impact modeli, eğitim sırasında kullanılan donanım, lokasyon ve enerji kaynağına göre karbon ayak izinin değişim gösterdiğini ortaya koymaktadır. Bu çalışma, YZ modellerinin enerji ve karbon muhasebesi yapılması gerektiğini vurgular. Önde gelen bazı LLM modellerine ait karşılaştırmalar Tablo 2'de verilmiştir.

Tablo 2 Önde Gelen LLM Modellerinin Eğitim Aşaması Enerji ve Karbon Ayak İzi Karşılaştırması (Luccioni et al., 2023)

Model	Parametre (Milyar)	Enerji Tüketimi (MWh)	CO2 Eşdeğer Emisyon (Ton)	Enerji Kaynağı / Lokasyon	Enerji Kullanım Verimliliği
GPT-3	175	1.287	502	Karışık / ABD	1.1
Gopher	280	1.066	352	Karışık	1.08
OPT	175	324	70	Karışık	1.09
BLOOM	176	433	25	Nükleer / Fransa	1.2

3.4. Çıkarım (Inference) Aşamasındaki Enerji Maliyeti

Yapay zekâ modellerinin eğitimi tek seferlik (veya periyodik) bir olayken, çıkarım aşaması modelin ömrü boyunca sürekli devam etmektedir. ChatGPT, Gemini veya Claude gibi modellerin milyonlarca kullanıcı tarafından her gün milyarlarca kez kullanılması, enerji tüketim dengesini eğitimden çıkarıma doğru kaydırmaktadır.

2024 ve 2025 yılı verilerine göre, popüler bir LLM'nin yaşam döngüsünde emisyonlarının %90'ına varan kısmı çıkarım aşamasında oluşmaktadır (Fu et al., 2024). Araştırmalar, ChatGPT ile yapılan tek bir sohbetin, standart bir Google aramasından 10 ila 20 kat daha fazla enerji tükettiğini ortaya koymaktadır. Bazı tahminlere göre, ChatGPT üzerindeki her bir sorgu yaklaşık 4.32 gram CO_2 salımına neden olurken, bir web araması yalnız 0.2 gram seviyesindedir (Plan Be Eco, 2024).

Çıkarım sırasındaki enerji tüketimini etkileyen faktörler şunlardır:

- İstem (Prompt) Uzunluğu ve Karmaşıklığı: Daha uzun girdiler ve çıktılar, daha fazla GPU döngüsü gerektirir.

- Model Boyutu: Daha büyük modeller, her token üretimi için daha fazla bellek bant genişliği ve işlem gücü harcar.
- Donanım Verimliliği: Kullanılan GPU'ların nesli (örneğin NVIDIA H100 vs A100) ve sunucu konfigürasyonu tüketilen enerjiyi etkiler.

3.5. Gömülü Karbon

Yazılımın tükettiği elektriğin ötesinde, yapay zekayı mümkün kılan fiziksel altyapının (GPU'lar, sunucular, soğutma sistemleri) üretimi de ciddi bir çevresel maliyet taşımaktadır. Dell ve HP gibi donanım üreticilerinin yaşam döngüsü analizleri, bir sunucunun toplam karbon ayak izinin önemli bir kısmının, cihazın operasyonel hale gelmesinden önceki madencilik, üretim ve lojistik süreçlerinde (Gömülü Karbon) oluştuğunu göstermektedir (Różycki et al., 2025).

BLOOM modelinin detaylı analizinde, toplam 50 tonluk emisyonun yaklaşık %22'sinin donanım üretiminden, %28'inin ise veri merkezinin âtıl (idle) enerji tüketiminden kaynaklandığı belirlenmiştir (Luccioni et al., 2023). Bu bulgu, donanım yenileme hızının yavaşlatılması ve cihaz kullanım ömrünün uzatılmasının, Yeşil YZ stratejilerinin ayrılmaz bir parçası olması gerektiğini vurgulamaktadır.

3.6. E-Atık Sorunu ve Donanım Ömrü

YZ uygulamalarının yaygınlaşması, yüksek performanslı donanım talebini artırmakta ve donanım yenileme döngülerini hızlandırmaktadır. Bu durum büyük miktarda elektronik atık (e-waste) üretimine yol açmaktadır.

Birleşmiş Milletler'e göre e-atık, dünyanın en hızlı büyüyen atık kategorisi olup yılda 50 milyon tonu aşmaktadır (UNEP, 2020). Veri merkezlerinde donanım yenileme döngülerinin 2–3 yıla kadar

düşmesi, YZ altyapısının atık üretiminde önemli bir paya sahip olmasına neden olmaktadır. E-atığın yalnızca hacmi değil; içerdiği toksik maddeler (arsenik, kurşun, kadmiyum) ekosistem için ciddi tehdit oluşturmaktadır.

4. Yeşil Yapay Zekâ İçin Yöntemler ve Teknolojiler

Yeşil Yapay Zekâ, yalnızca bir değerlendirme yaklaşımı değil; aynı zamanda enerji verimliliğini artırmayı, karbon ayak izini azaltmayı ve daha sürdürülebilir YZ ekosistemleri oluşturmayı amaçlayan yöntem ve teknolojiler bütünü olarak görülmektedir. Bu kısımda, Yeşil YZ uygulamalarını mümkün kılan optimizasyon teknikleri, model mimarileri, eğitim süreçleri ve altyapı stratejileri ele alınmaktadır.

4.1. Enerji Verimli Mimari Tasarımlar

Enerji verimliliğini artırmanın en temel yollarından biri, model mimarisinin hesaplama maliyetini azaltacak şekilde optimize edilmesidir. Transformer mimarisi çeşitli görevlerde yüksek performans sunmasına rağmen, hesaplama karmaşıklığı kuadratik düzeydedir (Vaswani et al., 2017). Bu nedenle literatürde daha verimli alternatif mimariler geliştirilmiştir:

- Efficient Transformers: Sparse attention, linear attention ve reformer yapıları gibi düşük karmaşıklığa sahip modeller. (Kitaev et al., 2020).
- Mobil cihazlar için optimize modeller: MobileNet (Howard et al., 2017), EfficientNet (Tan & Le, 2019).
- Daha az parametreyle benzer performans sunan kompakt mimariler.

Bu mimarilerin temel avantajı, hesaplamayı hem bellek hem de enerji açısından daha az maliyetli hâle getirmeleridir.

4.2. Model Sıkıştırma Yöntemleri: Budama, Nicemleme, Bilgi Damıtma

Model sıkıştırma teknikleri, Yeşil YZ teknolojilerinin temelini oluşturmaktadır. Bu yöntemler hem eğitim hem de çıkarım maliyetlerini azaltarak enerji verimliliği sağlar.

4.2.1. Budama

Budama (pruning), modelin gereksiz nöronlarını veya bağlantılarını kaldırarak daha hafif yapılar elde edilmesini sağlar (Han et al., 2015).

- Avantaj: Hesaplama yükünde %50–90 düşüş sağlanabilir.
- Dezavantaj: Aşırı budama performans kaybına yol açabilir.

4.2.2. Nicemleme

Nicemleme (quantization), 16/32-bit temsil yerine 8-bit veya daha düşük bit genişliği kullanarak hesaplamayı hızlandırır (Jacob et al., 2018). Bu yöntem özellikle uç (edge) cihazlarda büyük tasarruf sağlar.

4.2.3. Bilgi Damıtma

Bilgi damıtma (Distillation), büyük bir modelin bilgisinin daha küçük ve verimli bir modele aktarılması tekniğidir (Hinton et al., 2015). Damıtılmış modeller çoğu zaman daha düşük enerji tüketimine rağmen benzer performans sunabilir.

4.3. Veri Verimliliği: Sıfır Örnekli, Az Örnekli ve Kendinden Denetimli Öğrenme

Veri toplama, işaretleme ve işleme süreçleri de çevresel maliyete sahiptir. Bu nedenle veri verimliliği yüksek öğrenme yöntemleri Yeşil YZ için kritiktir.

4.3.1. Sıfır Örnekli Öğrenme

Sıfır örnekli (zero-shot) öğrenme, bir modelin daha önce hiç görmediği bir sınıf veya görev için yalnızca sahip olduğu genel bilgi temeliyle doğru çıktılar üretebilmesini ifade eder. Model, eğitim aşamasında temas etmediği yeni kavramları anlamsal ilişkiler, kelime gömme (embedding) uzayları, metin açıklamaları, etiket ilişkileri gibi semantik bilgilere dayanarak tahmin eder. Bu yöntemin avantajları şunlardır:

- Büyük etiketli veri ihtiyacını önemli ölçüde azaltır.
- Hızlı uyarlanabilir sistemler için uygundur (örneğin chatbot'lar, görüntü tanıma).
- Enerji maliyetini düşürerek Yeşil YZ'ye katkı sağlar.

4.3.2. Az Örnekli Öğrenme

Az örnekli (few-shot) öğrenme, bir modelin yalnızca 2–50 arası etiketli örnek ile yeni bir göreve hızla uyum sağlamasıdır. Temel amaç, mevcut bilgi temellerini kullanarak çok az veriyle yüksek performans elde etmektir. Az örnekli öğrenmede meta öğrenme (öğrenmeyi öğrenme), temsile dayalı öğrenme ve benzerlik tabanlı teknikler kullanılır. Bu yöntemin avantajları:

- Geniş veri toplama eforunu azaltır.
- İnsan öğrenmesine daha yakın bir yaklaşım sunar.
- Eğitim süresi düşüktür bu da enerji maliyetini düşürür.

4.3.3. Özdenetimli Öğrenme

Özdenetimli (self-supervised) öğrenme, verilerin kendi iç yapısından otomatik etiketler (pseudo-labels) oluşturularak eğitim yapılmasıdır. Yani model, dışarıdan etiketlenmiş veri gerektirmeden öğrenebilir. Kullanılan temel teknikler maskeli öğrenme, karşıtlık

öğrenmesi, otoenkoder temelli yöntemlerdir. Bu teknik Yeşil YZ için oldukça değerlidir. Avantajları:

- Etiketleme için insan gücü ihtiyacını azaltır.
- Milyonlarca verilik veri setlerinde düşük maliyetle eğitim yapılabilir.

4.4. Uç Yapay Zekâ ve Federe Öğrenme ile Enerji Tasarrufu

Uç Yapay Zekâ (Edge AI), yapay zekâ işlemlerinin buluttaki merkezi sunucular yerine verinin üretildiği cihaz üzerinde (uçta) gerçekleştirilmesidir. Bu cihazlara örnek:

- Akıllı telefonlar
- IoT sensörleri
- Kamera sistemleri
- Araç içi bilgisayarlar
- Medikal cihazlar

Uç Yapay Zekâ, veriyi işlem için buluta göndermek yerine cihaz üzerinde işler ve karar verir. Bu yöntem; bant genişliği tüketimini, veri merkezi bağımlılığını ve sadece karbon ayak izini azaltır (Sze et al., 2017).

Federe öğrenme (federated learning), yapay zekâ sürecinin verilerini merkezi bir sunucuya toplamadan, verilerin bulunduğu cihazlarda yerel olarak gerçekleştirilmesini sağlayan bir yaklaşımdır. Temel mantık:

Model cihazlara gönderilir.

Cihazlar verilerini kullanarak modeli yerelde eğitir.

Güncellenmiş model parametreleri (veri değil) merkeze gönderilir.

Merkezde bu güncellemeler birleştirilir.

Ortak bir küresel model elde edilir.

Federe öğrenme yönteminin avantajları şunlardır:

- Veri gizliliğini artırır,
- Veri iletim maliyetini düşürür,
- Büyük veri kümelerinin merkezi depolanmasının gerekmediği senaryolarda daha düşük enerji tüketimi sağlar.

4.5. Yenilenebilir Enerji ile Çalışan YZ Altyapıları

Yapay zekâ veri merkezlerinin çevresel etkisini azaltmanın bir diğer yolu, bu altyapıların yenilenebilir enerji kaynaklarıyla çalıştırılmasıdır. Google ve Microsoft gibi şirketler veri merkezlerinde %100 yenilenebilir enerji hedeflerine yönelmiştir (Google Sustainability Report, 2023).

Bunun yanında "karbon bilincine sahip zamanlama" (carbon-aware scheduling) yaklaşımı, model eğitimini enerji şebekesinin karbon yoğunluğunun düşük olduğu zaman dilimlerine planlayarak emisyonları azaltmayı hedefler.

5. Sürdürülebilirlik İçin Yapay Zekâ Uygulamaları

Sürdürülebilirlik; çevresel, ekonomik ve toplumsal bileşenleri kapsayan bütüncül bir dönüşüm gerektirmektedir. YZ, veri işleme kapasitesi, tahmin kabiliyeti ve otomasyon potansiyeli sayesinde sürdürülebilirlik hedeflerine doğrudan katkı sağlayabilecek temel teknolojiler arasında bulunmaktadır.

5.1. İklim Değişikliği Modellemesi ve Erken Uyarı Sistemleri

İklim sistemleri oldukça karmaşık, çok değişkenli ve uzun vadeli dinamiklere sahiptir. YZ, bu karmaşık yapıları daha doğru ve

hızlı şekilde modellemek için kullanılabilmektedir. Derin sinir ağları, tekrarlayan sinir ağları (RNN), evrişimli ağlar (CNN) ve artık son yıllarda transformatör temelli modeller ile iklim simülasyonlarının doğruluğu artmıştır (Reichstein et al., 2019).

YZ destekli iklim modelleri şu alanlarda öne çıkmaktadır:

- Aşırı hava olaylarının tahmini: Sel, fırtına, sıcak hava dalgaları.
- Uzun vadeli iklim projeksiyonları: Karbon emisyon senaryolarının değerlendirilmesi.
- Uydu verilerinin işlenmesi: Bulut, aerosol, bitki örtüsü ve deniz yüzeyi sıcaklık değişimlerinin analizi.

5.2. Akıllı Enerji Yönetimi ve Akıllı Şehir Uygulamaları

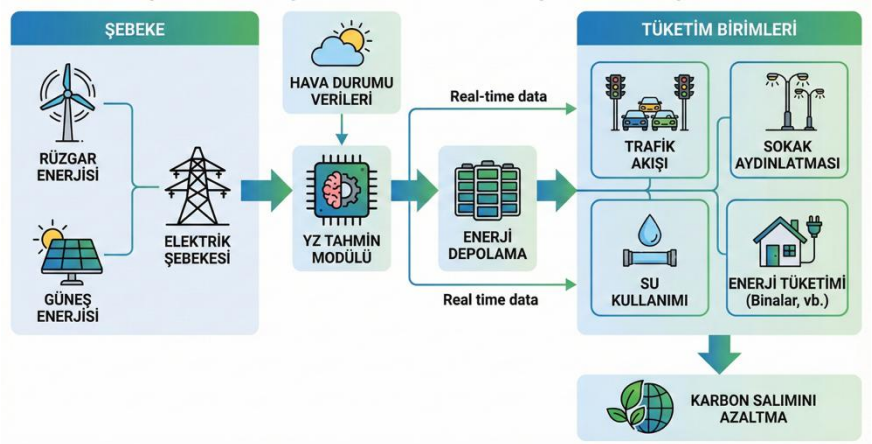
Enerji tüketiminin optimize edilmesi sürdürülebilirliğin temel bileşenlerinden biridir. YZ, enerji yönetiminde özellikle:

- Talep tahmini,
- Akıllı şebeke (smart grid) optimizasyonu,
- Enerji depolama planlaması,
- Binalarda enerji verimliliği,
- Yenilenebilir enerji üretiminin kestirimi

gibi alanlarda kullanılmaktadır (Alova, 2020).

Örneğin rüzgâr ve güneş enerjisi üretimi doğası gereği değişkendir. YZ, hava durumu verilerini kullanarak üretimi öngörebilir ve şebeke dengesini sağlamaya yardımcı olabilir. Akıllı şehir sistemlerinde ise YZ; trafik akışı, sokak aydınlatması, su kullanımı ve enerji tüketimini gerçek zamanlı yöneterek karbon salımını azaltır (Şekil 1).

