PESA Center for Political, Economic and Social Research PUBLICATIONS

INTERNATIONAL JOURNAL OF ECONOMIC STUDIES

ULUSLARARASI EKONOMİK ARAŞTIRMALAR DERGİSİ

December 2025, Vol:11, Issue:2 Aralık 2025, Cilt:11, Sayı:2
e-ISSN: 2149-8377 p-ISSN: 2528-9942
journal homepage: https://dergipark.org.tr/tr/pub/ead

Görünmez Tehditten Toplumsal Soruna: Yapay Zekâ Önyargısının Ekonomik ve Toplumsal Etkilerinin Değerlendirilmesi

From Invisible Threat to Social Problem: Assessing the Economic and Societal Impacts of Artificial Intelligence Bias

Deniz TURAN
Doç. Dr., Polis Akademisi Başkanlığı, Güvenlik Bilimleri Enstitüsü, Suç Araştırmaları ABD, ahmetdenizturan@gmail.com
Mehmet UĞURLU
Türk Havacılık ve Uzay Sanayii A.Ş.(TUSAŞ) , mehmet.ugurlu1@tai.com.tr
Ertuğrul KARATAY
Devlet Hava Meydanları İşletmesi, ertugrul.karatay@dhmi.gov.tr

MAKALE BİLGİSİ

Makale Geçmişi:
Geliş 18 Kasım 2025
Dzüeltme 27 Aralık
Kabul 28 Aralık 2025

Anahtar Kelimeler:
Yapay Zekâ, Önyargı, Ekonomik ve Sosyal Etkiler, Önyargı Azaltma

ÖZET

Yapay zekâ tabanlı sistemler günlük hayatımıza sıkı sıkıya yerleşmiştir. Yapay zekâ teknolojileri, ekonomik ve finansal işlemler, sağlık, eğitim, adalet ve kolluk kuvvetleri gibi birçok sektörlerde karar alma süreçlerinde giderek daha fazla kullanılmaktadır. Yapay zekâ sistemlerinin taşıdığı gizli önyargılar ise toplum hayatını olumsuz etkilemekte ve toplumun teknolojiye olan güvenini zedelemektedir. Yapılan çalışma, yapay zekâ sistemlerinin ürettiği önyargıların nedenlerini, ortaya çıkardığı etkileri ve alınabilecek teknik ve sosyopolitik önlemleri açıklığa kavuşturmayı amaçlamaktadır. Teoride ve uygulamada yapay zekâ önyargısının incelendiği çalışmada elde edilen bulgular şunlardır; Yapay zekâ önyargısı, alınan kararlarda toplumsal adaletsizliğe ve ayrımcılığa yol açmaktadır. Yapay zekâ önyargısı ile etkin bir mücadele politikası hem teknik hem de etik hususları içeren bütüncül bir yaklaşım gerektirir. Nitel ve çeşitlendirilmiş veri kümeleri ile verilerinin kalitesi ve temsiliyeti artırılmalıdır. Aynı zamanda yapay zekâ sistemlerini yöneten etik yönergeler ve düzenleyici standartlar oluşturulmalıdır. Standartların belirlenmesinde ve algoritmaların değerlendirilmesinde ulusal ve uluslararası düzeyde disiplinler arası iş birliğine gidilmelidir.

ARTICLE INFO

Article History:
Received 18 November 2025
Correction 27 December 2025
Accepted 28 December 2025

Keywords:
Artificial Intelligence, Bias, Economic and Social Impacts, Bias Mitigation

ABSTRACT

Artificial intelligence-based systems are firmly ingrained in our daily lives. Artificial intelligence technologies are increasingly being used in decision-making processes across numerous sectors, including economic and financial transactions, healthcare, education, justice, and law enforcement. The hidden biases inherent in artificial intelligence systems negatively impact public life and undermine public trust in technology. This study aims to clarify the causes of biases generated by artificial intelligence systems, their effects, and the potential technical and sociopolitical measures to address them. The findings of this study, which examines artificial intelligence bias in theory and practice, are as follows: Artificial intelligence bias leads to social injustice and discrimination in decision-making. An effective policy to combat artificial intelligence bias requires a holistic approach encompassing both technical and ethical considerations. The quality and representativeness of the data should be improved through the use of qualitative and diverse datasets. Ethical guidelines and regulatory standards governing artificial intelligence systems should also be established. Interdisciplinary collaboration should be fostered at the national and international levels to set standards and evaluate algorithms.

GİRİŞ

İnsanlık tarihi kadar eski olan ve nesiller arasında yazı ve dil ile aktarılan önyargılar, adil karar almayı engelleyen önemli bir sosyal olgudur. Büyük veriden beslenen yapay zekâ (YZ) teknolojileri de bu önyargılardan etkilenmektedir. YZ teknolojilerinin karmaşık ve uzmanlık gerektiren yapısı ise alınan kararlarda önyargıların tespit edilmesini zorlaştırmaktadır. Sonuçta insan kararlarına yardımcı olmak üzere tasarlanmış YZ'nin son on yıllarda, finans, adalet, kriminal olaylar, sağlık ve personel seçimi gibi riskli alanlarda karar alma mekanizmalarına dâhil olmaya başladıkları görülmektedir.

Önyargılı algoritmalar tarafından alınan kararlar, kredi başvuruları, mali suçlarla mücadelede, hastalık teşhisi, haksız tutuklama, bireylere veya gruplara karşı ayrımcı ve adil olmayan muameleye neden olmaktadır. Tüm bu olumsuzluklar, görünmez bir tehdit olan YZ önyargısının toplumsal ve hatta uluslararası bir güvenlik tehdidi haline gelmesine yol açmıştır. Bu kapsamda makalede, YZ sistemlerindeki önyargı sorunu, nedenleri ve olası çözüm önerileri üzerinde durulmakta ve disiplinler arası bakış açısıyla kapsamlı bir analitik analiz yapılmaktadır. Literatürdeki diğer çalışmalardan farklı olarak önyargıya yol açan nedenler risk temelinde analiz edilmiş ve disiplinler arası bir yöntem izlenerek hem teknik hem de sosyopolitik çözüm önerileri birlikte ele alınmıştır.

Bu çalışmanın amacı, YZ sistemlerinin yaşam döngüsü boyunca ortaya çıkan önyargıların nedenlerini ve ortaya çıkardığı ekonomik ve toplumsal sorunlar konusunda farkındalık yaratmak ve önyargı ile mücadelede alınması gereken hibrit tedbirleri açığa kavuşturmaktır. Bu kapsamda çalışmanın içeriği aşağıdaki gibi düzenlenmiştir. İlk olarak bu çalışmanın temel kavramları olan önyargı ve YZ tanıtılmış ve devamında YZ sistemlerinde önyargıya yol açan nedenler, YZ yaşam döngüsü çerçevesinde ayrıntılı olarak açıklanmıştır. Sonraki bölümlerde önyargının taşıdığı risklerle bağlantılı olarak ortaya çıkan temel sorunlar irdelenmiştir. Son olarak YZ önyargısını azaltmaya yönelik teknik ve sosyopolitik çözüm önerileri üzerinde durulmuş ve sonuçlar tartışılmıştır.