Şekil 1 Akıllı şehirlerde YZ destekli enerji yönetim bileşenleri



5.3. Sürdürülebilir Tarım ve Hassas Tarım Teknolojileri

YZ, tarımsal üretimin hem verimliliğini hem de sürdürülebilirliğini artırmak için kritik bir rol üstlenmektedir. Hassas tarım (precision agriculture) kapsamında YZ şu uygulamalarda kullanılmaktadır:

- Dron ve uydu görüntüleriyle ürün sağlığı analizi (Kamilaris & Prenafeta-Boldú, 2018)
- Sulama optimizasyonu
- Toprak nemi, pH ve besin seviyesine dayalı gübreleme planlaması
- Zararlı tespiti ve hastalık sınıflandırması
- Hasat zamanının tahmini

Bu uygulamalar kaynak kullanımını azaltarak çevresel sürdürülebilirliğe önemli katkı sağlar.

5.4. Su Kaynaklarının Yönetiminde YZ Uygulamaları

İklim değişikliği nedeniyle su kıtlığı ve kaynakların verimsiz yönetimi giderek kritik hâle gelmiştir. YZ, su kaynaklarının korunmasında şu alanlarda kullanılmaktadır:

- Su talebi tahmini
- Su kayıplarının (kaçak) tespiti
- Yeraltı su seviyelerinin izlenmesi
- Taşkın risk analizi
- Sulama sistemlerinin otonom kontrolü

Çeşitli çalışmalar, YZ modellerinin geleneksel hidrolik modellere kıyasla daha doğru tahminler sunduğunu göstermiştir (Mosavi, Ozturk & Chau, 2018).

5.5. Atık Yönetimi ve Geri Dönüşüm Süreçlerinde YZ

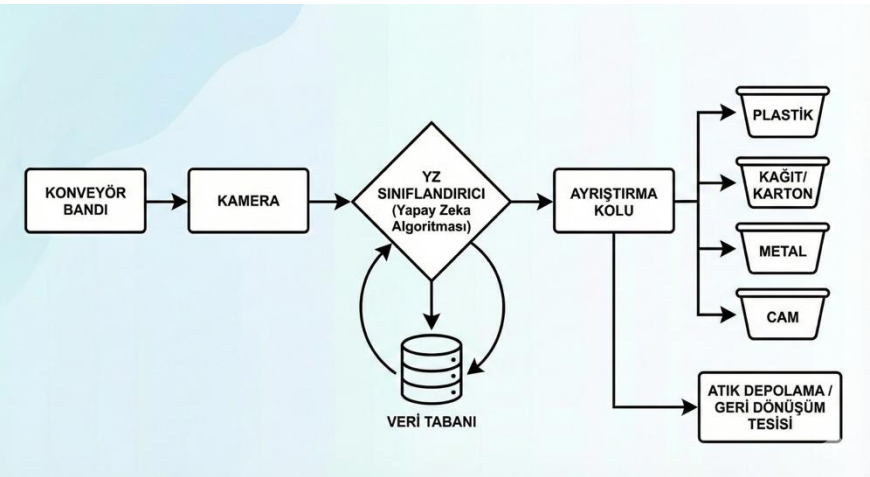
Atık yönetimi, sürdürülebilirlik politikalarının temel bileşenlerinden biridir. YZ atık yönetimi süreçlerinde:

- Geri dönüşebilir malzemelerin otomatik sınıflandırılması,
- Atık toplama güzergâhlarının optimizasyonu,
- Atık üretim miktarının tahmini,
- Atık ayrıştırma robotları,
- Geri dönüşüm tesislerinde kalite kontrol

gibi uygulamalarda kullanılmaktadır.

Özellikle derin öğrenme temelli görüntü işleme teknikleri, geri dönüşüm hatlarında doğruluk ve hız artışı sağlamaktadır (Şekil 2).

Şekil 2 YZ tabanlı geri dönüşüm ayrıştırma sistemi



5.6. Biyoçeşitliliğin Korunmasında YZ

Biyoçeşitlilik kaybı küresel ekosistem için önemli bir risk oluşturmaktadır. YZ, bu alanda aşağıdaki uygulamalarda kullanılmaktadır:

- Vahşi yaşam izleme (kamera tuzakları, akustik sensörler)
- Nesli tükenmekte olan türlerin tespiti
- Ekosistem değişikliklerinin uydu görüntüleriyle analizi
- Kaçak avcılık tespiti için erken uyarı sistemleri
- Tür sınıflandırma modelleri

6. Yeşil YZ Politikaları, Standartlar ve Regülasyonlar

YZ sistemlerinin artan enerji tüketimi ve çevresel etkileri hem ulusal hem de uluslararası düzeyde düzenleyici çerçevelerin oluşturulmasını zorunlu hâle getirmiştir. Yeşil YZ politikaları; yalnızca teknolojinin geliştirilme sürecini değil, aynı zamanda enerji altyapısı, veri yönetimi, etik ilkeler ve kurumsal sürdürülebilirlik

standartlarını da kapsamaktadır. Bu bağlamda, çeşitli ülkeler ve uluslararası kuruluşlar çevresel açıdan sorumlu YZ kullanımını teşvik eden stratejiler geliştirmektedir.

6.1. Uluslararası Stratejiler ve Politikalar

6.1.1. Avrupa Birliği (AB) Yaklaşımları

Avrupa Birliği, YZ'nin çevresel etkilerinin azaltılması konusunda en kapsamlı stratejilere sahip bölgesel aktörlerden biridir. Avrupa Yeşil Mutabakatı (European Green Deal) kapsamında veri merkezlerinin 2030 yılına kadar karbon-nötr olması hedeflenmiştir (European Commission, 2019).

Ayrıca AB Yapay Zekâ Yasası (AI Act), yüksek riskli YZ sistemleri için enerji verimliliği raporlama ve çevresel etki değerlendirmesi gibi yükümlülükleri gündeme getirmektedir. Bu düzenleme, YZ sistemlerinin yalnızca etik ve güvenlik açısından değil, çevresel sürdürülebilirlik açısından da hesap verebilir olmasını hedefler.

6.1.2. OECD Yapay Zekâ İlkeleri

OECD'nin 2019 yılında kabul ettiği Yapay Zekâ İlkeleri, çevresel sürdürülebilirliği doğrudan vurgulayan uluslararası standartlardan biridir. İlkelerden biri, YZ'nin "kapsayıcı büyüme, sürdürülebilir kalkınma ve refah" için kullanılmasını teşvik eder (OECD, 2019). Bu çerçeve, Yeşil YZ yaklaşımlarının küresel ölçekte benimsenmesini desteklemektedir.

6.1.3. Birleşmiş Milletler ve UNEP Yaklaşımları

Birleşmiş Milletler Çevre Programı (UNEP), dijitalleşmenin çevresel etkilerini azaltmaya yönelik çeşitli rehberler hazırlamıştır. Özellikle Dijital Sürdürülebilirlik Girişimleri kapsamında YZ'nin

enerji kullanımının şeffaflaştırılması ve sürdürülebilir altyapılarla desteklenmesi önerilmektedir (UNEP, 2020).

6.2. Kurumsal Sürdürülebilirlik Hedefleri

Küresel ölçekli teknoloji şirketleri, çevresel etkilerini azaltmak amacıyla Çevresel-Sosyal-Yönetişim (Environmental, Social, Governance – ESG) ilkelerini stratejik planlamalarına entegre etmektedir. Google, Microsoft ve Amazon gibi şirketler;

- %100 yenilenebilir enerji ile çalışma hedefleri,
- Karbon negatif veya karbon nötr olma taahhütleri,
- Veri merkezlerinde ısı geri-kazanım sistemleri,
- Algoritmik verimlilik artırma programları

gibi uygulamalar geliştirmektedir.

Bu yaklaşımlar YZ'nin yalnızca performans odaklı değil, sürdürülebilirlik odaklı değerlendirilmesi gerektiğini kurumsal açıdan pekiştirmektedir.

6.3. Karbon-Nötr YZ Girişimleri ve Standardizasyon Çabaları

YZ altyapısının karbon ayak izinin hesaplanabilir ve şeffaf olması için çeşitli araştırma grupları ve kurumlar standartlar geliştirmektedir.

6.3.1. Karbon Muhasebesi Araçları

Bu araçların ortak amacı, YZ çalışmalarının enerji ve karbon etkisini hesaplamayı standartlaştırmaktır. Bu araçlardan başlıcaları:

- ML CO_2 Impact (Lacoste et al., 2019): Model eğitim işlemlerinin karbon salımını hesaplamak için kullanılan en bilinen araçlardan biridir.

- Carbontracker (Anthony et al., 2020): YZ model eğitim süreçlerinin enerji kullanımını ölçmek için geliştirilen bir Python kütüphanesidir.
- CodeCarbon (Schmidt et al., 2021): Çeşitli makine öğrenimi framework'leri ile entegre olabilen açık kaynak bir karbon izleme aracıdır.

6.3.2. ISO Standartları ve Gelişen Regülasyonlar

ISO/IEC JTC 1/SC 42 komitesi, yapay zekâ için standartlar geliştirmekle sorumludur. Henüz Yeşil YZ'ye özel bir uluslararası standart bulunmasa da ISO 14064 (karbon ayak izi hesaplama), ISO 50001 (enerji yönetim sistemleri) standartları YZ altyapılarında dolaylı olarak kullanılmaktadır. Yakın gelecekte YZ karbon raporlama standartlarının daha net biçimde tanımlanması beklenmektedir.

6.4. Etik Çerçeveler ve Sorumlu YZ İlkeleri

Yeşil YZ, yalnızca teknik bir optimizasyon meselesi değil aynı zamanda etik bir gerekliliktir. Çevresel etkilerin göz ardı edilmesi, YZ sistemlerinin toplumsal eşitsizlikleri derinleştirmesine ve çevresel adaletsizliklere yol açabilir. Etik çerçevelerde öne çıkan ilkeler:

- Hesap verebilirlik: Çevresel etkilerin şeffaf raporlanması
- Sürdürülebilirlik: YZ modellerinin çevreye duyarlı geliştirilmesi
- Adalet: Çevresel yükün eşitsiz dağılımının önlenmesi
- Erişilebilirlik: Yüksek hesaplama gücüne dayalı YZ modellerinin yarattığı dijital uçurumun azaltılmasıdır.

6.5. Türkiye'de Yeşil YZ Politikaları ve Stratejiler

Türkiye, 2021 yılında yayımlanan Ulusal Yapay Zekâ Stratejisi ile YZ yatırımlarını artırmayı, YZ ekosistemini geliştirmeyi ve veri altyapısını güçlendirmeyi hedeflemiştir (T.C. Sanayi ve Teknoloji Bakanlığı, 2021). Ancak Yeşil YZ'ye yönelik düzenlemeler henüz başlangıç aşamasındadır. Türkiye'de çevresel sürdürülebilirlik ve dijitalleşme alanındaki politikalar;

- Yeşil Mutabakat Eylem Planı,
- Sıfır Atık Politikası,
- Dijital Türkiye girişimi,
- Enerji verimliliği eylem planları

ile şekillenmektedir. Bu stratejiler, Yeşil YZ uygulamalarının hayata geçirilmesi için önemli bir zemin sunmaktadır.

Türkiye'nin gelecekte atması gereken adımlar arasında:

- YZ projelerinde karbon raporlama zorunluluğu,
- Veri merkezlerinde yenilenebilir enerji kullanım oranlarının artırılması,
- Yerli Yeşil YZ araştırmalarının desteklenmesi

sıralanabilir.

7. Yeşil YZ Ekosistemi ve Endüstriyel Yaklaşımlar

Yeşil YZ ekosistemi; kamu kurumları, özel sektör kuruluşları, araştırma laboratuvarları, sivil toplum örgütleri ve uluslararası kuruluşların ortak katkılarıyla şekillenen çok bileşenli bir yapıdır. Bu ekosistem, hem YZ modellerinin enerji verimliliğini artırmayı hem de çevresel sürdürülebilirliği destekleyen uygulamaların yaygınlaştırılmasını hedeflemektedir.

7.1. Büyük Teknoloji Şirketlerinin Yeşil Dönüşüm Çalışmaları

7.1.1. Google

Google, 2030 yılına kadar 24 saat tamamen karbon-sıfır enerjiyle çalışma hedefi koymuştur (Google, 2023). Şirket, veri merkezlerinde;

- Karbon-farkındalıklı zamanlama (carbon-aware scheduling),
- Verimli soğutma sistemleri,
- YZ tabanlı enerji optimizasyonu

kullanmaktadır.

7.1.2. Microsoft

Microsoft, 2030 yılına kadar karbon-negatif olacağını açıklamış ve veri merkezlerinde enerji kullanımını YZ destekli izleme sistemleriyle optimize etmektedir (Microsoft, 2020).

7.1.3. Amazon

Amazon, 2025 yılına kadar tüm operasyonlarında %100 yenilenebilir enerjiye geçiş hedefini duyurmuştur. Ayrıca EC2 altyapısında daha enerji verimli Graviton işlemcileri kullanarak karbon yoğunluğunu azaltmaktadır.

7.2. Yeşil YZ Start-up Ekosistemi ve Yenilikçi Girişimler

Yeni nesil girişimler Yeşil YZ teknolojilerinin geliştirilmesinde yenilikçi çözümler sunmaktadır. Bu girişimler genellikle enerji izleme, karbon azaltım teknolojileri, verimli model eğitimi ve sürdürülebilir altyapılar üzerine yoğunlaşmaktadır.

Öne çıkan start-up türleri:

- Karbon izleme ve şeffaflık sistemleri geliştiren girişimler

- Verimli YZ modelleri ve optimizasyon araçları sunan girişimler
- Veri merkezleri için enerji optimizasyon yazılımları sunan şirketler
- Atık yönetimi ve geri dönüşüm için YZ çözümleri geliştiren girişimler

Bu girişimler, geleneksel büyük teknoloji şirketlerine göre daha çevik oldukları için Yeşil YZ inovasyonunun hızını artırmaktadır.

7.3. Bilimsel Araştırma Eğilimleri ve Akademik Laboratuvarlar

Yeşil YZ araştırmaları, özellikle son beş yılda hızla yükselen bir akademik alan haline gelmiştir. Öne çıkan araştırma toplulukları ve laboratuvarlar şunlardır:

- MILA – Quebec AI Institute: Enerji verimli öğrenme ve karbon ayak izi hesaplaması üzerine öncü çalışmalar (örn. CodeCarbon kütüphanesi) gerçekleştirmektedir.
- Allen Institute for AI (AI2): Yeşil Yapay Zekâ kavramını sistematik şekilde ilk ortaya koyan kurumlardan biridir (Schwartz et al., 2020).
- Stanford Sustainability & AI Lab: Sürdürülebilir tarım, enerji ve iklim araştırmaları için YZ modelleri geliştirmektedir.
- DeepMind: Veri merkezi soğutma sistemlerinde YZ tabanlı kontrol algoritmaları ile enerji tüketimini %40'a varan oranlarda azaltmıştır (Evans & Gao, 2016).

7.4. Açık Kaynak Araçlar ve Topluluk Destekleri

Açık kaynak ekosistemi, Yeşil YZ'nin yaygınlaşmasında kritik bir role sahiptir. Aşağıdaki araçlar ve topluluklar, enerji ve karbon odaklı YZ geliştirme süreçlerini desteklemektedir:

- CodeCarbon: ML işlemlerinin karbon etkisini izleyen açık kaynak araç (Schmidt et al., 2021).
- Carbontracker: Eğitim süresince enerji tüketimini tahmin eden ve raporlayan Python tabanlı araç.
- Hugging Face – Optimum Framework: Transformers modelleri için quantization, pruning, distillation gibi verimlilik tekniklerini kolaylaştırır.
- TensorFlow Model Optimization Toolkit: Küçültülmüş, düşük enerji tüketen YZ modelleri geliştirmek için kullanılan araç seti.
- PyTorch FX ve seyrekleştirme (sparsity) araçları: Model sıkıştırma ve düşük karmaşıklıklı mimari oluşturma için kullanılır.

Bu araçlar, araştırmacıların enerji verimli modeller geliştirmesini kolaylaştırarak Yeşil YZ'nin demokratikleşmesini sağlamaktadır.

7.5. Sürdürülebilir Donanım Tasarımı ve Yeni Nesil İşlemciler

Enerji verimli YZ sistemleri için donanım seviyesinde geliştirilen teknolojiler de ekosistemin kritik bir parçasıdır. Enerji verimli donanım yaklaşımlarına örnek olarak şunlar verilebilir:

- ARM tabanlı işlemciler (örn: AWS Graviton)
- Daha az güç tüketen GPU'lar
- Nöral işleme birimleri (NPU)

- Uç (edge) cihazlar için optimize edilmiş hızlandırıcılar
- Yaklaşık hesaplama (approximate computing) yaklaşımları

Donanım inovasyonunun amacı, daha düşük güç tüketimiyle daha yüksek hesaplama kapasitesi sunmaktır.