1. Yapay Zekâ Kavramı ve Yapay Zekâ Önyargısına Yol Açan Faktörler

YZ, genellikle makine öğrenimi, robotik, otonom sistemler ve diğer alt alanları kapsayan, zekâya sahip makineler yaratmayı hedefleyen geniş bir şemsiye terim olarak kullanılmaktadır (Mitchell, 2019: 8). İstatistik, dilbilim ve sinir bilimden beslenen ve veri, algoritmalar ve hesaplamayı içeren YZ sistemleri, ham verileri beslemek için bir algoritma kullanan ve modeller aracılığıyla anlamlı çıktılar üretebilen bir makine öğrenimidir (Heisler ve Grossman, 2024: 1). YZ sistemlerinin temelini, çok katmanlı veri temsiline dayanan ve derin öğrenme adı verilen bir teknik ailesini içeren makine öğrenimi oluşturmaktadır. Makine öğrenimi, insanların hata paylarını kademeli olarak azaltma yöntemlerini taklit etmek için veri ve algoritmaların kullanımını içermektedir. Makine öğrenimi, büyük veriyi toplama, işleme ve depolama teknolojilerinin ve nesnelerin internetinin ortaya çıkmasıyla güçlü bir gelişim ivmesi kazandı. Geçmiş deneyimler ve olgulardan oluşan veriler, makineye gelecekteki sonuçları tahmin etme talimatını vermek için temel olarak kullanılır (Kharitonova ve diğ., 2021: 492). YZ önyargısı ise sosyal ve teknik yönleri içeren çok yönlü bir olgudur. Genellikle bir grubun diğerine göre sistematik olarak kayırılmasına neden olan ve veri toplama, analiz etme, yorumlama ve yaymada yanlış sonuçlara yol açabilecek herhangi bir gerçeklikten sistematik bir sapma, YZ önyargısı olarak sınıflandırılır. Günümüzde YZ, modern teknolojinin önemli bir parçası halini almış ve ekonomi, finans, sağlık, hukuk, eğitim ve kolluk kuvvetleri gibi birçok kritik alanlardaki sektörde karar alma süreçlerini etkilemeye başlamıştır. Örneğin, kredi başvuruları ve işe alım gibi yüksek riskli kararların otomatikleştirilmesi ise mevcut insan önyargısından etkilenen YZ sistemlerinin sistematik ayrımcılığa ve gizli önyargılara yol açtığı görülmektedir

YZ teknolojisinin büyük miktarda veriyi işleme yeteneği, doğru tahminler yapmak ve karar alma süreçlerine yardımcı olmada giderek daha fazla makine öğrenimi sistemlerine güvenilmesine yol açmıştır. YZ sistemleri, sonuçları veya hedef değişkenin değerini tahmin etmek üzere eğitilirler. Ancak bu süreçte ortaya çıkan YZ önyargısı ele alınması karmaşık bir

sorundur. Yalnızca teknik tasarım hatalarını içeren bir mühendislik sorunu olmayıp aynı zamanda veri kaynaklarından, algoritma geliştirme süreçlerinden ve toplumsal-kurumsal yapılardan kaynaklı sosyal bilimler perspektifinden bütüncül şekilde ele alınması gerekmektedir. Bu başlık altında YZ önyargısının nedenleri, YZ'nin yaşam döngüsü temel alarak üç geniş faktör altında ele alınacaktır (Hanna ve diğ., 2025: 3).

- Eğitim verilerinde temsili olmayan verilerin kullanımından kaynaklanan veri kaynaklı önyargılar,
- Algoritma geliştirme sırasında ortaya çıkan geliştirme önyargısı,
- Operasyonel bağlamlarıyla modelle uygunsuz kullanıcı etkileşiminden kaynaklanan etkileşim önyargısı.

1.1. Eğitim Verilerinde Mevcut Önyargılar (Veri kaynaklı önyargılar)

Veri yanlılığı, önyargılı bir veri kümesine neden olan faktörlere odaklanır. YZ sistemlerindeki en yaygın ve temel önyargı kaynağını, eğitim aşamasında makine öğrenimi modelleri geliştirmek için kullanılan eğitim verileri oluşturur. Bu veri kümeleri, istemeden toplumsal önyargıları kodlayarak tarihsel eşitsizlikleri veya veri toplama sürecinde mevcut olan sistemik adaletsizlikleri yansıtabilir. "Giren Çöp, Çıkan Çöp" ilkesi, YZ çıktılarının kalitesinin doğrudan girdi eğitim verilerinin kalitesine bağlı olduğunu vurgular (Hanna ve diğ., 2025: 5). Bir YZ modelini eğitmek için kullanılan veriler önyargılı veya hatalıysa, ortaya çıkan model muhtemelen benzer önyargılar sergileyecektir. Eğitim verileri çeşitli şekillerde önyargılı olabilir.

- ***Bilişsel (Tarihsel) Önyargı****:* Tarihsel önyargı, makine öğrenimine yönelik modellerin eğitiminde kullanılan verilerinin toplumdaki geçmiş önyargıları ve eşitsizlikleri karar süreçlerine yansıtması durumunda ortaya çıkar. Tarihsel önyargılar, dolandırıcılık tespit modellerini, tıbbi teşhis sistemlerini ve ceza adaleti algoritmalarını bozarak geçmiş hataları gelecekteki kararlara dâhil etme riski taşımaktadır. Örneğin, yönetici pozisyonu için bir işe alım algoritmasının eğitiminde kullanılan veriler eğer ağırlıklı olarak erkekleri işe alan bir şirketin verileriyle eğitilirse, algoritma erkek adayları tercih etmeyi öğrenebilir ve bu da cinsiyet önyargısını pekiştirerek sürdürmesine yol açma riski taşımaktadır (González Sendino ve diğ., 2024: 8).

- ***Etiket Yanlılığı:*** Etiket yanlılığı, gözetimli öğrenmede kullanılan etiketlerin hatalı, tutarsız veya doğası gereği yanlı olması ve buna bağlı olarak modelin yanlış etiketleri içselleştirip sürdürmesine neden olması durumunda ortaya çıkar. Bu durum, insan önyargılarından etkilenmesi durumunda ortaya çıkabileceği gibi doğrudan bir insan kararına karşılık gelmeyen, gözlemlenen sonuçlara da karşılık gelebilir. Etiket önyargısının en temel nedenlerinden biri ise tarihsel önyargılardır. Örneğin, Sweeney (2013: 48-49) Afro-Amerikalı gibi algılanan isimlere sahip bireylerin, tutuklama kayıtları olmasa bile, bu tür kayıtlara sahip olduklarını ima eden çevrimiçi reklamlarla karşılaşma olasılıklarının daha yüksek olduğunu belirtmiştir. Bu önyargı, kültürel olarak Afro-Amerikalı veya beyaz bireylere ait olan isimlerle ilişkilendirilmiştir (Li, ve diğ., 2024: 1069). Tıbbi tanılarda hastalık sınıflandırmalarının atanmasındaki önyargılar ise hastalık tahmin modellerini etkilemektedir. Örneğin, biyopsi tümör derecelendirme sonuçlarının subjektif yorumlanması, hastalığın yanlış sınıflandırılmasına yol açmaktadır (Hanna ve diğ., 2025: 4). Bir modeli eğitmek için kullanılan etiketler yanlıysa, algoritma ne kadar gelişmiş olursa olsun, modelin kendisi de yanlı olabilmektedir.

- ***Ölçüm Yanlılığı:*** Ölçüm yanlılığı, toplanan verilerin veya ölçme şeklinin belirli grupları sistematik olarak aşırı veya yetersiz temsil etmesi durumunda ortaya çıkar. Örneğin ABD'de bir sanığın potansiyel tekrar suç işleme olasılığının tahmininde kullanılan Ceza İnfaz Suçlusu Yönetimi Profillemesi (COMPAS) uygulaması, sanığın tekrar suç işleme riskini değerlendirmek için bir algoritma kullanır. Önceki tutuklama geçmişi ile arkadaşların veya aile üyelerinin önceki tutuklamaları da risk değerlendirmesi ve suç davranışı miktarını tahmin etmede kusurlu temsili değişkenler olarak kullanılmıştır. Bu durum, azınlık gruplarının daha

sık denetlenip takip edildiği ve bunun sonucunda daha yüksek tutuklama oranlarına yol açtığı gerçeği karşısında, azınlık gruplarına dâhil kişilerin yalnızca daha yüksek tutuklama oranlarına sahip oldukları için daha tehlikeli oldukları varsayımı önyargıya yol açmaktadır (Dube ve Shafana N, 2021: 229). Başka bir örnek, damar ve kalp hastalık teşhisinde kullanılan kardiyak MRI'nın daha yüksek doğruluğuna rağmen, kadınlarda göğüs ağrısı değerlendirme algoritmaları için tek foton emisyonlu bilgisayar tomografisinin kullanılmasıdır (Mihan ve diğ., 2024: 751).