8. Zorluklar, Riskler ve Tartışmalar

Yeşil YZ yaklaşımı, sürdürülebilir kalkınmayı destekleyen önemli fırsatlar sunmasına rağmen çok yönlü zorluklar ve tartışmalara da sahiptir. Bu tartışmalar; teknik sınırlılıklar, ölçeklenebilirlik problemleri, enerji-performans dengesi, etik sorumluluklar, dijital eşitsizlikler ve "yeşil yıkama" (greenwashing) risklerini içermektedir.

8.1. Enerji – Performans İkilemi

YZ modellerinde performans, özellikle son yıllarda büyük ölçekli derin öğrenme modellerinin yaygınlaşmasıyla birlikte parametre sayısı, veri büyüklüğü ve hesaplama kapasitesine bağımlı hale gelmiştir (Kaplan et al., 2020). Bu durum, daha iyi sonuç elde etmek için daha büyük modellerin eğitilmesini teşvik etmektedir. Ancak model boyutunun artması; eğitim süresini, hesaplama maliyetini, enerji tüketimini ve karbon salımını doğrudan yükseltmektedir.

Bu nedenle Yeşil YZ'nin temel tartışmalarından biri, "Enerji verimliliği ile model performansını aynı anda artırmak mümkün müdür?" sorusuna odaklanmaktadır. Bazı durumlarda küçük ve optimize modeller, büyük modellerle benzer performans sağlayabilse de özellikle büyük dil modelleri gibi alanlarda performans–enerji dengesi hâlâ bilimsel tartışma konusudur.

8.2. Veri Merkezlerinin Sürdürülebilirlik Kapasitesindeki Sınırlılıklar

YZ'nin çevresel etkileri yalnızca model seviyesinde değil; veri merkezlerinin altyapısında da belirgin hale gelmektedir. Veri merkezlerinin büyüklüğü ve enerji ihtiyacı arttıkça;

- yenilenebilir enerjiye erişim sınırlı kalabilmekte,
- soğutma sistemleri aşırı enerji tüketebilmekte,
- atık ısı geri kazanımı her zaman uygulanamayabilmekte,
- konumsal eşitsizlikler nedeniyle bazı bölgelerde daha yüksek karbon yoğunluğuna bağlı enerji kullanılabilmektedir.

IEA (2022) raporlarına göre veri merkezlerinin küresel elektrik talebi sürekli artmakta ve sürdürülebilir enerji kaynakları bu talebi her zaman karşılayamamaktadır. Bu durum Yeşil YZ stratejilerinin hayata geçirilmesinde altyapı temelli engeller yaratmaktadır.

8.3. Algoritmik Sürdürülemezlik: "Bigger is Better" Paradigmasının Sınırları

YZ araştırmalarında uzun süre "bigger is better" (daha büyük modellerin daha iyi olduğu) yaklaşımı hâkim olmuştur. Bu yaklaşım; daha büyük veri setleri, daha derin ağlar, daha fazla parametre ve daha uzun eğitim süreleri ile daha iyi sonuç elde edileceğini varsaymaktadır. Ancak Strubell et al. (2019), dev modellerin makul performans artışına karşın aşırı ölçekte enerji maliyetlerinin sürdürülebilir olmadığını göstermiştir. Bu durumda iki temel risk ortaya çıkmaktadır:

- Ölçeksel verimsizlik: Model büyüdükçe performans artışı azalmakta, fakat karbon maliyeti katlanarak artmaktadır.
- Araştırma eşitsizliği: Sadece büyük şirketler bu modelleri eğitebildiği için akademik kuruluşlar ve küçük laboratuvarlar rekabet gücünü kaybetmektedir.

Bu sorunlar, Yeşil YZ'nin etik ve erişilebilirlik boyutlarını da gündeme getirmektedir.

8.4. Yeşil Yıkama (Greenwashing) Riski

Teknoloji şirketlerinin sürdürülebilirlik iddiaları her zaman gerçek veriler ile örtüşmeyebilir. Bazı şirketler;

- karbon nötr hedeflerini yalnızca karbon kredileriyle sağlamakta,
- veri merkezlerinde kullanılan toplam enerji miktarını paylaşmamakta,
- YZ modellerinin gerçek enerji-maliyet raporlarını açıklamamakta,
- optimizasyon tekniklerini yalnızca pazarlama aracı olarak sunmaktadır.

AI Now Institute (2021), birçok şirketin enerji tüketimini sistematik olarak raporlamadığını, çevresel etkilerin çoğu zaman şeffaflık eksikliği nedeniyle kamunun denetiminden uzak kaldığını belirtmektedir. Bu durum, Yeşil YZ çalışmalarının güvenilirliğini azaltan önemli bir tartışmadır.

8.5. Dijital Erişim, Adalet ve Eşitsizlik Sorunları

YZ altyapılarının yüksek enerji ve donanım gereksinimleri, teknolojinin erişilebilirliğini sınırlamakta ve dijital eşitsizliği

artırmaktadır. Büyük modelleri eğitmek için gereken donanım maliyetleri özellikle gelişmekte olan ülkeler, küçük araştırma laboratuvarları, kamu üniversiteleri ve sivil toplum kuruluşları için erişilemez hâle gelebilmektedir.

Bu durum, "teknolojik güç yoğunlaşması" (technology concentration) olarak adlandırılmakta ve çevresel sürdürülebilirlik ile toplumsal adalet arasındaki ilişkiyi tartışmaya açmaktadır (Birhane, 2022). Nihai olarak, YZ geliştirmenin maliyetinin yalnızca çevresel değil; toplumsal bir maliyet olarak da ele alınması gerekmektedir.

8.6. Düzenleyici Çerçevelerdeki Boşluklar ve Uygulama Zorlukları

Yeşil YZ politikaları gelişmekte olan bir alandır ve düzenleyici çerçeveler henüz tam olgunlaşmış değildir. Öne çıkan boşluklar şunlardır:

- Uluslararası düzeyde karbon muhasebesi standardı yoktur.
- YZ eğitim süreçlerinin enerji tüketimini raporlama zorunluluğu çoğu ülkede bulunmamaktadır.
- Şeffaf veri paylaşımı eksiktir.
- Donanım üretim döngüsünün çevresel etkileri çoğunlukla göz ardı edilmektedir.
- Yeşil YZ sertifikasyon süreçleri henüz oluşmamıştır.

Bu eksiklikler, Yeşil YZ uygulamalarının ölçeklenmesini ve hesap verebilirliğini zorlaştırmaktadır.

9. Gelecek Perspektifi

Yeşil YZ, enerji verimliliğini merkezine alan ve çevresel etkileri azaltmayı hedefleyen bütüncül bir yaklaşımdır. Bu yaklaşımın geleceği, temel araştırma alanlarının gelişmesi, donanım ve yazılım optimizasyonlarının ilerlemesi, düzenleyici çerçevelerin olgunlaşması ve küresel sürdürülebilirlik hedefleriyle bütünleşmesi ile şekillenecektir. Gelecekte Yeşil YZ'nin yalnızca teknik çözümlerden ibaret olmayacağı; etik, ekonomik ve toplumsal boyutları içeren kapsamlı bir dönüşüm yaratacağı öngörülmektedir.

9.1. Ultra Verimli Model Mimarileri ve Hesaplama Yaklaşımları

Önümüzdeki yıllarda YZ araştırmalarının en kritik odaklarından biri, ultra verimli model mimarileri geliştirmek olacaktır. Şu anda bile model sıkıştırma, düşük-bit nicemleme, budama ve bilgi damıtma teknikleri enerji kullanımını ciddi ölçüde azaltmaktadır. Ancak gelecekte bu tekniklerin ötesine geçilerek;

- Sıfır atık hesaplama yaklaşımları,
- Düşük enerji tüketimli nöral mimariler,
- Hafıza verimliliği yüksek YZ çekirdekleri,
- Adaptif karmaşıklığa sahip modeller (gerektikçe büyüyen veya küçülen modeller),
- Nöral ağlar yerine enerji verimli hesaplama paradigmaları

geliştirileceği öngörülmektedir.

9.2. YZ'de Karbon Muhasebesinin Standardizasyonu

Gelecekte YZ projelerinin karbon ayak izlerinin rutin olarak hesaplanması ve raporlanması beklenmektedir. Bugün hâlâ farklı

araçlar (CodeCarbon, Carbontracker) farklı ölçüm yöntemleri sunmakta, küresel bir standart bulunmamaktadır.

Uluslararası kuruluşların (ISO, OECD, Avrupa Komisyonu) enerji tüketimi ve karbon salımı raporlamasını zorunlu kılması hâlinde;

- her YZ modeli için "karbon etiketi",
- eğitim süreci boyunca gerçek zamanlı karbon izleme,
- veri merkezleri için karbon yoğunluk sertifikasyonu,
- araştırma yayınlarında zorunlu enerji raporlama bölümleri

gibi uygulamaların yaygınlaşması beklenmektedir.

Bu süreç, YZ araştırmalarının şeffaflığını artıracak ve sürdürülebilirlik konusunda hesap verebilirlik oluşturacaktır.

9.3. Otonom Sistemlerde Sürdürülebilirlik Odaklı Optimizasyon

Akıllı şehirler, ulaşım sistemleri, tarım robotları ve endüstriyel otomasyon sistemleri gibi otonom yapılar gelecekte YZ destekli sürdürülebilirlik uygulamalarının kritik bileşenleri olacaktır. Örneğin:

- Otonom araçlarda enerji verimli rota planlama,
- Akıllı şehirlerde karbon yoğunluğunu azaltan trafik optimizasyonu,
- Tarım robotlarında su ve gübre kullanımının en aza indirilmesi,
- Lojistikte karbon salımını azaltan dağıtım planlaması

gibi uygulamalar YZ sayesinde daha etkili hâle gelecektir.

Bu sistemlerde verimlilik yalnızca teknik değil, çevresel faydaya odaklı olacaktır. Yapılan çalışmalar, otonom sistemlerin enerji tüketimini optimize edebildiğini ve karbon salımını %20–40 oranında azaltabildiğini göstermektedir (Liu & Xu, 2021).

9.4. Döngüsel Ekonomi ve YZ Entegrasyonu

Döngüsel ekonomi, kaynak kullanımını azaltan ve atık üretimini minimize eden sürdürülebilir bir üretim yaklaşımıdır. YZ, döngüsel ekonomiye şu alanlarda büyük katkı sağlayacaktır:

- Ürün yaşam döngüsü analizinde otomatik veri işleme,
- Atıkların gerçek zamanlı ayrıştırılması,
- Yeniden kullanım ve geri dönüşüm için malzeme sınıflandırması,
- Arıza tahmini ile ürün ömrünün uzatılması,
- Döngüsel tedarik zinciri planlaması.

YZ'nin veri odaklı optimizasyon gücü, döngüsel ekonominin operasyonel süreçlerini güçlendirecek ve daha az kaynakla daha yüksek değer üretilmesini sağlayacaktır (Ellen MacArthur Foundation, 2022).

9.5. 2030 ve 2050 Ufuklarında Yeşil YZ Söylemi

Uzun vadede Yeşil YZ'nin küresel sürdürülebilirlik hedefleriyle doğrudan bağlantılı olacağı öngörülmektedir.

2030 Öngörüleri:

- Veri merkezlerinin büyük kısmının yenilenebilir enerji ile çalışması,
- YZ eğitim süreçlerinde karbon izleme zorunluluğu,

- Model eğitimi için enerji verimli donanımların standart hale gelmesi,
- Şehirlerde YZ tabanlı enerji yönetim sistemlerinin yaygınlaşması,
- Karbon-nötr YZ mimarilerinin ortaya çıkması.

2050 Öngörüleri:

- Karbonsuz (zero-carbon) YZ altyapılarının küresel norm olması,
- Nöromorfik ve kuantum tabanlı ultra düşük enerjili YZ sistemlerinin yaygınlaşması,
- Döngüsel ekonomi sistemlerinin tamamen otonom YZ tarafından yönetilmesi,
- Küresel YZ araştırmalarında enerji verimliliğinin temel performans ölçütlerinden biri olması.

10. Sonuç

Yeşil YZ yaklaşımı, günümüzün hızla gelişen dijital ekosisteminde sürdürülebilirlik, enerji verimliliği ve çevresel sorumluluk gereksinimlerinin bir sonucu olarak ortaya çıkmıştır. YZ sistemlerinin giderek artan hesaplama gücü ihtiyacı, veri merkezlerinin enerji tüketimindeki büyüme ve büyük ölçekli modellerin karbon ayak izi, klasik YZ araştırma ve uygulama paradigmalarını yeniden değerlendirmeyi zorunlu kılmaktadır. Bu bağlamda Yeşil YZ, yalnızca teknik bir optimizasyon süreci değil; aynı zamanda etik, politik, toplumsal ve ekonomik boyutları olan çok katmanlı bir dönüşüm alanıdır.

YZ sistemlerinin çevresel etkileri model eğitimi, çıkarım süreçleri, donanım üretimi ve veri merkezi altyapılarının tamamını

içeren geniş bir yelpazeye yayılmaktadır. Bu etkilerin azaltılabilmesi için geliştirilen teknik çözümler arasında enerji verimli mimariler, model sıkıştırma yöntemleri, özdenetimli öğrenme yaklaşımları, uç YZ ve federe öğrenme gibi dağıtık yöntemler öne çıkmaktadır. Bunlar hem enerji tüketimini hem de veri hareketini azaltarak çevresel maliyetleri düşürmektedir.

Yeşil YZ'nin yalnızca teknik düzeyde değil, uygulama alanlarında da sürdürülebilirliğe çok yönlü katkılar sunduğu görülmektedir. İklim modelleme, akıllı enerji yönetimi, sürdürülebilir tarım, su kaynakları yönetimi, atık yönetimi, biyoçeşitlilik izleme ve döngüsel ekonomi gibi alanlarda YZ, çevresel problemlerin çözümünde güçlü bir araç haline gelmiştir. Bu katkılar, gelecekte YZ'nin sürdürülebilir kalkınmanın temel bileşenlerinden biri olacağını göstermektedir.

Ancak Yeşil YZ yolculuğu önemli zorluklar da barındırmaktadır. Enerji–performans ikilemi, veri merkezlerinin sürdürülebilirlik kapasitesi, büyük ölçekli modellerin yarattığı çevresel ve toplumsal eşitsizlikler, düzenleyici çerçevelerin yetersizliği ve yeşil yıkama riskleri bu alandaki kritik tartışma noktalarıdır. Bu zorlukların aşılabilmesi için şeffaf karbon muhasebesi, zorunlu enerji raporlama standartları, sürdürülebilir donanım tasarımları ve etik yönetişim yaklaşımları öncelikli ihtiyaçlar arasındadır.

Geleceğe yönelik değerlendirmeler, Yeşil YZ'nin 2030 ve 2050 yıllarına doğru giderek daha merkezi bir konuma yerleşeceğini göstermektedir. Ultra verimli model mimarilerinin yaygınlaşması, karbon-nötr YZ altyapılarının norm haline gelmesi, döngüsel ekonomi süreçlerinin YZ tarafından desteklenmesi ve otonom sistemlerde enerji odaklı optimizasyon yaklaşımlarının benimsenmesi, bu dönüşümün somut göstergeleri olacaktır. Ayrıca,

uluslararası kuruluşların ve devletlerin sürdürülebilir dijitalleşme politikaları, Yeşil YZ'nin standartlaşmasını hızlandıracaktır.

Sonuç olarak Yeşil YZ, modern YZ araştırma ve uygulamalarının yönünü belirleyen stratejik bir alan olarak ortaya çıkmıştır. YZ sistemlerinin çevresel etkilerinin azaltılması, yalnızca teknoloji geliştiren kuruluşlar için değil; politika yapıcılar, akademisyenler, endüstri uzmanları ve toplumun tüm aktörleri için ortak bir sorumluluktur. Hem yerel hem de küresel düzeyde sürdürülebilir kalkınma hedeflerine ulaşılabilmesi, Yeşil YZ yaklaşımlarının bütüncül biçimde benimsenmesine bağlıdır. Bu nedenle Yeşil YZ, geleceğin yalnızca daha verimli değil, aynı zamanda daha adil, daha şeffaf ve daha sürdürülebilir bir dijital ekosistemi için vazgeçilmez bir yol haritası sunmaktadır.