- ***Örnekleme (Temsil) Önyargısı:*** Temsil önyargısı, bir popülasyonu nasıl tanımladığı ve örneklendiğinden kaynaklanır. Örnekleme önyargısı, eğitim ve test verilerinin modellemek istediği bir nüfus grubunu temsil etmediği veya eksik temsil ettiğinde durumunda ortaya çıkar. Örneğin, ağırlıklı olarak açık tenli bireylerin görüntüleri üzerinde eğitilen yüz sınıflandırma algoritmalarının, eğitim setinde çeşitlilik temsilinin olmaması nedeniyle daha koyu tenli kadınlar söz konusu olduğunda daha düşük doğruluk gösterdiğini ortaya koydu (Ulnicane ve Aden, 2023: 666).

1.2. Algoritma Geliştirme Sırasında Ortaya Çıkan Yanlılıklar (Gelişim Yanlılığı)

Eğitim öncesi aşamada tarafsız ve kaliteli eğitim verileri olsa bile, eğitim aşamasında yani algoritma geliştirme sürecinde yine de önyargılar ortaya çıkabilir. Gelişim yanlılığının başlıca türleri şunlardır:

- ***Algoritmik Yanlılık****:* Algoritmik yanlılık, bir problemin çözümüne yönelik algoritmaların tasarımı ve seçiminden kaynaklanır. Farklı algoritmalar, verilerdeki yanlılıklara karşı farklı düzeylerde duyarlılığa sahiptir. Bazı algoritmalar azınlıklara kıyasla çoğunluğun lehine karar verecek şekilde bir istatistiksel örüntü öğrenir ve model zamanla önyargılı hale gelebilir. Örneğin kredi puanı hesaplamasında geri ödeme olasılığını tahmin etme amacıyla sadece önceki kredi notlarına dikkat edilecek şekilde formüle edilen bir YZ programı, az ya da hiç kredi kullanmayanlara düşük kredi puan verecek ve dezavantajlı duruma düşürecektir (Chadha, 2024: 39).

- ***Model Seçimleri***: Modelleme sürecinde özelliklerin, model mimarisinin ve hiper parametrelerin seçimi gibi hususlarda geliştiricilerin aldıkları kararlar da YZ'ye önyargı katabilir. Modellerin eğitim verilerindeki önyargıları güçlendirmesinin kökeninde genellikle bazı grupların aşırı temsil edilmesi, diğerlerinin ise yetersiz temsil edilmesi yatar. Eğer bir demografik grup verilere hâkimse diğer gruplar için yeterli bilgi olmadığında algoritma bir araya gelemez Verilerdeki kalıpları aşırı veya yetersiz vurgulayan ve alt grup farklılıklarını hesaba katmayan model mimarisi, dezavantajlı gruplar arasındaki mevcut eşitsizlikleri artırıp güçlendirerek bu grupları olumsuz etkileyen önyargılı tahminlerde bulunabilir. Örneğin, bir öngörücü model için seçilen özellikler ırk veya cinsiyet gibi hassas niteliklerle ilişkiliyse, model istemeden bu önyargıları öğrenebilir ve çoğaltabilir (Baeza-Yates ve Murgai, 2024: 448).

- ***Şeffaflık Eksikliğinden (Yorumlanamazlık) Kaynaklanan Önyargı*****:** Derin sinir ağları gibi bazı gelişmiş makine öğrenimi modelleri tahminler yapmada çok başarılı olsa da karmaşık ve opak yapıları nedeniyle insan kullanıcıları için genellikle anlaşılmazdır ve nasıl karar aldıkları bilinmemektedir. Yorumlanamazlık, önyargıların modelde nasıl yayıldığını tespit etmeyi ve anlamayı zorlaştırır. Makine öğrenimi algoritmalarının kullanışlılığı, bunların anlamlandırılması için kavramsal bir çerçeve geliştirilmeden önce benimsenmesine yol açmıştır. Önyargı riskine olan farkındalığın artması, makine öğreniminin bir "kara kutu sorunundan" muzdarip olduğu söylemini güçlendirmiştir (Krishnan, 2020: 489). Kara kutu algoritmaları, yeterli şeffaflık (açıklanabilirlik) sağlamayan algoritmalardır. Karmaşık YZ modellerinin yorumlanamazlığı nedeniyle, güvenilirlik her zaman inceleme altındadır ve YZ modellerinde şeffaflık ve güvenilirliği sağlayabilecek güçlü bir geliştirme ve bakım teknik önlemlerinin alınmasını gerektirir.

- ***Döngüdeki İnsan Önyargısı****:* YZ sistemlerinin geliştirilmesi ve dağıtımında geliştiricilerin verilere veya modele yansıyan kendi önyargıları ve varsayımları, YZ modellerinin tasarımını ve eğitimini etkileyebilir. Eğitim verileri belirli gruplara veya bireylere karşı önyargılar içeriyorsa, YZ bu önyargıları kopyalayabilir ve ayrımcılığı sürdürebilir. Bu

önyargılar genellikle örtüktür ve bir bireyin veya grubun karar vermek veya eksik ya da bilinmeyen bilgileri doldurmak için bilgiyi nasıl algıladığıyla ilgilidir (Gray ve diğ., 2024: 688).

1.3. Etkileşim Yanlılığı

Bu önyargılar, eğitim sonrasında, YZ sistemleri ile operasyonel ortamları arasındaki etkileşimden ortaya çıkabilen, kullanıcı etkileşimleri veya geri bildirimleri aracılığıyla modele dâhil olan önyargılardır. Gelişim yanlılığının başlıca türleri şunlardır:

- ***Dağıtım Yanlılığı:*** YZ yaşam döngüsünde veri hattının dışında ortaya çıkan bir önyargı türüdür. Dağıtım yanlılığı, geliştirme ortamında iyi performans gösteren bir makine öğrenimi modelinin gerçek vakalara uygulandığında başarısız olması veya beklenmedik davranışlar sergilemesi durumudur. Eğitildiği ortam dağıtıldığı koşullarla uyuşmamaktadır. Bu durum, gerçek dünya uygulamalarında istenmeyen ve haksız sonuçlara yol açabilir. Örneğin, kentsel alanlardan alınan verilerle eğitilen YZ sistemi kırsal alanlarda iyi performans göstermeyebilir ve bu da önyargılı sonuçlara yol açabilir. Dağıtım sırasında ortaya çıkabilecek önyargıları belirlemek ve ele almak için sürekli izleme ve geri bildirim sistemleri kurmak gerekmektedir (Chadha, 2024: 39-40).