Kaynakça

Alova, G. (2020). A global analysis of the progress and failure of electric utilities to adapt their portfolios to the energy transition. Nature Energy, 5(11), 920–927.

Andrae, A. (2020). Hypotheses for primary energy use, electricity use and CO_2 emissions of global computing and its shares of the total between 2020 and 2030. Challenges, 11(2), 31.

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv. https://arxiv.org/abs/2007.03051

Belkhir, L., & Elmeligi, A. (2018). Assessing ICT global emissions footprint: Trends to 2040 & recommendations. Journal of Cleaner Production, 177, 448–463.

Birhane, A. (2022). Automating ambiguity: Challenges and pitfalls of artificial intelligence. arXiv preprint arXiv:2206.04179.

Ellen MacArthur Foundation. (2022). AI and the circular economy: Opportunities and implications. EMF.

European Commission. (2019). The European Green Deal. (8/12/2025 tarihinde https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:640:FIN adresinden ulaşılmıştır)

Evans, R., & Gao, J. (2016, July 20). DeepMind AI reduces Google data centre cooling bill. Google DeepMind Blog. (7/12/2025 tarihinde https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill/ adresinden ulaşılmıştır).

Fu, Z., Chen, F., Zhou, S., Li, H., & Jiang, L. (2024). LLMCO2: Advancing Accurate Carbon Footprint Prediction for LLM Inferences. arXiv. https://arxiv.org/html/2410.02950v1

Google (2023). Sustainability Report. Google Inc.

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. arXiv:1506.02626.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1–43.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.

Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

IEA – International Energy Agency. (2022). Data centres and data transmission networks. IEA.

Jacob, B., Kligys, S., Chen, B., et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. CVPR 2018, 2704–2713.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv:2001.04451.

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv:1910.09700.

Liu, X., & Xu, Y. (2021). Energy-efficient optimization for autonomous systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4007–4020.

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Journal of Machine Learning Research, 24, 1–15.

Microsoft (2020). Microsoft sustainability commitments. (10/12/2025 tarihinde https://www.microsoft.com/sustainability adresinden ulaşılmıştır).

Mosavi, A., Ozturk, P., & Chau, K.-W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536.

OECD (2019). OECD principles on artificial intelligence. OECD Publishing.

Patterson, D., Gonzalez, J., Le, Q. V., Liang, C., Munguia, L. M., Rothchild, D., So, D. R., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

Plan Be Eco (2024). AI's carbon footprint – how does the popularity of artificial intelligence affect the climate?. (7/12/2025 tarihinde https://planbe.eco/en/blog/ais-carbon-footprint-how-does-the-popularity-of-artificial-intelligence-affect-the-climate/ adresinden ulaşılmıştır).

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204.

Różycki, R., Solarska, D. A., & Waligóra, G. (2025). Energy-Aware Machine Learning Models—A Review of Recent Techniques and Perspectives. Energies, 18(11), 2810. https://doi.org/10.3390/en18112810

Schmidt, V., Goyal, K., Joshi, A., Feld, B., Conell, L., Laskaris, N., Blank, D., Wilson, J., Friedler, S., Luccioni, S.: CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing (2021).

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. ICML 2019, 6105–6114.

T.C. Sanayi ve Teknoloji Bakanlığı (2021). Ulusal Yapay Zekâ Stratejisi (2021–2025).

UNEP - United Nations Environment Programme (2020). Digital transformation and sustainability. 2020.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. NeurIPS 2017, 5998–6008.

Visual Capitalist (2024). The Carbon Emissions of Training AI Models. (7/12/2025 tarihinde https://www.voronoiapp.com/technology/The-Carbon-Emissions-of-Training-AI-Models-1447 adresinden ulaşılmıştır).

AÇIKLANABİLİR YAPAY ZEKÂ: KURAMSAL TEMELLER VE YAKLAŞIMLAR

OSMAN EROL[3]

1. Giriş

Yapay zekâ (YZ) teknolojileri, son yıllarda sağlık, finans, eğitim, kamu yönetimi ve siber güvenlik gibi birçok alanda karar verme süreçlerinin merkezine yerleşmiştir. Makine öğrenmesi ve özellikle derin öğrenme tabanlı modeller, büyük veri kümeleri üzerinde yüksek doğruluk ve performans sağlayarak insan yetkinliklerini aşan sonuçlar üretebilmektedir. Ancak bu teknolojik ilerleme, beraberinde önemli bir soruyu da gündeme getirmiştir: Yapay zekâ sistemleri yalnızca doğru sonuçlar üretmekle yetinmeli midir, yoksa bu sonuçların nasıl üretildiği de anlaşılabilir olmalı mıdır? Bu soru, günümüzde yapay zekâ araştırmalarının ve uygulamalarının en kritik tartışma alanlarından birini oluşturmaktadır.

Özellikle karmaşık ve çok katmanlı modellerin yaygınlaşmasıyla birlikte, yapay zekâ sistemlerinin karar

[3] Doç.Dr. Burdur Mehmet Akif Ersoy Üniversitesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi, Orcid: 0000-0002-9920-5211

mekanizmaları giderek daha opak hâle gelmiş; bu durum literatürde "kara kutu problemi" olarak adlandırılmıştır. Kara kutu niteliğindeki sistemler, yüksek doğruluk oranlarına rağmen, neden belirli bir çıktıyı ürettiklerini açıklamakta yetersiz kalmaktadır. Bu durum, kullanıcı güveninin zedelenmesine, hatalı kararların fark edilememesine ve etik ya da hukuki sorumlulukların belirsizleşmesine yol açabilmektedir. Özellikle insan hayatını, temel hakları veya toplumsal adaleti doğrudan etkileyen alanlarda, açıklanamayan algoritmik kararlar ciddi riskler barındırmaktadır.

Bu bağlamda, güvenilir yapay zekâ kavramı ön plana çıkmaktadır. Güvenilir yapay zekâ; yalnızca teknik olarak doğru çalışan sistemleri değil, aynı zamanda adil, şeffaf, hesap verebilir, güvenli ve insan merkezli yapıları ifade etmektedir. Uluslararası kuruluşlar ve düzenleyici kurumlar, yapay zekâ sistemlerinin bu nitelikleri taşımasını giderek daha açık biçimde zorunlu hâle getirmektedir. Avrupa Birliği'nin Yapay Zekâ Yasası (AI Act) ve Genel Veri Koruma Tüzüğü (GDPR) gibi düzenlemeler, özellikle yüksek riskli uygulamalarda açıklanabilirliği temel bir gereklilik olarak konumlandırmaktadır.

Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence – XAI), tam da bu noktada ortaya çıkan kritik bir yaklaşım olarak değerlendirilmektedir. XAI, yapay zekâ modellerinin ürettiği kararların insanlar tarafından anlaşılabilir, yorumlanabilir ve denetlenebilir biçimde açıklanmasını amaçlayan yöntemler, teknikler ve tasarım ilkeleri bütünüdür. Açıklanabilirlik; kullanıcı güvenini artırmanın yanı sıra, model hatalarının tespit edilmesini, önyargıların görünür hâle gelmesini ve sistemlerin etik ve hukuki denetime açılmasını mümkün kılmaktadır. Bu yönüyle XAI, teknik bir araç setinin ötesinde, yapay zekâ ile toplum arasındaki ilişkiyi yeniden tanımlayan çok boyutlu bir çerçeve sunmaktadır.

2. Yapay Zekâ ve Güvenilirlik Temelleri

Yapay zekâ sistemleri günümüzün dijital dönüşümünün temel bileşeni hâline gelirken, bu teknolojilerin güvenilirliği hem teknik hem de etik açıdan kritik bir gereklilik olarak öne çıkmaktadır. Özellikle sağlık, finans, eğitim ve kamu yönetimi gibi yüksek etkili alanlarda kullanılan algoritmaların yalnızca doğru sonuç üretmesi yeterli değildir; aynı zamanda şeffaf, izlenebilir, adil ve güvenli şekilde çalışması beklenmektedir. Güvenilirlik, yapay zekânın toplum tarafından kabul edilmesini sağlayan temel unsurlardan biri olmakla birlikte, hatalı kararların, önyargıların veya kötü niyetli manipülasyonların önüne geçebilmek için gerekli bir çerçevedir. Bu bölüm, yapay zekâ sistemlerinin güvenilirlik kavramını, bileşenlerini ve neden günümüz teknolojik ekosisteminde vazgeçilmez bir rol üstlendiğini temellendirerek kitabın sonraki bölümlerine teorik bir zemin sunmayı amaçlamaktadır.

2.1. Kara Kutu Problemi ve Açıklanabilirlik İhtiyacı

Günümüzde kullanılan makine öğrenmesi modelleri, özellikle derin öğrenme tabanlı yapılar, yüksek doğruluk sağlamalarına rağmen genellikle “kara kutu” (black box) olarak tanımlanmaktadır. Bunun nedeni, modelin iç işleyişinin, hangi girdinin hangi karara nasıl etki ettiğinin çoğu zaman insan tarafından açıkça görülememesidir. Modelin milyonlarca parametresi arasındaki karmaşık etkileşimler, çıkan kararların nedenlerini yorumlamayı güçleştirmektedir (Burrell, 2016). Kara kutu problemine ilişkin temel kaygılar şöyle sıralanabilir:

- **Karar Sürecinin Opaklığı:** Kullanıcıların, modelin verdiği kararın gerekçesini anlamaması güven sorununa yol açabilir.

- **Etik ve Hukuki Riskler:** Özellikle sağlık, adalet ve kamu yönetimi gibi alanlarda, açıklanamayan kararlar hatalı sonuçlar doğurabilir ve ciddi etik sorunlara neden olabilir.
- **Önyargı ve Ayrımcılık:** Eğitim verilerinde bulunan toplumsal önyargılar model kararlarına yansıyabilir; bu durum fark edilmezse ayrımcılık gizli biçimde otomatikleşebilir (O'Neil, 2016).
- **Model Güvenilirliğinin Değerlendirilememesi:** Açıklanamayan bir sistemde hata kaynakları, sınırlılıklar ve riskler belirlenemez.

2.2. Güvenilir Yapay Zekâ: Kavramlar ve İlkeler

Güvenilir yapay zekâ, yalnızca doğru sonuçlar üreten sistemleri değil, aynı zamanda güvenli, adil, şeffaf ve hesap verebilir olan sistemleri ifade eder. Avrupa Birliği'nin "Trustworthy AI" çerçevesine göre bir yapay zekâ sisteminin güvenilir kabul edilebilmesi için yedi temel ilkeyi karşılaması gerekmektedir:

İnsan kontrolü,

Teknik sağlamlık,

Gizlilik ve veri yönetişimi,

Şeffaflık,

Çeşitlilik ve ayrımcılık karşıtlığı,

Toplumsal refaha katkı,

Hesap verebilirlik (European Commission, 2019).

Bu ilkeler, YZ sistemlerinin yalnızca teknik performanslarıyla değil, aynı zamanda toplumsal etkilere ve etik değerlere göre de değerlendirilmesi gerektiğini göstermektedir (Tablo 1). Güvenilir YZ yaklaşımı, mühendislik süreçlerine etik

sorumlulukların dahil edilmesini ve risk temelli bir denetim anlayışının geliştirilmesini zorunlu kılmaktadır.

Tablo 1. Güvenilir Yapay Zekâ İlkeleri ve Örnek Uygulamalar

İlke	**Açıklama**	**Örnek**
Şeffaflık	Model kararlarının anlaşılabilir olması	Kredi puanlama sistemi
Adalet	Önyargısız karar üretimi	İşe alım algoritmaları
Teknik sağlamlık	Modelin saldırılara karşı dayanıklılığı	Siber güvenlik tespit sistemleri

2.3. Şeffaflık Kavramı ve Yapay Zekâda Önemi

Şeffaflık, yapay zekâ sisteminin işleyişinin, kullandığı verilerin, karar süreçlerinin ve sınırlılıklarının anlaşılabilir olması anlamına gelir. Şeffaflık yalnızca model mimarisinin teknik düzeyde açıklanması değil, aynı zamanda sistemin kullanıcıya neden belirli bir çıktı sunduğunu kavramsal açıdan ifade edebilmesi anlamına gelir. Araştırmalar, insanların bir YZ sistemine duyduğu güvenin, modelin performansından ziyade modelin ne ölçüde anlaşılabilir ve açıklanabilir olduğuna bağlı olduğunu göstermektedir (Guidotti vd., 2019). Bu nedenle şeffaflık, hem tasarım aşamasında hem de kullanıcı etkileşimi bağlamında stratejik bir gereklilik hâline gelmiştir. Şeffaf olmayan sistemlerde:

- Hatalar fark edilmeyebilir,
- Modelin güncellenmesi zorlaşabilir,
- Regülasyonlara uyum güçleşebilir,
- Kullanıcı kabulü azalabilir.

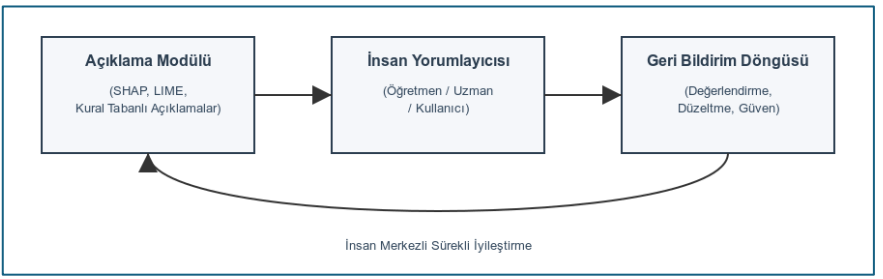
Bu nedenle şeffaflık, güvenilir yapay zekânın temel unsurlarından biri olarak kabul edilmektedir.

3.1. Açıklanabilir Yapay Zekâ (XAI) Nedir?

Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence – XAI), bir yapay zekâ modelinin ürettiği kararların, tahminlerin veya sınıflandırmaların insanlar tarafından anlaşılabilir biçimde açıklanmasını amaçlayan yöntem, teknik ve araçlar bütünüdür. XAI, klasik makine öğrenmesi yaklaşımlarından çok daha fazla parametre içeren karmaşık modellerin giderek yaygınlaşması nedeniyle kritik bir ihtiyaç hâline gelmiştir. Özellikle yüksek kapasiteli derin öğrenme mimarileri, yüksek doğruluk üretmelerine karşın karar süreçlerinin şeffaf olmaması nedeniyle "kara kutu" nitelendirmesiyle eleştirilmektedir (Doshi-Velez & Kim, 2017).

XAI, yalnızca modelin nasıl çalıştığını açıklamakla kalmaz; aynı zamanda modelin güçlü ve zayıf yönlerini anlamayı, hatalı kararların nedenlerini ortaya çıkarmayı ve sistemin güvenilirliğini artırmayı amaçlar. Bu yönüyle XAI, hem teknik hem etik hem de hukuki açıdan yapay zekâ çalışmalarının geleceğinde temel bir bileşen hâline gelmiştir.

Şekil 1. Açıklanabilir Yapay Zekâ Kavramsal Diyagramı



Modelde, yapay zekâ sisteminin ürettiği kararlar açıklama modülü aracılığıyla insan yorumlayıcısına sunulmakta; kullanıcı geri bildirimleri doğrultusunda sistem sürekli olarak iyileştirilmektedir. Bu yapı, şeffaflık, güven ve insan-merkezli yapay zekâ tasarımını desteklemektedir (Şekil 1).

3.2. Açıklanabilirlik Türleri: Küresel ve Yerel Açıklamalar

XAI literatürü açıklanabilirliği genel olarak iki kategoride ele alır: **küresel (global)** ve **yerel (local)** açıklamalar.

3.2.1. Küresel Açıklanabilirlik

Küresel açıklamalar, modelin genel çalışma prensiplerini, değişkenler arasındaki ilişkileri ve karar kurallarının yapısını ortaya koymayı amaçlar. Karar ağaçları, kural tabanlı sistemler gibi şeffaf modeller küresel açıklanabilirliği doğal olarak sağlar. Ancak derin öğrenme gibi karmaşık modeller için küresel açıklama üretmek daha zordur (Molnar, 2022). Örneğin; bir karar ağacında hangi özelliğin hangi karara yol açtığı hiyerarşik biçimde görülebilir.