- ***Geri Bildirim Döngüleri****:* Bu döngüler, bir YZ sistemi tarafından yapılan tahminlerin aynı sistemi güncellemek için kullanılan verileri etkilemesiyle oluşur. Toplu modeller veya çevrimiçi öğrenme modelleri olarak adlandırılan bazı makine öğrenimi modelleri, dağıtımdan sonra da "öğrenmeye" devam eder. Sistem tarafından yapılan tahminler, gelecekteki tahminler için sisteme geri beslenen verileri etkiler. Geri bildirim döngüleri, “kendini gerçekleştiren kehanet” olarak adlandırılan bir durum yaratabilir. Bu durum, zamanla önyargıyı artırabilir ve pekiştirebilir. Örneğin, mevcut suç verilerine dayanarak oluşturulan öngörücü polislik sistemleri, polisin hangi mahallelerin devriye gezilmesi gerektiğini belirlemeye yardımcı olmaktadır. Sistem (A) bölgesinde daha fazla suç tespit ederek oraya daha fazla devriye göndermeye karar verirse, bu bölgede daha fazla suç kaydedilir. Bu veriler sisteme geri beslenerek sistemin (A) bölgesinde daha fazla suç olduğuna olan önyargıyı pekiştirilecektir. Bu durum, bazı mahallelerde aşırı polis denetimi yapılırken diğerlerinde yetersiz polis denetimi yapılmasına yol açabilir (FRA, 2022: 29-30).

- ***Kullanıcı Etkileşimi Önyargısı:*** Kullanıcıların bir YZ sisteminin veri, çıktı veya sonuçlarıyla etkileşim kurarken YZ sisteminin insanlarla önyargılı bir şekilde etkileşime girmesi ve sistemin kendi seçtiği önyargıları ve davranışları dayatmasıyla ortaya çıkar. Örneğin, bir öneri sistemi, sürekli olarak benzer içerikler önererek kullanıcıların mevcut tercihlerini ve önyargılarını güçlendirebilir ve bir yankı odası etkisi yaratabilir.]. Bu tür önyargı, sunum önyargısı (web kullanıcılarının yalnızca gördükleri içeriğe tıklaması) ve sıralama önyargıları (Kullanıcıların bir arama motoru sonuçları listesinde en üst sıradaki sonuçların en alakalı ve önemli olduğuna inanması ve en üstteki sonuca diğerlerinden daha fazla tıklama eğiliminde olması) gibi diğer tür ve alt türlerden etkilenebilir. (Baeza-Yates ve Murgai, 2024: 453).

- ***Veri ve Ekonomik Kaynak Kısıtlamaları:*** Belirli veri türlerinin veya hesaplama kaynaklarının kullanılabilirliği üzerindeki pratik kısıtlamalar ve sınırlamalar da bir çalışmanın istatistiksel gücünü azaltabilir ve önyargılı tahminler üreterek geçersiz sonuçlara yol açabilir. Bütçe kısıtlamaları, veri toplama sürecinde belirli grupların yeterince temsil edilmemesine neden olabilir. Yetersiz verinin modeli daha kolay önyargılı hale gelmektedir. Örneğin, bir model öncelikle Batı ülkelerinden alınan verilerle eğitilmişse, Batı dışındaki bağlamlarda uygulandığında düşük performans gösterebilir (Chadha, 2024: 39).

2. Yapay Zekâ Önyargısının Yol Açtığı Temel Sorunlar

İçine doğduğumuz zaman diliminde karar alma süreçlerinde yoğun şekilde kullanılmaya başlayan YZ teknolojisinin taşıdığı önyargılar kamu düzenini ve güvenliğini olumsuz etkilemektedir. Özellikle kamu yönetiminde YZ'nin hâkim olmaya başlaması ve giderek bürokrasinin yerini algoritmaların alması, “algokrasi” (Algoritma ile yönetim) kavramının

kullanılmasına yol açmıştır. İçerisinde önyargılar barındıran algokrasi anlayışının ise kamu düzenine önemli bir risk oluşturacağı aşikârdır[1].

YZ sistemlerinin yol açtığı ayrımcılık, güven kaybı ve hesap verilemezlik riskleri ekonomik kamu düzenine zarar vermekte, sağlık, ceza adaleti, eğitim ve sosyal refah gibi birçok alanda hem bireysel hem de toplumsal ekonomik maliyetlere ve sorunlara yol açmaktadır. Önyargılı YZ sistemleri, bu teknolojileri kullanan her sektör ve alanda önyargılı kararların oluşma riskini barındırmaktadır. Ancak bu başlık altında YZ önyargısının taşıdığı risklerden hareketle, günümüzde YZ önyargılı kararlarının en yoğun görüldüğü ve toplumun geniş kesimlerini ilgilendiren, kamu düzeni ve güvenliğini önemli ölçüde etkileyen, yüksek riskli karar alma süreçleri barındıran adalet, sağlık ve ekonomik alanlarında ortaya çıkan ve/veya çıkma potansiyeli yüksek sorunlar ele alınacaktır.

2.1. Ekonomik ve Finansal Sorunlar

Finans sektöründe YZ yoğun şekilde kullanılmaktadır ve birçok finansal hizmet günümüzde çevrimiçi olarak sunulmaktadır. Kredi onayları, sigorta ve ipotek işlemleri gibi işlemlerde otomatik algoritmalar tarafından alınan kararlar finansal adaletsizliğe, fırsat kayıplarına ve ayrımcılığa yol açacak önyargılı sonuçlar üretebilir. Kredi başvurularında bulunanların kredibilitesinin değerlendirilmesinde önyargıların mevcudiyeti, düşük kredi puanlarının oluşmasına ve finansal hizmetlere erişiminin kısıtlanması ile sonuçlanabilmektedir. Geçmişteki önyargılı kredi verme modelleri ile eğitilen ve yetersiz verileri kullanan bir kredi algoritması, düşük sosyoekonomik geçmişe sahip kişiler gibi belirli grupların ya da beyaz ve erkek olmayanların orantısız bir şekilde kredi veya ipoteklere erişimini zorlaştırabilir (World Economic Forum, 2023; Ferrara, 2024: 4). Apple cihazında kullanılmak için tasarlanan bir kredi kartı olan Apple Card'ın algoritması, erkeklerle benzer ya da daha yüksek gelir ve kredi puanı olan kadınlara, erkeklere göre daha düşük harcama limiti verdiği için inceleme altına alınmıştır (Upadhyay, 2023).

Büyük hacimli verileri analiz eden YZ sistemleri, karapara aklama, terörün finansmanı ve dolandırıcılık gibi ekonomik ve finansal suç ve suçluların tespitinde yoğun şekilde kullanılmaktadır. Ancak şeffaf olmayan algoritmalara güvenmek, adalet veya hesap verebilirlik konusunda yeni riskler ortaya çıkararak güveni zedeleyebilir. Örneğin dolandırıcılık tespitinde ya da karapara aklama ile mücadelede bir bankanın kullandığı YZ sistemi belli bölgelerde ikamet eden kişi ve kurumları ya da belirli bir demografik kitle haksız yere şüpheli olarak işaretleyen verilerle eğitilmişse, masum kişilerin haksız yere potansiyel suçlu olarak işaretlenmesine ve kırmızı listeye girmelerine neden olacaktır. Marjinal geçmişe sahip bazı bireylerin kredi başvurularının reddedilmesi ise aslında YZ destekli kredi modellerinin kullandığı geçmiş verilerin bu kararı alırken finansal davranışları göz ardı ederek diğer verilere göre karar almasından kaynaklanabilmektedir (Iddenden, 2025). Risk anlayışını çarpıtması, gerçek mali suçların göz ardı edilmesine ve hatalı sonuçlara ulaşılmasına yol açacaktır (Butvinik, 2022). YZ'nin kullanımında müşterileri korumak ve suçla mücadele çabası arasında dengenin kurulması önem taşımaktadır.

2.2. Adalet Sistemde Ortaya Çıkan Sorunlar

Önyargılı YZ modellerinin etkilediği kritik alanlardan biri de mahkemeler, polis, savcılık ve cezaevleri gibi kurumları içeren ceza adalet sistemidir. Veri, gelişim ya da etkileşim yanlılığı sonucu ortaya çıkan önyargılı kararların ve modellerin ceza adaletinde kullanımı, haksız suçlamalara ve tutuklamalara, sınır dışı işlemlerinde haksız kararlara, daha ağır cezalar almaya veya kefaletle serbest bırakıp bırakmama gibi ciddi sonuçlara yol açabilir. Ayrımcılık taşıyan önyargılı algoritmalar, belirli bölgelere ya da etnik gruplara dahil kişilere veya cinsiyet temelinde kadın ve erkekler arasında adil olmayan muameleye yol açabilmektedir (Saha ve diğ., 2023: 1). Stanford Hukuk Fakültesi Göçmen Hakları Kliniği'nin yayınladığı raporda (NIPNLG, 2024: 3), ABD'de Afro-Amerikalılar ve etnik azınlıklara karşı ayrımcılığın ve yaygın bir eşitsizliğin bulunduğu belirtilmiştir. Özellikle küçük çaplı trafik suçları, uyuşturucu ve

[1] Algokrasi kavramı ve özgürlükler üzerine etkisi konusunda ayrıntılı bilgi için bkz: (Danaher, 2020)

mülkiyet suçları hakkındaki ceza yargılamasının her aşamasında önyargılı kararların yaygın olarak ortaya çıkabileceği belirtilmektedir[2].

ABD'de hapis cezaları ve tekrar suç işleme oranları hakkında karar vermek için kullanılan duruşma öncesi ve ceza infaz kurumlarında yaygın olarak kullanılan "Ceza İnfaz Suçlusu Yönetimi Profillemesi" (COMPAS) algoritmasının önyargılı olduğu ortaya konmuştur (Upadhyay, 2023). Model, Afro-Amerikalı suçlular için tekrarlanan suçlarda (%45) beyaz suçlulara (%23) kıyasla iki kat daha fazla yanlış pozitif tahminde bulunmuştur (Bansal ve diğ., 2023: 374). Benzer şekilde, ABD'nin çeşitli eyaletlerindeki polis teşkilatları tarafından 2013 yılından günümüze aktif kullanılan ve öngörücü bir polislik algoritması olan PredPol'ü (Tahmine Dayalı Polislik) Oakland şehrinde kullanılmıştır. Irksal önyargılı tutuklama verilerine dayanan önyargılı yazılımın, Afro-Amerikalıların olduğu mahalleleri beyaz mahallelere kıyasla 2 kat daha fazla hedef alarak oldukça ayrımcı olduğu bulunmuştur (Moya ve Le, 2021: 8).

2.3. Sağlık Hizmetlerinde Ortaya Çıkan Sorunlar

YZ sistemlerinin yoğun olarak kullanıldığı sağlık hizmetleri, YZ kararlarının yaşam standardını kökten etkileme gücüne sahip olduğu riskli bir alandır. YZ, belirli insan grupları için sağlık sorunları riskini tespit etmede ve en etkili tedavi seçeneklerini belirlemede kullanılabilir. Diğer taraftan, YZ destekli tanı ve tedavi süreçlerinin önyargılar barındırması durumunda, önyargılı çıktıların nispeten az temsil edilen belirli gruplar için tedaviye eşitsiz erişime yol açmasının yanı sıra hastalara yanlış ve/veya yetersiz tanı ve tedavi planları yaparak eşitsizlikleri artırma ve zarar verme potansiyeli bulunmaktadır (Chen vd., 2023: 2). Örneğin, açık ten tonları üzerinde eğitilmiş bir cildiye algoritmasının koyu tenli kişilerdeki rahatsızlık tanısı koymasında sorunlar yaşanacaktır. Diğer bir örnekte, Amerikan sağlık şirketi Optum'un, ek klinik değerlendirme için hasta esmer hastalar yerine beyaz hastaları önerdiği tespit edilmiştir. Ek değerlendirmeye çağrılması gereken esmer hastaların oranı %46,5 olmasına karşılık sadece %17,7'si ek değerlendirmeye çağrılmıştır. Başka bir çalışmada ise, kandaki oksijen satürasyonunu ölçen cihazlar olan nabız oksimetrelerinin, beyaz tenli insanlara kıyasla Afro-Amerikalı insanlarda daha az doğru sonuçlar verdiği tespit edilmiştir (Bansal ve diğ., 2023: 375).

Sağlık alanında genellikle geçmiş insan kararlarına dayanan verilerle eğitilen karar destek sistem algoritmaları, geçmişte verilmiş yanlış teşhislerin yani sistematik hataların önyargısını yansıtacaktır. Ancak tıp alanında nesnel ve analitik görevlerde objektif olduğu düşünülen algoritmik tavsiyelere aşırı güvenilmesi, YZ önyargılarının kullanıcılar tarafından sorgulanmasını engelleme riski taşımaktadır. Bu nedenle, sağlık hizmetlerinde kullanılan modellerin önyargılarının azaltılmasında, eğitim verilerindeki dengesizlikleri gidermek önem taşımaktadır (Vicente ve Matute, 2023: 2).

3. Yapay Zekâ Önyargısını Azaltmaya Yönelik Teknik Çözüm Önerileri

YZ sistemlerinde önyargıları azaltarak adil ve etik kararların elde edilmesi için alınabilecek tedbirlerin alınması, bütünsel bir strateji gerektiren, kapsamlı ve çok aşamalı bir süreçtir. YZ önyargısını azaltmaya yönelik çözüm önerileri literatürde birçok farklı yöntemle tasnif edilmesine karşılık bu çalışmada çözüm önerileri hibrit bir yaklaşımla ele alınmış ve mühendislik boyutu ile teknik çözüm önerileri ve sosyopolitik çözüm önerileri olarak iki kısımda irdelenmiştir. YZ'nin yaşam döngüsünde yer alan önyargı türleri dikkate alınarak çözüm önerileri üç ana kategoride ele alınacaktır.

3.1. Eğitim Öncesi Ön İşleme Teknikleri

[2] ABD'de Afro-Amerikalı erkekler, benzer suçlar işleyen Beyaz erkeklere göre ortalama %19,1 daha uzun cezalar almaktadır. Kaliforniya Oakland'da yapılan araştırmaya göre ırksal sınırlar arasında uyuşturucu kullanım oranları genel olarak aynı olmasına rağmen Afro-Amerikan mahallelerinde, diğer Oakland mahallelerine göre 200 kat daha fazla uyuşturucu tutuklaması olmaktadır. Sonuçta, yalnızca tutuklama verileriyle eğitilmiş bir polislik algoritması, Afro-Amerikan mahallelerinin daha fazla uyuşturucu kullandığı sonucuna varabilme riski taşımaktadır (Moya ve Le, 2021: 8).

Eğitim öncesi ön işleme aşaması, eğitim verilerindeki temsili olmayan verilerin ayıklanma ve nötralize etmeye yönelik çözüm tekniklerini içermektedir. "Giren Çöp, Çıkan Çöp" ilkesi gereği, temsili ve adil verilerin kullanılarak, YZ modelleri eğitim almadan önce verilerin ayıklanması önem arz etmektedir. Etkili önyargı azaltma, veri toplama ve işleme aşamasından başlar. Ön işleme tedbirleri ile girdi olarak sağlanan verilerin adil ve eşit olmasını sağlanır. Eğitim öncesi alınabilecek tedbirler maddeler halinde verilmiştir.