3.2.2. Yerel Açıklanabilirlik

Yerel açıklamalar, belirli bir örneğe ilişkin model kararının nedenlerini ortaya koyar. Karmaşık modellerde "niçin bu tahmini yaptı?" sorusuna verilen yanıtlar yerel açıklamaya girer. Literatürde SHAP ve LIME gibi yöntemler sıklıkla yerel açıklamalar üretmek için kullanılmaktadır (Ribeiro, Singh & Guestrin, 2016).

Tablo 2. Yerel ve Küresel Açıklanabilirliğin Karşılaştırılması

Açıklama Türü	**Amaç**	**Kullanım Alanı**
Küresel	Modelin genel yapısını anlamak	Model denetimi, politika geliştirme
Yerel	Tek bir kararın nedenlerini açıklamak	Kullanıcı geri bildirimi, hata tespiti

Her iki açıklama türü de farklı ihtiyaçlara hizmet ettiği için XAI sistemlerinin çoğu hem küresel hem yerel açıklama üretebilecek şekilde tasarlanmaktadır (Tablo 2).

3.3. Model-Agnostic ve Model-Specific Yaklaşımlar

Açıklanabilirlik yöntemleri genellikle iki ana gruba ayrılır: model-agnostic (modelden bağımsız) ve model-specific (modele özgü) yaklaşımlar.

3.3.1. Model-Agnostic Yöntemler

Bu yöntemler, modelin iç yapısını bilmeye gerek duymadan, girdiler ve çıktılar üzerinden açıklama üretir. Bu yönüyle herhangi bir makine öğrenmesi modeline uygulanabilirler.

Başlıca model-agnostic yöntemler:

- LIME (yerel lineer açıklamalar)
- SHAP (özellik katkı değerleri)
- Permütasyon önemlilik analizleri
- Kısmi bağımlılık grafikleri (PDP)

Bu yöntemlerin en büyük avantajı esnek olmalarıdır. Bir başka ifadeyle, derin öğrenme modellerinden regresyon modellerine kadar geniş bir alanda kullanılabilirler (Molnar, 2022).

3.3.2. Model-Specific Yöntemler

Bu yöntemler yalnızca belirli model türleri için açıklama üretir. Derin sinir ağları için geliştirilen Grad-CAM veya karar ağaçlarındaki kural görselleştirme teknikleri bu kategoriye örnektir. Başlıca model-specific yöntemler:

- Grad-CAM (görüntü sınıflandırma için)
- Saliency Maps
- Decision Rules Extraction
- Attention görselleştirme (Transformer modelleri için)

Model-specific yöntemlerin avantajı, hedef modele derinlemesine uyarlanmış olmaları nedeniyle çok daha ayrıntılı açıklamalar sunabilmeleridir (Selvaraju vd., 2017).

3.4. Açıklanabilirlik, Güven ve Hesap Verebilirlik İlişkisi

Açıklanabilirlik, güvenilir yapay zekânın temel bileşenlerinden biridir. Kullanıcıların bir YZ sistemine güvenebilmesi için, sistemin neden belirli bir karar ürettiğini makul düzeyde anlaması gerekir. Yapılan araştırmalar, kullanıcı güveninin yalnızca model performansına değil, aynı zamanda modelin açıklanabilirliğine bağlı olduğunu göstermektedir (Suresh & Guttag, 2021). Açıklanabilirlik ayrıca hesap verebilirlik ile doğrudan ilişkilidir. Bir karar mekanizması şeffaf değilse:

- hataların kaynağı belirlenemez,
- sorumluluk atanamaz,
- yasal denetim yapılamaz.

Örneğin; bir yapay zekâ destekli ölçme-değerlendirme sisteminin bir öğrenciyi "başarısız" olarak sınıflandırmasının gerekçeleri açıklanabilir değilse, bu kararın pedagojik geçerliliği sorgulanır. Öğrenci, öğretmen ve kurum açısından hatanın kaynağı belirlenemez; dolayısıyla hem akademik sorumluluk hem de kurumsal hesap verebilirlik zayıflar.

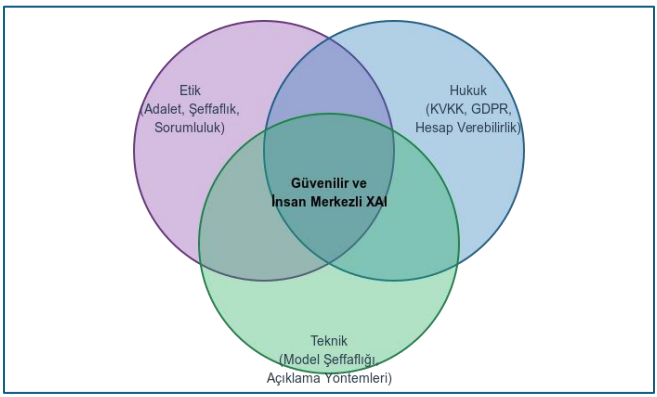
3.5. XAI'ın Yasal, Etik ve Politik Temelleri

Son yıllarda yapay zekâya yönelik regülasyonlar, özellikle açıklanabilirlik bağlamında belirgin bir artış göstermektedir. Avrupa Birliği'nin Yapay Zekâ Yasası (AI Act), yüksek riskli YZ sistemlerinde açıklanabilirliği zorunlu unsurlardan biri olarak tanımlar. Benzer şekilde GDPR, bireylerin kendileriyle ilgili

algoritmik kararların "mantıklı bir açıklamasını" talep hakkı bulunduğunu belirtmektedir (Wachter, Mittelstadt & Floridi, 2017).

Etik açıdan bakıldığında açıklanabilirlik; adalet, şeffaflık, önyargısızlık ve hesap verebilirlik ilkeleriyle doğrudan ilişkilidir. Toplumsal düzeyde ise XAI, insanların teknolojiye duyduğu güveni artırmakta, yüksek riskli uygulamalardaki hataların önüne geçmekte ve karar mekanizmalarını daha demokratik hâle getirmektedir.

Şekil 2. Yasal – Etik – Teknik XAI Kesişimi Diyagramı



Etik ilkeler, hukuki düzenlemeler ve teknik açıklama yöntemlerinin birlikte ele alınması, güvenilir, şeffaf ve insan-merkezli yapay zekâ sistemlerinin geliştirilmesini mümkün kılmaktadır (Şekil 2).

4. Açıklanabilir Yapay Zekâ Teknik ve Yöntemleri

Açıklanabilir Yapay Zekâ (XAI), modern yapay zekâ modellerinin karmaşık yapıları nedeniyle ortaya çıkan belirsizlikleri azaltmak amacıyla geliştirilmiş yöntemler bütünüdür. Bu yöntemler, modelin karar mantığını anlaşılabilir hâle getirerek hem kullanıcı güvenini artırmayı hem de hatalı veya önyargılı kararların tespitini mümkün kılmayı amaçlar.

4.1. Özellik Önemlilik Analizleri

Özellik önemlilik analizleri, modelin çıktısına en fazla etki eden değişkenlerin belirlenmesine yönelik yöntemlerdir. Bu teknikler, modelin hangi girdileri daha kritik gördüğünü ortaya çıkarır ve özellikle yüksek boyutlu veri kümelerinde karar mantığının aydınlatılması açısından oldukça değerlidir.

4.1.1. Permütasyon (Permutation) Önemlilik

Modelden bağımsız (model-agnostic) bir yöntem olan permütasyon önemlilik analizi, bir özelliğin değerleri rastgele karıştırıldığında modelin performansının ne ölçüde düştüğünü inceleyerek o özelliğin önemini belirler (Breiman, 2001). Performanstaki düşüş arttıkça söz konusu özelliğin model için daha kritik olduğu yorumlanır. Avantajları:

- Her model türüne uygulanabilir.
- Basit ve yorumlaması kolaydır.

Sınırlılıkları:

- Özellikle çok korelasyonlu özelliklerde yanlış değerlendirmelere yol açabilir.

4.1.2. SHAP (SHapley Additive exPlanations)

SHAP yöntemi, bir özelliğin modele katkısını *Shapley değerleri* üzerinden hesaplar ve her bir özellik için sayısal, tutarlı ve yorumlanabilir katkı skorları üretir (Lundberg & Lee, 2017). SHAP, açıklanabilirlik alanında şu anda en yaygın kullanılan ve en güçlü yöntemlerden biridir. SHAP'ın güçlü yönleri:

- Hem küresel hem yerel açıklama sağlar.
- Özelliklerin birbirine etkisi istatistiksel olarak modellenir.
- Görsel araçları yüksek yorumlanabilirlik sunar.

4.1.3. LIME (Local Interpretable Model-Agnostic Explanations)

LIME, belirli bir örneğe ilişkin yerel açıklamalar üretir. Bu yöntem, modelin yereldeki davranışını anlamak için o örneğin çevresinde oluşturulan sentetik veri noktaları üzerinde basit bir lineer model kurar (Ribeiro, Singh & Guestrin, 2016). LIME'ın temel özellikleri:

- Modelden bağımsızdır.
- Yerel açıklamalar sunar.
- Özellikle kullanıcı arayüzlerinde anlaşılır açıklamalar için uygundur.

Şekil 3. SHAP ve LIME Yöntemlerinin Kavramsal Karşılaştırılması

SHAP, oyun teorisine dayalı yapısı sayesinde hem küresel hem de yerel açıklamalar sunarken, LIME yerel doğrusal yaklaşımlar aracılığıyla hızlı ve esnek açıklamalar üretmektedir. Her iki yöntem de modelden bağımsız açıklanabilirlik sağlamakla birlikte, hesaplama maliyeti ve tutarlılık açısından farklılıklar göstermektedir (Şekil 3).

4.2. Basitleştirilmiş ve Kural Tabanlı Açıklamalar

Bazı modeller, açıklanabilirliği sistemin içinde doğrudan barındırır. Bu durum özellikle *interpretable-by-design* olarak tasarlanan modellerde görülür.

42.1. Karar Ağaçları ve Açıklanabilirlik

Karar ağaçları, hiyerarşik kurallar üzerinden çalışan şeffaf modellerdir. Her düğümde bir karar kuralı bulunur ve modelin nasıl karar verdiği adım adım izlenebilir (Quinlan, 1993). Örneğin; "Yaş < 30 ise reddet, değilse değerlendir" gibi açıklamalar doğrudan görülebilir. Avantajı:

- Tam açıklama sunar.
- Yorumlanması kolaydır.

Dezavantajı:

- Karmaşık problemler için çok büyük ağaçlara dönüşebilir.

4.2.2. Kural Tabanlı Modeller ve Rule Extraction

Kural tabanlı modeller, karar sürecini mantıksal ifadelerle temsil eder. Ayrıca, kara kutu modellerden kural çıkarmayı amaçlayan yöntemler de literatürde gelişmiştir (Andrews, Diederich & Tickle, 1995).

Tablo 3. Kural Tabanlı Açıklama Örneği

Koşul	**Sonuç**
Gelir < 20.000 ve Yaş < 25	Kredi reddi
Gelir > 50.000	Kredi kabulü

4.3. Görsel Açıklama Teknikleri

Görsel açıklama teknikleri özellikle görüntü işleme modellerinde (CNN modelleri) modelin hangi bölgelere odaklanarak karar verdiğini göstermeyi amaçlar.

4.3.1. CAM (Class Activation Mapping)

CNN modellerinin belirli sınıfları tanımlarken hangi görüntü bölgelerini kullandığını gösterir. Modelin yanlış davranışlarını tespit etmek için oldukça yararlıdır.

4.3.2. Grad-CAM (Gradient-Weighted Class Activation Mapping)

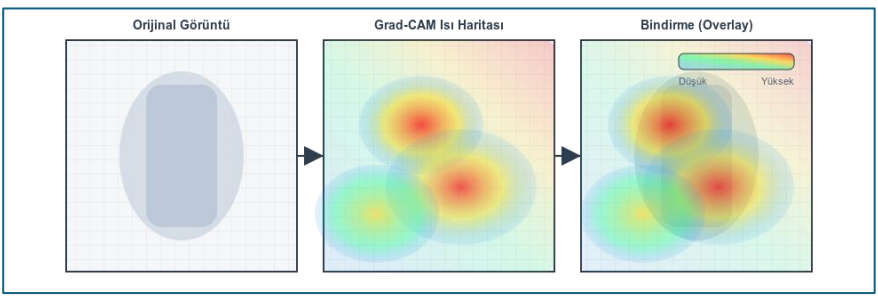
Grad-CAM, gradyan bilgilerini kullanarak görüntünün hangi kısımlarının sınıflandırma sonucuna en fazla katkı sağladığını harita biçiminde gösterir (Selvaraju vd., 2017). Örnek kullanım alanı:

- X-ray görüntülerinde modelin gerçekten hastalıklı bölgeye mi odaklandığını değerlendirmek.

4.3.3. Saliency Maps

Bu yöntem, girdideki her pikselin çıktı üzerindeki duyarlılığını görselleştirir. Piksel bazlı bir önemlilik haritası sunar.

Şekil 4. Grad-CAM Uygulama Örneği — Görüntü Üzerinde Isı Haritası



Grad-CAM örneği: orijinal görüntüden elde edilen sınıf-odaklı gradyan bilgisi kullanılarak ısı haritası üretilmiş ve modelin kararına en yüksek katkıyı sağlayan bölgeler görüntü üzerinde bindirme (overlay) ile görselleştirilmiştir. Kırmızı/sıcak alanlar

yüksek katkıyı, mavi/soğuk alanlar düşük katkıyı temsil etmektedir (Şekil 4).

4.4. Doğallaştırılmış (Textual) Açıklamalar

Son yıllarda açıklanabilirlik alanında metin tabanlı açıklamaların önemi giderek artmaktadır. Büyük Dil Modelleri (LLM) sayesinde modellerin kararlarını doğal dilde ifade eden sistemler geliştirilebilmiştir (Hendricks vd., 2016). Örneğin; "Model bu sınıflandırmayı yaptı çünkü görüntüde belirgin kenar hatları ve yüksek kontrastlı bölgeler bulunmaktadır. Bu tür açıklamalar:

- Kullanıcı dostudur,
- Teknik olmayan kişilere daha anlaşılabilir bilgi sunar,
- Eğitim ve kamu hizmeti gibi alanlarda idealdir.

4.5. Nedensellik (Causality) Temelli Açıklamalar

Klasik XAI yöntemleri korelasyon temellidir; ancak bazı uygulamalarda, özellikle sağlık ve sosyal bilimlerde *nedensellik* temelli açıklamalar gereklidir. Pearl'in (2009) önerdiği nedensel grafik modeller, bir özelliğin yalnızca modelin çıktısıyla ilişkisini değil, aynı zamanda nedensel etkisini ortaya koyar. Nedensel açıklamaların avantajları:

- "Bu karar neden verildi?" sorusunu gerçekten yanıtlar.
- Müdahale senaryoları (counterfactual) üretilebilir.
- Adalet analizlerinde güçlüdür.

4.6. XAI İçin Değerlendirme Ölçütleri

Açıklanabilirlik yöntemlerinin başarısını ölçmek oldukça zor bir problemdir. Literatürde farklı değerlendirme ölçütleri geliştirilmiştir. Bu ölçütler:

- **Sadakat (Fidelity):** Açıklamanın gerçek modeli ne ölçüde doğru temsil ettiğini ölçer.
- **Tutarlılık (Consistency):** Benzer girdilere benzer açıklamalar üretilip üretilmediğini değerlendirir.
- **Anlaşılabilirlik (Interpretability):** Kullanıcıların açıklamayı ne ölçüde anlayabildiğini değerlendirir. Hem objektif hem subjektif ölçütler içerir.
- **Kararlılık (Stability):** Model küçük bir değişiklik yaptığında açıklamanın nasıl değiştiğini inceler.
- **Kullanıcı Güveni Ölçümleri:** İnsan deneklerle yapılan çalışmalarda açıklamanın güven artırıcı etkisi ölçülür (Poursabzi-Sangdeh vd., 2021).