- ***Veri Artırma ve Yeniden Örnekleme (Data Augmentation and Re-sampling)***: Önyargı ile mücadeleye, ayrımcılık içermeyen ve nüfusu doğru bir şekilde temsil eden adil bir veri kümesi parametrelerinin tanımlanması ile başlamalıdır. Önyargılı veri kümeleri tarafından üretilen kararlar, etkilenen tarafları ayrımcı zararlardan yeterince korunamayacaktır. Adil temsil öğrenimi sağlandıktan sonra bu parametreleri karşılayan veri kümesinin nitel ve nicel boyutu değiştirilerek kümelerinin çeşitlendirilmesi, zenginleştirilmesi ve hedef kitlenin tamamını temsil edecek şekilde seçilmesi[3] yoluna gidilebilir (De ve diğ., 2023: 108-109). Örneğin, dezavantajlı grupları aşırı örnekleyerek veya baskın grupları yetersiz örnekleyerek veri kümesi dengeli hale getirilmesi, algoritmanın yapacağı tahminleri olumsuz etkileyebilecek önyargıları azaltacaktır.

- ***Yeniden Ağırlıklandırma (Re-weighting)*:** Veri noktalarına temsiliyetlerine göre farklı ağırlıklar atamak, önyargıyı azaltmaya yardımcı olabilir. Farklı demografik özellikleri, coğrafi konumları ve sosyoekonomik geçmişleri olan dezavantajlı gruplardan alınan örneklerin ağırlıklarını artırarak ve baskın gruplardan alınan örneklerin ağırlıkları azaltarak eğitim verilerine yerleşmiş önyargıları düzeltmeyi amaçlar. Az temsil edilen sınıfların aşırı örneklemesi, modelin eğitim aşamasında bunlara daha fazla dikkat etmesine yol açacaktır. Bu tedbir, modelin gruplar arasında ayrımcılık yapmayarak tüm gruplara daha eşit davranmayı öğrenmesini sağlayacaktır (Hanna vd., 2025: 8; Ulnicane ve Aden, 2023: 678).

- ***Adil Temsil Öğrenimi (Fair Representation Learning)*:** Adil temsil, bir kişiyi adil olmayan (ayrımcı) bir gruba bağlayabilecek bilgilerin elenmesiyle elde edilir. Örneğin, belirli bir demografik grubun dezavantajlı veri örnekleminde aşırı temsil edildiği bir veri kümesinde, bu grup için ek sentetik veri noktaları oluşturulabilir veya dengeli bir veri kümesi elde etmek için mevcut veriler yeniden örneklenebilir (Chadha, 2024: 42). Temsil kusurlarını düzeltmede, veriler ile modellenecek temel popülasyon arasındaki uyumu ahlaki ve etik boyutlarıyla değerlendirmede uzmanlardan yararlanmak gerekmektedir. Bu tedbir, sonuçların açıklanabilirliğini karmaşık hale getirebilir ancak adilliği artırır (González Sendino ve diğ., 2024: 10).

3.2. Eğitim Aşamasında İşleme İçi Teknikler

Eğitim, önyargıyla başa çıkmanın en verimli aşamasıdır. Burada kullanılan teknikler verileri olduğu gibi alır ve model eğitim sürecini kendisi ayarlar. Bu aşamada model eğitim sürecine, öğrenme algoritmalarının adalet kısıtlamaları dâhil edilir ve algoritmalar adaleti de artıracak şekilde değiştirilir. Grup adaleti veya bireysel adalete duyarlı öğrenme ve önyargı tespit algoritmaları, model seçimi ve eğitiminden kaynaklanabilecek önyargıları azaltmaya yardımcı olabilir. Böylece eğitim sonrası dağıtım aşamasına geçilmeden modelde ayarlamalar ve iyileştirmeler yapılır ve önyargılar azaltılabilir. Eğitim aşamasında önyargıyı azaltmaya yönelik başlıca tedbirler şunlardır:

- ***Düzenleme Yöntemleri (Regularization)*:** Model eğitiminde bir modelin geçmiş verilere yetersiz veya aşırı uyumu düzeltmek için kullanılan bu teknik, adalet hususlarını tahmin modellerinin eğitim sürecine açıkça entegre ederek önyargıları azaltmak için de kullanılabilir. Bu kapsamda düzenlemeyi uygulamanın bir yolu, model tahminleri ile hassas nitelikler arasındaki ilişkiyi çözen kısıtlamaların örtük olarak eklenmesidir. Diğer bir yöntem

[3] Yeniden örnekleme yöntemleri, alt örnekleme ve üst örnekleme olarak ikiye ayrılır. Üst örnekleme, orijinal veri kümesini zenginleştirmek için veri örneklerinin oluşturulması, alt örnekleme teknikleri ise veri kümesinden çıkarılan örneklere dayanır. Üst örneklemeye örnek olarak demografik eşitliği artırmak için sentetik tablo verileri üretmek amacıyla kullanılan Üretken Çelişkili Ağlar (GAN'lar) ile Sentetik Azınlık Üst Örnekleme Tekniği (SMOTE) verilebilir (González Sendino ve diğ., 2024: 10)

ise, model kayıp fonksiyonunu güncellemektir. Modelinin çıktılarındaki hata derecesini izleyen ve model performansını ölçen kayıp fonksiyonu, farklı korunan gruplar arasındaki performans farkını en aza indirmek güncellenebilir. Örneğin, fırsat eşitliği veya demografik eşitlik gibi kısıtlamalar, modeli önyargılı tahminler nedeniyle cezalandırmak için dahil edilebilir (González Sendino ve diğ., 2024: 10; Chadha, 2024: 42).

- ***Adil Kümeleme (Fair Clustering)*:** Orijinal veri kümesindeki önyargıyı azaltmak amacıyla bir veri kümesini bölmeyi amaçlayan bir tekniktir. Eğitim sürecine entegre edilmiş bu teknik, öğrenilen temsillerin adil olmasını sağlar. Geleneksel kümeleme algoritmaları verilerdeki korunan niteliklerle (örneğin ırk ve cinsiyet) ilişkili olan önyargıyı göz ardı edebilir ve sadece verilerin içyapısını hesaba katar. Adil bir kümeleme ise ırk, ten rengi vb. farklı korunan nitelik değerlerine sahip nesnelerin her kümede tekdüze bir dağılıma sahip olmasını gerektirir. Örneğin, varyasyonel adil otomatik kodlayıcılar (variational fair autoencoders), eğitim sırasında hassas nitelikleri diğer özelliklerden ayırarak adil temsiller öğrenmek için kullanılabilir (Chadha, 2024: 42). Elektronik multimedyada yüz duygusu tanıma için özellik çıkarma örneğinde adil kümelemeye gidilmesi, önyargıları azaltacaktır. Adil kümeleme yöntemleri ile burun, göz ve ağız gibi organlar benzerliğe göre gruplanabilir ve korunan nitelik olarak kabul edilen ırk tespitinde daha adil ve önyargısız sonuçlar elde edilecektir (Pan ve Zhong, 2023: 2-3).

- ***Rakip Eğitim (Adversarial Training)*:** Rakip (mücadeleci) eğitim, modellerin sağlamlığını özünde geliştirmeyi amaçlar. Matematiksel olarak, rekabetçi eğitim, en kötü durum optimumuna en iyi çözümü arayan bir min-maks problemi olarak formüle edilir. YZ alanında rekabetçi eğitim, her eğitim döngüsünde rekabetçi örneklerle eğitim verilerini zenginleştirilmesi ile başlar. İkincil bir model (rakip), birincil modelin tahminlerinden hassas niteliği tahmin etmek üzere eğitilir. Böylece ilk ağı önyargıya karşı daha dayanıklı ve dirençli hale gelmeye zorlar (Bai ve diğ., 2021: 4312). Rakip öğrenme kullanılarak makine öğrenimi modellerinin adaleti, önyargıların azaltılmasıyla iyileştirilebilir. Rakip" olarak da bilinen ikinci ağ, ilk tahminlerdeki zayıflıkları bulmaya çalışır. Korunan niteliklere dair kanıtlar azaltılabilir ve demografik eşitlik artırılabilir (Chadha, 2024: 42).

3.3. Eğitim Sonrası Son İşleme Teknikleri

Son işlem yaklaşımları, önyargıyı azaltmak için orijinal modelin yeniden eğitilmesi yerine modelin yapacağı tahminleri (çıktıları) ayarlar. Bu aşamada, dağıtımdan önce tespit edilen önyargıları azaltmak için modelde ayarlamalar ve iyileştirmeler yapılmasına olanak tanır. Bu kapsamda farklı alt grupların adil temsiline zarar verebilecek eğitilmiş bir modelin sonuçları düzeltilerek değiştirilir. Metin üreten bir dil modelinin, nefret söylemini içeren önyargılı ifadelerin kullanımını engellemek ve filtrelemek amacıyla bir tarayıcı kullanması örnek verilebilir (SAP, 2024). Ancak son işleme tekniklerinin doğruluğu azaltma veya ilk sınıflandırıcı tarafından elde edilen herhangi bir genellemeyi tehlikeye atma riski bulunmaktadır (Dube ve Shafana N, 2021: 231).

- ***Yeniden Sıralama (Re-ranking):*** Çapraz kodlayıcı olarak da bilinen bu teknikte önyargıyı azaltmak ve adaleti sağlamak için modelin tahminleri yeniden sıralanmaktadır. Örneğin, bir sıralama probleminde sonuçlar, listenin en üstünde farklı grupların adil bir şekilde temsil edilmesini sağlamak için ayarlanabilir (Chadha, 2024: 43). Çevrimiçi uygulamalarda kullanıcılara kişiselleştirilmiş öneriler sunmak amacıyla kullanılan öneri sistemleri popülerlik yanlılığından sıklıkla mustariptirler. Popüler ürünlerin daha fazla öneri almasına yol açan önyargıyı aşmak ve her kullanıcının öznel ilgi alanlarına göre kullanıcı odaklı yeniden sıralamaya gidilebilir (Gulsoy ve diğ., 2025).

- ***Eşik Ayarlaması (Threshold Adjustment):*** Modelin karar eşiklerinin farklı gruplar için değiştirmek, önyargıyı azaltmaya yardımcı olabilir. Demografik gruplar arasında yanlış pozitif ve yanlış negatif oranlarının eşit dağılmasını sağlayan ve böylece eşit olasılıklar elde etmek için bir model tarafından alınan kararları ayarlayan eşik ayarlaması yöntemi, demografik eşitlik veya eşit fırsat gibi daha adil sonuçlara ulaşılmasına yol açacaktır (Ferrara, 2024: 6). Örneğin ders geçme oranları veya kredi puanı tahmininde, sıralama algoritmalarının

içine açık adalet kısıtlamaları yerleştirerek farklı bir eşik kullanılması, önyargıların giderilerek adil ve doğru tahmin üretilmesine imkân tanır (Diana, 2025).

- ***Kalibrasyon*:** Farklı gruplar arasında modelin tahmin edilen olasılıkları ile gerçek dünya gözlemleri uyumlu hale getirilmesini ifade eden kalibrasyon, tahmin modellerinde adaleti ve karar güvenilirliğini artırır (Nikoli´c, 2025: 1-2). Kalibre edilmiş eşitlenmiş oranlar, çıktı etiketlerini değiştirmek için en uygun olasılıkları belirlemeyi amaçlar. Bu, tahmin edilen olasılıkların, tüm gruplar arasında sonuçların gerçek olasılığını tutarlı bir şekilde yansıtacak şekilde ayarlanmasını içerir (Chadha, 2024: 43).

4. Yapay Zekâ Önyargısını Azaltmaya Yönelik Sosyopolitik Çözüm Önerileri

YZ önyargısı ile etkili bir mücadelenin ikinci ayağını sosyal ve politik önlemler oluşturmalıdır. Yalnızca teknik yöntemlerle önyargı sorunları ile mücadele etmek istenen sonuçları vermeyecektir. YZ kararlarında önyargıların giderilmesi ve adil sonuçlar üretmesinin sağlanması amacıyla öncelikli olarak YZ döngüsünün her aşamasında etik ilkelerin göz önünde bulundurulmasını gerektirir. Özellikle adalet, sağlık ve finans alanlarında algoritmaları ahlaki standartlar ile uyumlu hale getirmek gerekmektedir. Adil, hesap verilebilir ve şeffaf sonuçların üretilmesi için etik yönergeler ve düzenleyici politika ve prosedürler geliştirerek ulusal ve uluslararası düzeyde standartlaştırılmış ölçütler oluşturulmalıdır. Bu sayede YZ uygulamaları ortak bir dil ve çerçeve oluşacak ve tutarlı adalet karşılaştırmaları yapılabilecektir (Chadha, 2024: 45-46).

YZ sistemlerini oluşturan algoritmaların opak yapıları, alınan kararlara nasıl ve neden ulaştığı konusunda bilinmezliğe yol açmaktadır. Öz-yansıma (iç gözlem) sağlayan makine öğrenimi modelleri geliştirmek, önyargıyı üreten faktörleri belirleyerek karar alma süreçlerini model tarafından gözden geçirmelerini sağlayacaktır. Kullanılan verileri eğitmek üzere erişime açmak, daha fazla şeffaflık ve güvenilirlik sağlayacaktır. Algoritmik hesap verebilirliğe ulaşmak için şeffaflığın sağlanması gerekmektedir (González Sendino ve diğ., 2024: 12).

YZ sistemlerinde ortaya çıkan önyargıların tespiti için düzenli bir gözetim ve denetim mekanizması kurulmalıdır. Önyargı tespit edildiğinde ilgili personeli uyarabilen gerçek zamanlı önyargı denetim araçları oluşturulması, sürekli iyileştirmeleri mümkün kılacaktır. Ancak sadece makine gözetimine güvenilmemeli ve her zaman bir miktar insan gözetimi mevcut olmalıdır. Makine ve insan gözetimi ile sürekli ayarlamalar ve güncellemeler yapılmalıdır. Söz konusu iç izlemenin yanı sıra, bağımsız üçüncü taraflarca dış denetimlere gidilmelidir (Schwartz, ve diğ, 2022: 42-43).

Standartların belirlenmesinde ve algoritmaların değerlendirilmesinde tasarımcıların bakış açıları tek başına yeterli olmayacaktır. Bu nedenle veri bilimciler, sosyal bilimciler, siyaset bilimciler, kalite ve etik uzmanları ile kullanıcı veya dış paydaş gruplarının temsilcilerinden oluşan disiplinler arası hibrit ekiplerin kurulması ve işbirliğine gidilmesi, önyargı ve adaletsizliği azaltmada etkili olacaktır (Silberg ve Manyika, 2019: 6). Diğer taraftan YZ'nin uluslararası düzeyde kullanımının yaygınlaşmış olması, ulusötesi etik endişeleri dikkate almayı gerekli kılmaktadır. Özellikle finansal suçların önlenmesinde ve sağlık sektörlerinde amacıyla uluslararası kuruluşlarla düzenleyici işbirliğinin kurulması gerekmektedir (Orso ve Medeiros, 2023).