5. Derin Öğrenmede Açıklanabilirlik

Derin öğrenme (Deep Learning), günümüzde yapay zekâ alanının en güçlü paradigmalarından biridir. Büyük veri setleri ve yüksek kapasiteli modeller sayesinde görüntü işleme, doğal dil işleme, ses analizi ve biyomedikal tanı gibi birçok alanda insan performansını aşan başarılar elde edilmiştir. Ancak derin öğrenme modelleri yüksek performanslarına rağmen, açıklanabilirlik açısından ciddi sınırlılıklar taşır. Bu nedenle, bu bölümde derin öğrenme modellerinin neden açıklanmaya ihtiyaç duyduğu, ne tür açıklanabilirlik tekniklerinin geliştirildiği ve mevcut zorluklar ele alınmaktadır.

5.1. Derin Öğrenme Modellerinin Açıklanmasının Zorlukları

Derin öğrenme modelleri, çok katmanlı yapıları ve milyonlarca parametre içermeleri nedeniyle oldukça karmaşıktır. Bu karmaşıklık, modelin girdiler ile çıktılar arasındaki ilişkileri nasıl

yapılandırdığı konusunda önemli bir belirsizlik yaratır. Derin öğrenmenin açıklanmasını güçleştiren temel faktörler şunlardır:

- **Yüksek Parametre Sayısı:** Derin sinir ağları (Deep Neural Networks – DNN), genellikle milyonlarca ağırlık ve parametre içerir. Bu parametrelerin her birinin çıktıya etkisini yorumlamak neredeyse imkânsızdır (Goodfellow, Bengio & Courville, 2016).
- **Karar Mekanizmasının Dağıtık Yapısı:** DNN'lerde bilgi belirli bir düğümde değil, ağ boyunca dağılmış şekilde temsil edilir. Bu da modelin belirli bir kararı neden verdiğini izlemeyi zorlaştırır.
- **Özellik Öğreniminin Soyut Düzeyi:** Erken katmanlar düşük seviyeli özellikler (kenarlar, dokular) öğrenirken, ileri katmanlar soyut kavramlara (nesne, duygu, dil yapısı) karşılık gelir. Bu soyutluk, insan tarafından anlaşılabilir açıklama üretmeyi zorlaştırır.
- **Hiperparametre Hassasiyeti:** Derin öğrenme modelleri, optimizasyon sürecinde hiperparametrelere oldukça duyarlıdır. Farklı hiperparametre seçimi farklı açıklamalar doğurabilir.

5.2. CNN Modellerinde Açıklanabilirlik

Evrimsel sinir ağları (Convolutional Neural Networks – CNN), özellikle görüntü işleme alanında yaygın olarak kullanılmaktadır. CNN'lerin iç yapısı, belirli filtrelerin görüntünün hangi kısımlarını yakaladığını anlamaya imkân tanıdığı için açıklanabilirlik burada görece daha güçlüdür. Ancak yine de derinlik arttıkça yorumlama zorlaşır.

5.2.1. Aktivasyon Haritaları (Feature Maps)

CNN katmanlarında üretilen aktivasyon haritaları, modelin hangi özelliklere duyarlı olduğunu gösterir. Örneğin belirli bir filtre kenarları yakalarken, başka bir filtre doku farklarını yakalayabilir.

5.2.2. CAM ve Grad-CAM

Class Activation Mapping (CAM) ve Grad-CAM teknikleri, modelin görüntü üzerinde hangi bölgelere odaklandığını ısı haritası şeklinde gösterir (Selvaraju vd., 2017). Bu teknikler özellikle tıbbi görüntü analizi gibi kritik alanlarda hatalı odaklanmaları tespit etmek için kullanılır. Örneğin; bir nömoni tespit modelinin akciğer grafisinde anormal olmayan bölgelere odaklanması modelin güvenilir olmadığını gösterebilir.

5.2.3. Occlusion Sensitivity

Girdinin belirli bölümleri kapatılarak modelin çıktısının nasıl değiştiği incelenir (Zeiler & Fergus, 2014). Bu yöntem, görüntünün hangi kısımlarının kritik olduğunu gösterir.

5.3. RNN ve NLP Modellerinde Açıklanabilirlik

Doğal dil işleme alanında kullanılan RNN, LSTM ve GRU gibi modeller, sıralı veriyi işlemeleri nedeniyle farklı açıklanabilirlik tekniklerine ihtiyaç duyar.

5.3.1. Attention Mekanizmaları

Attention mekanizması, modelin hangi kelimelere daha fazla ağırlık verdiğini göstererek doğal açıklamalar üretir (Bahdanau, Cho & Bengio, 2015). Transformer mimarisinin başarısının temel nedenlerinden biridir.

5.3.2. Kelime Duyarlılık Analizi

Girdideki belirli bir kelimenin kaldırılması veya değiştirilmesiyle model çıktısının ne kadar değiştiği analiz edilir. Örneğin; bir duygu analizi modelinde "harika" kelimesi çıkarıldığında sınıf değişiyorsa, bu kelimenin etkisi yüksektir.

5.3.3. Dil Modellemede Temsillerin Görselleştirilmesi

Embedding vektörlerinin t-SNE ile iki boyutta gösterilmesi, kelimeler arası anlamsal benzerlikleri açıklanabilir hâle getirir.

5.4. Transformer Modellerinde Açıklanabilirlik

Transformer mimarisi, NLP'de devrim yaratarak BERT, GPT ve diğer büyük dil modellerinin (LLM) temel yapı taşı hâline gelmiştir. Transformer'ların açıklanabilirliği özellikle önemlidir çünkü karar mekanizmaları oldukça karmaşık ve soyuttur.

5.4.1. Attention Head Analizi

Her bir attention head farklı bir fonksiyon öğrenir:

- Sentaktik ilişkiler
- Uzun mesafe bağımlılıkları
- Anlamsal bağlam

5.4.2. Attention ≠ Explanation Tartışması

Bazı araştırmacılar, attention ağırlıklarının açıklama olarak yorumlanmasının hatalı olabileceğini belirtmiştir (Jain & Wallace, 2019). Bu nedenle transformer açıklamalarının dikkatli değerlendirilmesi gerekir.

5.4.3. Gradient-Based NLP Açıklamaları

Integrated Gradients gibi yöntemler, dil modellerinde girdilerin çıktı üzerindeki etkilerini hesaplamak için kullanılmaktadır (Sundararajan vd., 2017).

5.5. Büyük Dil Modellerinde (LLM) Açıklanabilirlik

GPT, Claude, PaLM ve diğer LLM'ler, milyarlarca parametreye sahip oldukları için açıklanabilirlik daha karmaşık bir problem hâline gelir.

5.5.1. LLM'lerde Şeffaflık Sınırlılıkları

Parametre sayısının devasa olması, eğitildikleri verilerin genellikle gizli tutulması ve model içi temsillerin soyut ve çok boyutlu olması LLM' lerin açıklanabilirliğini zorlaştıran unsurlardır.

5.5.2. Zincirleme Düşünme Açıklamaları (Chain-of-Thought Explanations)

LLM'lerin adım adım akıl yürütme açıklamaları üretmesi, insan benzeri açıklamalar sağlar (Wei vd., 2022). Ancak bu açıklamalar modelin gerçek karar mantığını yansıtmayabilir.

6. Güvenilir Yapay Zekâ Mimarileri

Güvenilir yapay zekâ (Trustworthy Artificial Intelligence), yalnızca performansı yüksek değil, aynı zamanda güvenli, adil, şeffaf, hesap verebilir ve dayanıklı sistemler tasarlamayı hedefleyen bir yaklaşımdır. Yüksek etkili uygulamalarda—örneğin sağlık, finans, eğitim ve adalet sistemlerinde—yapay zekânın oluşturduğu kararların güvenilir olması, teknik yetkinlik kadar etik ve sosyal sorumluluk açısından da zorunludur. Bu bölümde, güvenilir YZ sistemlerinin nasıl tasarlanması gerektiği, kullanılan temel ilkeler, mimari yaklaşımlar ve teknik stratejiler ayrıntılı biçimde ele alınmaktadır.

6.1. Güvenilirlik Ölçütleri ve Standartlar

Güvenilir YZ, belirli teknik ve etik kriterleri karşılayan sistemler olarak tanımlanır. Uluslararası kuruluşlar (EU, OECD, UNESCO) güvenilir YZ için çerçeveler geliştirmiştir. Bu çerçevelerin çoğu ortak bazı ilkelere vurgu yapar:

- Güvenlik ve dayanıklılık (robustness)
- Adalet ve önyargısızlık (fairness)
- Şeffaflık ve açıklanabilirlik (transparency & explainability)
- Hesap verebilirlik (accountability)
- Gizlilik ve veri yönetişimi (privacy & governance)
- İnsan merkezlilik (human-centricity)

6.2. Adil ve Önyargısız Yapay Zekâ (Fair AI)

Yapay zekâ sistemleri, eğitildikleri verilerdeki tarihsel örüntüleri öğrendikleri için, veri setinde bulunan önyargılar model kararlarına yansıyabilir. Bu durum, özellikle işe alım, kredi değerlendirme, adalet sistemi, sağlık hizmetleri gibi kritik alanlarda ayrımcılık riskini artırır (Mehrabi vd., 2021). Önyargı Türleri şunlardır:

- Veri önyargısı (data bias)
- Ölçüm önyargısı (measurement bias)
- Algoritmik önyargı (algorithmic bias)
- Toplumsal önyargıların otomatikleşmesi (societal bias)

Adaleti sağlama yaklaşımları bağlamında adil YZ tasarımı üç aşamada ele alınabilir:

- **Pre-processing yaklaşımları:** Veri setindeki önyargıları azaltma (reweighing, debiasing).
- **In-processing yaklaşımları:** Öğrenme sürecine adalet kısıtları ekleme (fairness constraints).
- **Post-processing yaklaşımları:** Model çıktısını dengeleme (equalized odds, demographic parity).

Tablo 4. Adalet Sağlama Yaklaşımlarının Karşılaştırılması

Yaklaşım Türü	Örnek Teknikler	Avantaj	Dezavantaj
Pre-processing	Reweighing	Veri düzeyinde çözüm	Karmaşık veri gerektirir
In-processing	Adalet kısıtlamaları	Etkili	Performans düşebilir
Post-processing	Equalized odds	Esnek	Model değişmez (sınırlı etki)

6.3. Model Denetimi, İzlenebilirlik ve Hesap Verebilirlik

Güvenilir yapay zekâ mimarilerinde model denetimi kritik bir bileşendir. Denetim;

- modelin nasıl çalıştığının izlenmesini,
- performansının düzenli ölçülmesini,
- risklerin ve hataların erken tespit edilmesini sağlar.

6.3.1. İzlenebilirlik (Traceability)

İzlenebilirlik, modelin hangi verilerle eğitildiği, hangi sürümlerinin kullanıldığı, hangi girdilere hangi kararları verdiği gibi bilgilerin kayıt altına alınmasıdır. Bir başka deyişle modelin "denetim izi" kayıt altına alınmalıdır (Mittelstadt vd., 2016).

6.3.2. Hesap Verebilirlik (Accountability)

Hesap verebilirlik, model kararlarının sorumluluğunun kimde olduğunu ve denetim süreçlerinin nasıl işleyeceğini belirler. Örneğin bir sağlık tanı sisteminde hatalı kararın sorumluluğu sadece modelde değil, sistem tasarımcısı ve kullanıcıda da olabilir.

6.4. İnsan Merkezli Yapay Zekâ (Human-in-the-Loop AI)

Güvenilir YZ yaklaşımı, insan ile yapay zekânın birlikte değerlendirilmesini zorunlu kılar. İnsan-merkezli sistemlerde model kararları tamamen otomatikleştirilmez, kritik noktalarda insan denetimi devreye girer. İnsan denetiminin önemi:

- Kullanıcı güvenini artırır.
- Hatalı sonuçların etkisini azaltır.
- Etik kararların değerlendirilmesini sağlar.

Örneğin sağlık alanında kullanılan yapay zekâ destekli teşhis sistemlerinde son karar her zaman uzman doktora bırakılmalıdır.

6.5. Güvenilir YZ için Mimarilerin Birleştirilmesi: Bütüncül Yaklaşım

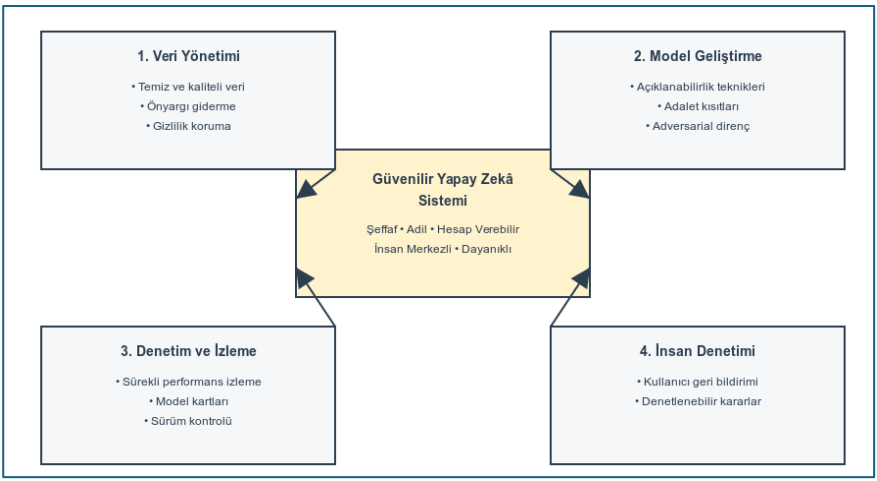
Modern yaklaşımlar, güvenilir yapay zekânın tek bir yöntemle değil, bütünleşik bir mimari ile sağlanabileceğini göstermektedir. Bütüncül bir güvenilir YZ mimarisi şunları içerir:

- **Veri yönetimi**: Temiz veri, Önyargı giderme, gizlilik koruma
- **Model geliştirme:** Açıklanabilirlik teknikleri, Adalet kısıtları, Adversarial direnç
- **Denetim ve izleme:** Sürekli performans izleme, Model kartları, Sürüm kontrolü

- **İnsan denetimi:** Kullanıcı geri bildirimi, Denetlenebilir karar mekanizmaları

Bu yapı, yüksek riskli uygulamalarda güvenilirliğin sürdürülebilir biçimde sağlanmasına imkân tanır.

Şekil 5. Bütüncül Güvenilir Yapay Zekâ Mimarisi



Güvenilir YZ sistemleri; veri yönetimi, model geliştirme, denetim ve izleme ile insan denetiminin birlikte ele alındığı bütünleşik bir mimari yaklaşım gerektirmektedir. Bu bileşenlerin eşgüdümlü çalışması, şeffaf, adil, dayanıklı ve hesap verebilir yapay zekâ sistemlerinin geliştirilmesini mümkün kılmaktadır (Şekil 5).

7. Açıklanabilir Yapay Zekâ Araçları ve Yazılımlar

Açıklanabilir yapay zekâ (XAI) araştırmalarının hızla gelişmesiyle birlikte, hem akademik hem de endüstriyel kullanım için çeşitli yazılım araçları ve kütüphaneler geliştirilmiştir. Bu araçlar, modellerin daha şeffaf, anlaşılabilir ve denetlenebilir hâle gelmesini sağlamanın yanı sıra mühendislerin, araştırmacıların ve veri bilimcilerin açıklamaları sistematik bir biçimde üretmesine olanak tanır.

7.1. Python Tabanlı XAI Kütüphaneleri

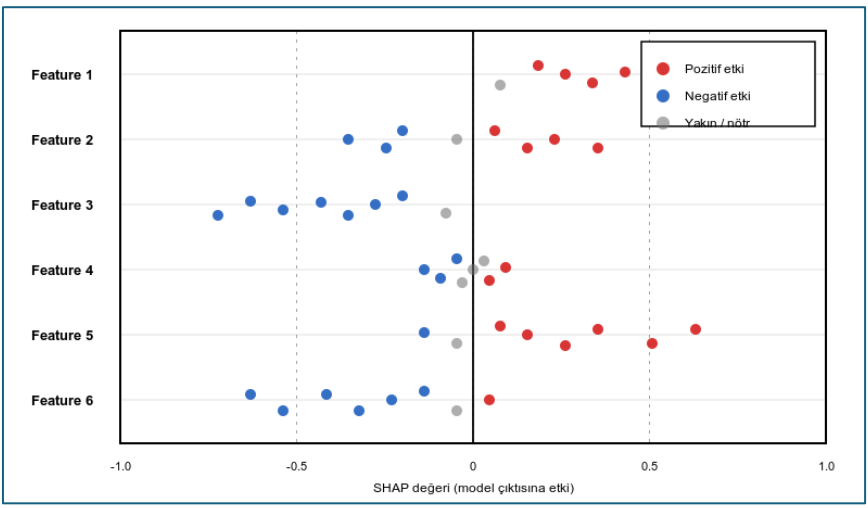
Python, veri bilimi ve makine öğrenmesi alanında en yaygın kullanılan programlama dili olduğu için XAI kütüphanelerinin büyük çoğunluğu Python ekosistemi içinde geliştirilmiştir.

7.1.1. SHAP (SHapley Additive exPlanations)

SHAP, bireysel özelliklerin karar üzerindeki katkısını hesaplayan ve hem küresel hem yerel açıklamalar sunan güçlü bir kütüphanedir (Lundberg & Lee, 2017). SHAP; decision trees, linear models, ensemble yöntemleri ve derin öğrenme modelleri ile uyumludur. Önemli SHAP araçları:

- **summary_plot:** Özellik önemliğini toplu olarak gösterir.
- **force_plot:** Tek bir örneğin açıklamasını görselleştirir.
- **dependence_plot:** Belirli özelliklerin etkileşimlerini ortaya koyar.

Şekil 6. SHAP Summary Plot Örneği



Yatay eksen, her bir özelliğin model çıktısına katkısını gösteren SHAP değerlerini temsil etmektedir (Şekil 6). Pozitif yöndeki katkılar kırmızı, negatif yöndeki katkılar mavi ile gösterilmiştir; noktaların dağılımı özellik etkilerinin örnekler arasındaki değişkenliğini yansıtmaktadır.

7.1.2. LIME (Local Interpretable Model-Agnostic Explanations)

LIME, tek bir örneğe ilişkin yerel açıklamalar üretmek için basit lineer modeller kullanır (Ribeiro, Singh & Guestrin, 2016). Karmaşık modelleri kullanıcıya açıklarken oldukça pratiktir. LIME'ın kullanım alanları:

- Metin sınıflandırma
- Görüntü sınıflandırma
- Tablosal veri modellemeleri

Tablo 5. LIME ve SHAP'ın Karşılaştırılması

Yöntem	Avantaj	Dezavantaj
SHAP	Tutarlı, teorik olarak güçlü	Hesaplama maliyeti yüksek
LIME	Hızlı ve sezgisel	Sonuçlar örneğe göre değişken olabilir

7.1.3. ELI5 (Explain Like I'm 5)

ELI5, özellikle scikit-learn modelleri için sade ve anlaşılır açıklamalar üretir. Özellik önemlilikleri, karar ağaçları ve lineer modellerin açıklamaları kolayca görselleştirilebilir.

7.1.4. InterpretML

Microsoft tarafından geliştirilen InterpretML, hem açıklanabilir modeller (Explainable Boosting Machines - EBM) hem

de kara kutu modeller için açıklama araçları sunar (Nori vd., 2019). Özellikleri:

- EBM modelleri yüksek doğruluk ve yüksek açıklanabilirlik sağlar.
- Model karşılaştırma araçları bulunur.

7.1.5. Captum (PyTorch İçin XAI Araçları)

Facebook (Meta) tarafından geliştirilen Captum, PyTorch tabanlı modeller için açıklama üretir. Derin öğrenme modelleriyle çalışan araştırmacılar için en zengin XAI araçlarından biridir. Başlıca yöntemler:

- Integrated Gradients
- DeepLift
- Gradient Shap
- Occlusion

7.2. Endüstriyel XAI Araçları ve Platformları

Endüstride geliştirilen XAI araçları, yalnızca teknik bir açıklama sunmakla kalmaz; aynı zamanda regülasyon uyumluluğu, raporlama, kullanıcı denetimi ve operasyonel izleme için kapsamlı çözümler içerir.

7.2.1. IBM Watson OpenScale

OpenScale, modellerin performansını izleyen, önyargı tespiti yapan ve gerçek zamanlı açıklamalar sunan gelişmiş bir platformdur (IBM, 2022). Özellikler:

- AI fairness monitoring
- Drift detection

- Real-time explainability

7.2.2. Google What-If Tool

Google'ın açık kaynaklı bu aracı, görsel arayüz üzerinden model davranışlarını incelemeyi sağlar. Yetenekleri:

- Veri manipülasyonu
- Karşı-olgusal senaryolar
- Sınıflandırma karar sınırları

7.2.3. H2O Driverless AI

H2O.ai'nin ticari platformu, otomatik makine öğrenimi (AutoML) süreçleri içinde açıklanabilirlik özelliklerini barındırır. XAI bileşenleri:

- SHAP açıklamaları
- Model dokümantasyonu
- Feature drift tespiti

7.2.4. Fiddler AI

Fiddler AI, kurumsal seviyede açıklanabilirlik ve model izleme platformudur. Sunduğu özellikler:

- Model performans izleme
- Veri kayması (drift) analizi
- Önyargı tespiti
- Karar açıklamaları

8. Sektörel Açıklanabilir Yapay Zekâ (XAI) Uygulamaları

Açıklanabilir Yapay Zekâ (XAI), yalnızca akademik bir araştırma alanı değil, aynı zamanda farklı sektörlerde operasyonel

verimliliği ve güvenilirliği artıran kritik bir bileşen hâline gelmiştir. Özellikle yüksek risk içeren alanlarda—sağlık, eğitim, finans, hukuk, kamu yönetimi ve siber güvenlik—kararların açıklanabilir olması hem etik açıdan hem de yasal gereklilikler açısından zorunlu hâle gelmiştir. Bu bölüm, XAI'ın çeşitli sektörlerdeki kullanım biçimlerini, fırsatlarını ve sınırlılıklarını ele almaktadır.

8.1. Sağlık Alanında Açıklanabilir Yapay Zekâ

Sağlık sektörü, XAI'ın en kritik uygulama alanlarından biridir. Yapay zekâ destekli tanı sistemleri, görüntü analizi, hastalık tahmini ve klinik karar destek sistemleri günümüzde yaygın olarak kullanılmaktadır. Ancak sağlık alanında yapılan her kararın insan hayatını etkiliyor olması, modellerin açıklanabilir olmasını zorunlu kılar (Tonekaboni, Joshi, McCradden & Goldenberg, 2019).

8.1.1. Tıbbi Görüntüleme

CNN tabanlı modeller birçok görüntüleme görevinde insan uzmanlarla yarışır performans göstermektedir. XAI burada üç amaçla kullanılır:

- Modelin doğru yere bakıp bakmadığını kontrol etmek: Grad-CAM ile akciğer grafilerinde modelin patolojik bölgeleri takip edip etmediği anlaşılabilir.
- Hataların açıklanması: "Bu görüntü neden yanlış sınıflandırıldı?" sorusu, saliency map ve occlusion teknikleriyle yanıtlanabilir.
- Klinisyen güvenini artırmak: Klinik uzmanlar, model kararlarını bağımsız doğrulama imkânı bulur.

8.1.2. Klinik Karar Destek Sistemleri

Klinik kararlarda XAI sayesinde:

- risk skorlarının hangi faktörlerden etkilendiği,
- hangi tedavi önerisinin neden üretildiği,
- tahmin edilen hastalık riskinin hangi değişkenlerle ilişkili olduğu açıklanabilir.

Örneğin; bir hastanın sepsise girme olasılığı tahmin edilirken en kritik değişkenlerin (ateş, solunum hızı, lökosit seviyesi) SHAP ile belirlenmesi.

8.1.3. Sınırlılıklar

- Tıbbi verilerin heterojen yapısı açıklamayı zorlaştırır.
- Fazla açıklama klinisyeni yanıltabilir
- Hatalı açıklama güveni azaltır.

8.2. Eğitim Alanında Açıklanabilir Yapay Zekâ

Eğitimde yapay zekâ sistemleri, öğrenci performans tahmini, öğrenme analitiği, kişiselleştirilmiş eğitim ve ölçme değerlendirme süreçlerinde kullanılmaktadır. Eğitim sürecinde şeffaflık özellikle önemlidir çünkü YZ sistemleri öğrenci notlarını, yönlendirmelerini ve eğitim çıktılarını etkileyebilmektedir (Holmes, Bialik & Fadel, 2019).

8.2.1. Öğrenme Analitiği

Öğrencilerin başarı riskini tahmin eden modellerde XAI şu avantajları sağlar:

- Öğrencinin neden risk grubunda olduğu açıklanır.
- Öğretmenler müdahale için doğru stratejiyi seçebilir.
- Eğitim eşitliği desteklenir.

Örneğin; SHAP değerleriyle bir öğrencinin matematik başarısının düşmesinde en etkili faktörlerin belirlenmesi (çalışma süresi, devamsızlık, önceki sınav puanları vb.).

8.2.2. Öğretmen Destek Sistemleri

XAI, öğretmenlere otomatik önerilerin arka planını göstererek sistemin pedagojik açıdan daha kabul edilebilir olmasını sağlar.

8.2.3. Ölçme ve Değerlendirme Sistemleri

Otomatik açık uçlu soru değerlendirme modellerinde açıklanabilirlik:

- verilen notun gerekçesini,
- hangi kelime ve ifadelerin etkili olduğunu gösterir.

Tablo 6. Eğitimde XAI Kullanım Alanları ve Sağladığı Faydalar

Kullanım Alanı	**XAI'nin Sağladığı Açıklanabilirlik**	**Eğitsel Fayda / Sonuç**
Öğrenme Analitiği (başarı ve risk tahmini)	• Öğrencinin neden *risk grubunda* yer aldığı açıklanır. • Özellik katkıları (örn. çalışma süresi, devamsızlık, önceki sınav puanları) görünür hâle getirilir. • SHAP gibi yöntemlerle bireysel öğrenci düzeyinde karar gerekçesi sunulur.	• Öğretmenlerin doğru müdahale stratejisini seçmesi kolaylaşır. • Erken uyarı ve destek mekanizmaları güçlenir. • Eğitimde eşitlik ve adalet desteklenir.
Öğretmen Destek Sistemleri	• Otomatik önerilerin hangi verilere dayanarak üretildiği açıklanır. • Önerileri etkileyen faktörler öğretmene açık biçimde sunulur. • Karar destek süreci pedagojik gerekçelerle temellendirilir.	• Öğretmenlerin sisteme duyduğu güven artar. • Kararların sınıf ve öğrenci bağlamına uyarlanması kolaylaşır. • Hesap verebilir ve bilinçli öğretim kararları desteklenir.
Ölçme ve Değerlendirme Sistemleri	• Verilen notun gerekçesi açık biçimde gösterilir. • Hangi kelime, ifade veya cevap unsurlarının puanı etkilediği açıklanır. • Otomatik değerlendirme süreçleri izlenebilir hâle gelir.	• Adil ve şeffaf değerlendirme algısı güçlenir. • Öğrencilere sunulan geri bildirimin kalitesi artar. • İtiraz ve denetim süreçleri kolaylaşır.

8.3. Finans ve Bankacılıkta XAI

Finans sektörü, yapay zekâ kullanımının en yaygın olduğu alanlardan biridir. Kredi skorlamadan dolandırıcılık tespitine kadar birçok kritik karar YZ sistemleriyle yapılmaktadır. Bu nedenle regülasyonlar finans sistemlerinde şeffaflığı zorunlu tutmaktadır (Rudin & Radin, 2019).

8.3.1. Kredi Skorlama

Kredi başvurularının reddedilmesi durumunda yasal olarak bireylere karar gerekçesi sunulmalıdır. XAI bu süreci destekler:

- SHAP değerleri ile kararın hangi faktörlerden etkilendiği gösterilir.
- LIME ile başvuru sahibine anlaşılır açıklamalar oluşturulur.
- Önyargı analizi yapılabilir.

8.3.2. Dolandırıcılık Tespiti

Açıklanabilirlik, durumun neden şüpheli görüldüğünü göstererek soruşturmaların hızlanmasını sağlar.

8.3.3. Risk Yönetimi

Model dokümantasyonu, denetim izleri ve XAI raporları bankacılık düzenleyicilerinin (BDDK, ECB) zorunlu kıldığı süreçlerdir.

8.4. Hukuk ve Kamu Yönetiminde XAI

Kamu yönetiminde karar destek sistemlerinin yaygınlaşması, açıklanabilirliği zorunlu hâle getirmiştir. Özellikle adalet sistemi gibi yüksek etkili alanlarda, algoritmaların neden belirli bir karara yöneldiğini açıklamak gerekir (Wachter, Mittelstadt & Floridi, 2017).

8.4.1. Ceza Adaleti Sistemleri

Yapay zekâ sistemleri bazı ülkelerde; suç tekrar etme olasılığı (recidivism) tahmini ve risk değerlendirme skorları üretmektedir. Bu modellerde şeffaflık eksikliği büyük tartışmalara yol açmıştır. COMPAS vakasında olduğu gibi, açıklanabilirlik sağlanamadığında algoritmik ayrımcılık tartışmaları büyüyebilir.

8.4.2. Kamu Hizmetlerinde Önceliklendirme

XAI, kamu kaynaklarının daha adil dağıtılması için kullanılabilir. Örneğin; sosyal yardım başvurularının puanlanmasında açıklanabilirlik, ayrımcılığı azaltabilir.

8.5. Siber Güvenlikte Açıklanabilir Yapay Zekâ

Siber güvenlik uygulamalarında anomali tespiti, saldırı tespiti (IDS/IPS), kötü amaçlı yazılım analizi ve tehdit sınıflandırma gibi alanlarda yapay zekâ yoğun şekilde kullanılmaktadır. Ancak YZ kararlarının açıklanamaması, kritik güvenlik açıklarına yol açabilir (Gül, 2022).

5.5.1. Anomali Tespit Sistemleri

Modelin neden belirli bir trafiği anormal olarak işaretlediğini açıklamak; yanlış alarmları azaltır ve gerçek tehditlere odaklanmayı kolaylaştırır.

8.5.2. Kötü Amaçlı Yazılım Tespiti

XAI, bir dosyanın neden zararlı kabul edildiğini açıklayarak analiz uzmanlarının karar vermesini kolaylaştırır.

8.5.3. Olay Müdahale Süreçleri

Açıklanabilirlik, güvenlik ekiplerinin saldırı vektörünü daha hızlı anlamasını sağlar.

8.6. Endüstri 4.0 ve Akıllı Sistemlerde XAI

Fabrikalarda kullanılan otomasyon sistemleri, robotik süreç otomasyonu (RPA), kalite kontrol ve kestirimci bakım uygulamalarında XAI önemli rol oynar.

8.6.1. Otonom Sistemler

Drone'lar ve otonom araçlar için açıklanabilirlik, güvenlik açısından zorunludur. Örneğin; otonom aracın neden fren yaptığı, hangi nesneye tepki verdiği açıklanabilir olmalıdır.

8.6.2. Kalite Kontrol

Görüntü tabanlı kalite kontrol sistemlerinde Grad-CAM ile hatalı bölgeler belirgin şekilde gösterilebilir.

9. Açıklanabilir Yapay Zekânın Etik, Hukuki ve Toplumsal Boyutu

Açıklanabilir yapay zekâ (Explainable AI – XAI), yalnızca teknik bir iyileştirme alanı değil; aynı zamanda etik normlar, hukuki düzenlemeler ve toplumsal beklentilerle şekillenen çok boyutlu bir kavramdır. Bir yapay zekâ sisteminin neden belirli bir karara ulaştığının açıklanabilir olması, kullanıcı güveninden hukuki hesap verebilirliğe kadar geniş bir çerçevede kritik öneme sahiptir. Bu bölümde XAI'ın etik, hukuki ve toplumsal etkileri sistematik bir biçimde ele alınmaktadır.

9.1. Etik Boyut: Algoritmik Adalet ve Sorumluluk

Etik ilkeler, yapay zekâ sistemlerinin güvenilir ve insan merkezli olmasını sağlayan temel kılavuzlardır. Son yıllarda yaşanan birçok olay, algoritmaların açıklanamaz olduğunda ayrımcılığa, haksızlığa ve toplumsal zararlara yol açabileceğini göstermiştir (O'Neil, 2016).

9.1.1. Algoritmik Önyargı ve Ayrımcılık

Veri setlerinde yer alan tarihsel eşitsizlikler, algoritmalar tarafından öğrenilerek yeniden üretilebilir. Bu durum özellikle şu alanlarda kritik riskler oluşturur:

- İşe alım sistemleri
- Kredi değerlendirme
- Sağlık hizmetleri
- Ceza adaleti
- Eğitimde başarı tahminleri

Örneğin, COMPAS risk değerlendirme sistemi Afro-Amerikan bireylere karşı sistematik önyargı taşıdığı için eleştirilmiştir (Angwin vd., 2016).

9.1.2. Etik Sorumluluk ve Hesap Verebilirlik

Bir YZ sisteminin zarara yol açması durumunda sorumluluğun kimde olduğu karmaşık bir sorudur.

- Yazılımcı mı?
- Modeli kullanan uzman mı?
- Modeli yöneten kurum mu?

XAI, karar sürecinin izlenebilirliğini sağlayarak hesap verebilirlik tartışmalarını destekler.

9.1.3. Etik İlkeler: Adalet, Fayda, Şeffaflık

Etik çerçeveler genellikle şu ilkelere dayanır (Floridi & Cowls, 2019):

- **Adalet (fairness):** Ayrımcılık yaratmayan sistemler.
- **Fayda (beneficence):** Zararı önleme, toplum yararını gözetme.
- **Şeffaflık (transparency):** Kararların anlaşılabilir olması.
- **Hesap verebilirlik:** Hataların izlenebilir olması.

9.2. Hukuki Boyut: Regülasyonlar, Haklar ve Yükümlülükler

Yapay zekânın yaygınlaşması, birçok ülkede hukuki çerçevelerin güncellenmesini gerektirmiştir. Açıklanabilirlik, birçok yasal düzenlemede merkezi bir gereklilik hâline gelmiştir.

9.2.1. GDPR ve "Açıklama Hakkı" Tartışması

Avrupa Birliği'nin Genel Veri Koruma Tüzüğü (GDPR), otomatik karar verme süreçlerinde bireylerin belirli haklara sahip olduğunu belirtir. Sık tartışılan Madde 22, bireyin "yalnızca otomatik işleme dayalı bir karara tabi tutulmama hakkı"nı tanımlar. Wachter, Mittelstadt ve Floridi (2017), GDPR'ın doğrudan bir "açıklama hakkı" getirmediğini, ancak kararların mantıklı bir açıklamasının pratikte zorunlu olduğunu savunur.

Tablo 7. GDPR'da Açıklanabilirliği Etkileyen Temel Maddeler

Madde	İçerik	XAI ile İlişkisi
13–15	Bilgilendirme yükümlülüğü	Karar sürecinin açıklanması
22	Otomatik karar verme	Açıklanabilirliğin önemi
35	Etki değerlendirmesi (DPIA)	Risk analizi

9.3. Toplumsal Boyut: Güven, Kullanıcı Kabulü ve Dijital Adalet

Yapay zekâ sistemlerinin toplumsal etkisi geniştir ve açıklanabilirlik bu etkiyi yöneten en önemli araçlardan biridir.

9.3.1. Kullanıcı Güveni

Kullanıcıların bir yapay zekâya güvenmesi, çoğu zaman model doğruluğundan çok, modelin neden belirli bir sonuca ulaştığını anlamasına bağlıdır (Suresh & Guttag, 2021). Örneğin; bir öğrencinin notunu tahmin eden sistem, neden düşük not verdiğini açıklamazsa öğretmen ve öğrenci sisteme güvenmez.

9.3.2. Dijital Adalet ve Eşitlik

Yapay zekâ tarafından verilen hatalı ya da önyargılı bir kararın sonuçları, dijital eşitsizlikleri artırabilir. XAI, yanlış kararları erken tespit ederek dijital adaleti destekler.

9.4. XAI'ın Etik ve Hukuki Sınırlılıkları

Her ne kadar XAI birçok fayda sunsa da, etik ve hukuki açıdan çeşitli sınırlılıklar bulunmaktadır.

9.4.1. Açıklamaların Gerçekliği Yansıtıp Yansıtmaması

Bazı açıklamalar, modeli gerçekten açıklamaktan ziyade insanı ikna etmek için "ikame açıklama" niteliğinde olabilir (Lipton, 2018). Bu durum etik açıdan sorun yaratır.

9.4.2. Gizlilik Riskleri

Açıklamalar, istemeden hassas bilgileri açığa çıkarabilir. Örneğin, doktor raporlarını sınıflandıran bir modelin açıklaması, hastaya ait gizli bilgileri açığa çıkarabilir.

9.4.3. Ticari Sırlar ve Şeffaflık Dengesi

Kurumlar, modellerinin tamamen açılmasını istemeyebilir—çünkü bu durum ticari sırları zedeleyebilir.

9.4.4. Sorumluluk Belirsizliği

Hatalı davranışlarda sorumluluğun kimde olduğu tam olarak belirlenemeyebilir.

10. Açıklanabilir Yapay Zekâda Gelecek Yönelimler

Açıklanabilir Yapay Zekâ (XAI), yapay zekâ sistemlerinin giderek daha karmaşık hâle gelmesiyle birlikte yalnızca teknik bir araştırma alanı olmaktan çıkmış; güven, etik, hukuk ve toplumsal kabul açısından temel bir gereklilik hâline gelmiştir. Özellikle derin

öğrenme ve büyük dil modellerinin (LLM) yaygınlaşması, karar süreçlerinin neden ve nasıl oluştuğuna ilişkin soruları daha görünür kılmıştır. Bu bağlamda XAI'ın geleceği, yalnızca "nasıl açıklama yapılacağı" sorusundan çok, "kimin için, hangi bağlamda ve hangi amaçla açıklama yapılacağı" sorularına odaklanmaktadır.

10.1. Doğruluk ve Açıklanabilirlik Arasındaki Gerilim

XAI alanındaki en temel sorunlardan biri, doğruluk ile açıklanabilirlik arasındaki dengedir. Günümüzde yüksek performanslı modeller genellikle daha opak yapılar sunarken, açıklanabilirliği yüksek modeller performans açısından sınırlı kalabilmektedir. Bu durum literatürde "açıklanabilirlik paradoksu" olarak ele alınmakta ve henüz evrensel bir çözüm üretilememiştir. Gelecekteki çalışmaların, bu ikiliği aşan daha dengeli ve bütünleşik yaklaşımlara yönelmesi beklenmektedir.

10.2. Açıklamaların Güvenilirliği ve Anlamlılığı Sorunu

Bir diğer önemli tartışma alanı, üretilen açıklamaların ne ölçüde güvenilir olduğudur. Çoğu XAI yöntemi, model eğitildikten sonra üretilen post-hoc açıklamalara dayanmaktadır. Bu açıklamaların, modelin gerçek karar mekanizmasını mı yoksa yalnızca ikna edici bir anlatıyı mı yansıttığı sorusu, XAI'ın sınırlarını gündeme getirmektedir. Bu nedenle gelecekte açıklamaların doğrulanabilirliği ve tutarlılığı, teknik performans kadar önemli bir ölçüt hâline gelecektir.

10.3. Büyük Dil Modelleri ve Yeni Açıklanabilirlik İhtiyaçları

Büyük dil modellerinin yükselişi, XAI için yeni fırsatlar kadar yeni riskler de doğurmuştur. Milyarlarca parametreye sahip bu modellerin iç işleyişleri büyük ölçüde opaktır. Ürettikleri akıl yürütme açıklamaları, her ne kadar ikna edici görünse de, her zaman gerçek içsel süreci yansıtmayabilir. Bu durum, özellikle eğitim,

sağlık ve hukuk gibi yüksek riskli alanlarda model çıktılarının daha dikkatli değerlendirilmesini zorunlu kılmaktadır.

10.4. Otonom ve Gerçek Zamanlı Sistemlerde Açıklanabilirlik

XAI'ın geleceğinde öne çıkan bir diğer alan, otonom ve gerçek zamanlı sistemlerdir. Otonom araçlar, robotlar ve endüstriyel otomasyon sistemleri çok kısa sürede karar vermek zorundadır. Bu nedenle açıklamaların karar sonrasında değil, karar süreciyle eş zamanlı olarak üretilebilmesi kritik bir gereksinimdir. Ayrıca bu sistemlerin çoklu veri kaynaklarıyla çalışması, açıklamaların da çok modlu bir yapıya sahip olmasını zorunlu kılmaktadır.

10.5. İnsan–Yapay Zekâ Etkileşiminde Açıklanabilirliğin Önemi

İnsan–yapay zekâ işbirliği bağlamında XAI'ın rolü de dönüşmektedir. Geleceğin sistemleri, tüm kullanıcılar için tek tip açıklamalar sunmak yerine, kullanıcının bilgi düzeyi ve ihtiyacına göre uyarlanmış açıklamalar üretecektir. Açıklamanın içeriği kadar, nasıl sunulduğu da kullanıcı güvenini doğrudan etkilemektedir. Bu nedenle açıklanabilirlik, teknik bir özellikten ziyade bir etkileşim ve tasarım problemi olarak ele alınmaktadır.

11. Sonuç ve Öneriler

Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence – XAI), modern yapay zekâ uygulamalarında güven, şeffaflık, adalet ve hesap verebilirliğin sağlanmasında stratejik bir rol üstlenmektedir. Bu bağlamda, XAI'ın teknik temelleri ile etik, hukuki ve toplumsal boyutları bütüncül bir bakış açısıyla ele alınmıştır. Bu kapsamda:

- Kara kutu probleminin kavramsal temelleri açıklanmış,

- Açıklanabilirlik türleri ve XAI yaklaşımları sistematik biçimde sınıflandırılmış,
- Güvenilir yapay zekânın etik ve hukuki arka planı ortaya konmuş,
- Nedensellik temelli açıklamaların yükselişi değerlendirilmiştir.

Uygulamalı düzeyde ise XAI araçları tanıtılmış, farklı derin öğrenme mimarileri için açıklanabilirlik yöntemleri örneklendirilmiş ve çeşitli sektörlerdeki kullanım alanları ele alınmıştır. Ayrıca AB AI Act, GDPR ve ilgili standartlar çerçevesinde, açıklanabilirlik ile hukuki sorumluluk arasındaki ilişki tartışılmıştır. Tüm bunlar ele alındığında aşağıdaki sonuçlara ulaşılmıştır:

- Derin öğrenme ve büyük dil modellerinin artan karmaşıklığı, yapay zekâ sistemlerinin karar süreçlerini giderek daha opak hâle getirmektedir (Goodfellow, Bengio & Courville, 2016). Bu durum, özellikle yüksek riskli alanlarda açıklanabilirliği zorunlu kılmaktadır. XAI, bu bağlamda isteğe bağlı bir özellik değil, güvenilir yapay zekânın temel bileşenlerinden biridir.
- Mevcut XAI yöntemleri, modellerin kararlarını hem lokal hem de küresel düzeyde anlamayı mümkün kılmaktadır. LIME, SHAP ve Grad-CAM gibi teknikler önemli katkılar sunsa da, hiçbir yöntemin tüm açıklama ihtiyaçlarını tek başına karşılamadığı görülmektedir (Ribeiro vd., 2016; Lundberg & Lee, 2017). Bu durum, açıklanabilirliğin bağlama ve amaca göre ele alınması gerektiğini göstermektedir.
- Güvenilir yapay zekâ mimarileri yalnızca teknik doğrulukla sınırlı değildir. Etik ilkeler, adalet mekanizmaları, yasal

düzenlemeler ve insan denetimi, açıklanabilirlikle birlikte değerlendirilmelidir (European Commission, 2019). XAI'ın değeri, özellikle sağlık, finans, eğitim ve siber güvenlik gibi kritik sektörlerde daha görünür hâle gelmektedir.

- Öte yandan mevcut açıklanabilirlik yaklaşımları, büyük dil modelleri ve otonom sistemler karşısında sınırlı kalmaktadır. Bu nedenle gelecekte XAI araştırmalarının; nedensellik temelli açıklamalar, çok modlu yaklaşımlar ve insan–yapay zekâ etkileşimi üzerine yoğunlaşması beklenmektedir.

Sonuç olarak XAI, yapay zekânın toplumsal kabulü, güvenilirliği ve sürdürülebilirliği açısından vazgeçilmez bir bileşen hâline gelmiştir. Gelecekte açıklanabilir yapay zekâ yaklaşımlarının daha insan-merkezli, bağlama duyarlı ve regülasyonlarla uyumlu biçimde gelişmesi, yapay zekâ sistemlerinin sorumlu kullanımını güçlendirecektir.

11.1. Öneriler

Açıklanabilir yapay zekâ gelişiminin sürdürülebilir olması için farklı paydaşların sorumlulukları vardır. Aşağıda, araştırmacılar, uygulayıcılar, öğretmenler, sağlık çalışanları, politika yapıcılar ve teknoloji şirketleri için özel öneriler sunulmuştur.

11.1.1. Araştırmacılar İçin Öneriler

- XAI yöntemlerinin doğruluk ve güvenilirlik ölçütleri geliştirilmelidir.
- Post-hoc açıklamalar yerine doğrudan açıklanabilir modeller tasarlanmalıdır (Rudin, 2019).
- LLM'ler için özel açıklanabilirlik teknikleri geliştirilmelidir.
- Nedensellik temelli açıklama yaklaşımı daha geniş veri ve model türlerine uygulanmalıdır.

11.1.2. Mühendisler ve Veri Bilimciler İçin Öneriler

- Model geliştirme sürecinde açıklanabilirlik başlangıçtan itibaren dahil edilmelidir.
- Model kartları ve veri kartları oluşturularak dokümantasyon güçlendirilmelidir (Mitchell vd., 2019).
- SHAP, LIME ve counterfactual yöntemler karar süreçlerinde standart hâle getirilmelidir.
- Kullanıcı dostu arayüzlerde açıklamaların sunumu iyileştirilmelidir.

11.1.3. Eğitim Sektörü İçin Öneriler

- Öğretmen ve öğrencilerin YZ okuryazarlığı artırılmalıdır.
- Öğrenci değerlendirmelerinde kullanılan YZ sistemleri şeffaf olmalıdır.
- Risk grubundaki öğrencileri belirleyen sistemlerde mutlaka açıklama modülleri yer almalıdır.

11.1.4. Sağlık Sektörü İçin Öneriler

- Tanı sistemleri açıklanabilirlik modülü olmadan klinik uygulamaya sokulmamalıdır.
- Çıktılar klinisyene açık, anlaşılır ve güvenilir şekilde sunulmalıdır.
- Açıklama sonuçları, klinik karar verme sürecine entegre edilmelidir (Tonekaboni vd., 2019).

11.1.5. Finans ve Kamu Politikası İçin Öneriler

- Kredi kararları, sosyal yardım algoritmaları ve risk analizi modellerinde şeffaflık zorunlu hâle getirilmelidir.

- Algoritmik etki değerlendirmeleri (AIA), proje başlangıcında yapılmalıdır.
- Denetim kurumları açıklanabilirliği düzenleyici çerçevelere dâhil etmelidir.

11.1.6. Toplum ve Kullanıcılar İçin Öneriler

- Dijital okuryazarlık artırılmalı, kullanıcılar algoritmik kararları sorgulama hakkını kullanmalıdır.
- Kamu kurumları açıklanabilirlik ilkelerini halka açık biçimde paylaşmalıdır.

Kaynakça

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems, 8*(6), 373–389.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *ICLR*.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.

European Commission. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

Gül, M. (2022). Explainable artificial intelligence applications in cybersecurity: A survey. *Journal of Cybersecurity Studies, 5*(1), 34–56.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016, September). Generating visual explanations. In *European conference on computer vision* (pp. 3-19). Cham: Springer International Publishing.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

IBM. (2022). *Watson OpenScale documentation*. https://www.ibm.com

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *NAACL*, 3543–3556.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mittelstadt, B., Russell, C., & Wachter, S. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv:1909.09223*.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-52).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD Conference*, 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? *Harvard Data Science Review, 1*(2).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from

deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *ICML*, 3319–3328.

Suresh, H., & Guttag, J. (2021). A framework for understanding unintended consequences of machine learning. *Communications of the ACM, 64*(4), 66–73.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical decision support. *Proceedings of Machine Learning Research, 106*, 359–380.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the GDPR. *International Data Privacy Law, 7*(2), 76–99.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *ECCV*, 818–833.