SONUÇ

Bu ayrıntılı çalışmada, YZ önyargılarının nedenleri ve ortaya çıkardığı derin toplumsal etkilerinin risklerin yanı sıra önyargıları azaltmaya yönelik çözüm stratejileri üzerinde durulmuştur. İçine doğduğumuz zaman diliminde YZ'nin ulusal ve uluslararası arenada güvenlikten eğitime, sağlıktan ekonomi ve finans dünyasına kadar her alanda karar alma süreçlerinde yoğun ve giderek artan bir şekilde kullanımıyla birlikte YZ sistemlerin neden olduğu önyargılı kararlardan kaynaklanan bireysel ve toplumsal sorunlar artmaya başlamıştır. Özellikle hastalık tanısı, mahkûmiyet kararları, risk değerlendirme, mali suçların tespiti gibi isan hayatını doğrudan etkileyen kritik süreçlerde alınan önyargılı kararlar, toplumsal adaletsizliğin ve ayrımcılığın artmasına, önyargıların yaygınlaşmasına ve gelecekteki

kararların da bu önyargılı kararlardan olumsuz etkilenmesine yol açmaktadır. Artan eşitsizlikler ve insan hakları ihlalleri, yeni bir araştırma alanı olarak YZ önyargısı ve taşıdığı risklerin literatürde daha fazla ele alınması zorunluluğu doğurmuştur.

Görülmektedir ki YZ önyargısı ile mücadele yalnızca mühendislik boyutuyla çözümlenecek bir sorun değildir ve sosyopolitik bakış açısı ile disiplinler arası iş birliğinin teşvik edilmesiyle desteklenmesi gerekmektedir. Bu kapsamda adil ve etkili bir önyargı azaltma politikası hem teknik hem de etik hususları içeren bütüncül bir yaklaşım gerektirir. Gelecekteki yapay zekâ çalışmalarının, YZ önyargısının ekonomik ve toplumsal etkilerini de dikkate almaları ve önyargıyla mücadele sürecinde ilgili paydaşlarla ulusal ve uluslararası düzeyde işbirliğinin gerekliliği her türlü izahtan varestedir.

KAYNAKÇA

Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q. (2021), "Recent Advances in Adversarial Training for Adversarial Robustness", *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, https://www.ijcai.org/proceedings/2021/0591.pdf 05.11.2025.

Bansal, C., Pandey, K. K., Goel, R. ve Sharma, A. (2023), "Srinivas Jangirala Artificial Intelligence (AI) Bias İmpacts: Classification Framework for Effective Mitigation", *Issues in Information Systems*, 24(4): 367-389.

Baeza-Yates, R. ve Murgai, L. (2024), "Bias and the Web", *Introduction to Digital Humanism,* Editors: Hannes Werthner, Carlo Ghezzi, Jeff Kramer, Julian Nida-Rümelin, Bashar Nuseibeh, Erich Prem, Allison Stanger, (435-462), Springer, India.

Butvinik, D., (25.07.2022), "Bias and Fairness of AI-Based Systems within Financial Crime", *NICE Actimize*, https://www.niceactimize.com/blog/fraud-bias-and-fairness-of-ai-based-systems-within-financial-crime/ 05.09.2025.

Chadha, K. S., (2024), "Bias and Fairness in Artificial Intelligence: Methods and Mitigation Strategies", *International Journal for Research Publication and Seminar*, 15(3): 36-49.

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S. ve Malin, B. (2023), "Human-Centered Design to Address Biases in Artificial Intelligence", *Journal of Medical Internet Research*, 25: 1-10.

Danaher, J (2020), "Freedom in an Age of Algocracy", *Oxford Handbook of Philosophy of Technology*, Editör: Shannon Vallor, (250-272), Oxford University Press, Oxford, UK.

De, S., Jangra, S., Agarwal, V., Johnson, J. ve Sastry, N. (2023), "Biases and Ethical Considerations for Machine Learning Pipelines in the Computational Social Sciences", *Ethics in Artificial Intelligence: Bias, Fairness and Beyond,* Editör: Animesh Mukherjee, Juhi Kulshrestha, Abhijnan Chakraborty, Srijan Kumar, (99-113), Springer.

Diana, E., (17.07.2025), "Building AI Fairness by Reducing Algorithmic Bias, Tepperspectives", *Tepper School of Business at Carnegie Mellon University*, https://tepperspectives.cmu.edu/all-articles/building-ai-fairness-by-reducing-algorithmic-bias/ 05.10.2025.

Dube, R. ve Shafana, J. N. (2021). "Bias in Artificial Intelligence and Machine Learning", *Bioscience Biotechnology Research Communications*, Special Issue 14(9):227-234. |

Ferrara, E. (2024), "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies", *Sci*, 6(1):1-15.

FRA, (European Union Agency for Fundamental Rights), (2022), Bias in Algorithms –Artificial Intelligence and Discrimination, Report, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf 05.10.2025.

González Sendino, R., Serrano, E., Bajo, J., ve Novais, P. (2024), "A Review of Bias and Fairness in Artificial Intelligence", *International Journal of Interactive Multimedia and Artificial Intelligence*, 9(1): 5–17.

Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, W. ve Wu, L. (2024), "Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science", Clinical Pharmacology & Therapeutics, 115(4): 687-697.

Gulsoy, M., Yalcin, E., Tacli, Y., Bilge, A., (2025), "DUoR: Dynamic User-oriented re-Ranking Calibration Strategy for Popularity Bias Treatment of Recommendation Algorithms", *International Journal of Human-Computer Studies*, 203:103578.

Hanna, M. G., Pantanowitza, L., Jacksonc, B., Palmera, O., Visweswarane, S., Pantanowitzf, J., Deebajahg, M. ve Rashidi, H. H. (2025), "Ethical and Bias Considerations in Artificial Intelligence/Machine Learning", *Modern Pathology,* 38: 1-13.

Heisler, N. ve Grossman, M. R. (2024), *Standards for the Control of Algorithmic Bias:The Canadian Administrative Context,* CRC Press, USA.

Iddenden, G. (20.03.2025), "Algorithmic Gatekeepers: The Hidden Bias in AI Payments", *The Payments Association*, https://thepaymentsassociation.org/article/algorithmic-gatekeepers-the-hidden-bias-in-ai-payments/ 02.10.2025.

Kharitonova, Y., Savina, V S. ve Pagnini, F. (2021), "Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues", *Perm University Herald. Juridical Sciences,* 53: 488–515.

Krishnan, M. (2020), "Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning", *Philosophy & Technology,* 33:487–502.

Li, Yunyi; Arteaga, M. ve Saar-Tsechansky, M. (2024), "Label Bias: A Pervasive and Invisibilized Problem", *Notices of the American Mathematical Society*, 71(8): 1069-1077.

Mihan, A., Pandey, A. ve Van Spall H. (2024), "Mitigating the Risk of Artificial Intelligence Bias in Cardiovascular Care", *Lancet Digit Health*, 6: 749–754.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans,* Farrar, Straus and Giroux,nUSA.

Moya, G. ve Le, V. (18.02.2021), "Algorithmic Bias: How Automated Decision-Making Becomes Automated Discrimination", *The Greenlining Institute*, https://greenlining.org/wp-content/uploads/2021/04/Greenlining-Institute-Algorithmic-Bias-Explained-Report-Feb-2021.pdf 12.09.2025.

NIPNLG, (National Immigration Project), (2024), *Bias in the Criminal Legal System: A Report on Racial Bias in the Criminal Process and its Impact on Noncitizens of Color in Removal Proceedings*, Immigrants' Rights Clinic, Stanford Law School, USA. https://law.stanford.edu/wp-content/uploads/2024/06/2024-Bias-Criminal-Legal-System.pdf 09.08.2025.

Nikoli´c, M.; Nikoli´c, D.; Stefanovi´c, M.; Koprivica, S.; Stefanovi´c, D. (2025), "Mitigating Algorithmic Bias Through Probability Calibration: A Case Study on Lead Generation Data", *Mathematics*, 13: 2183.

Orso, M. ve Medeiros, A. (22.08.2023), "The Uses and Risks of AI in BSA/AML Compliance: Navigating the Future of Financial Crime Prevention", Troutman Pepper Locke's Financial Services Group, https://www.troutmanfinancialservices.com/2023/08/the-uses-and-risks-of-ai-in-bsa-aml-compliance-navigating-the-future-of-financial-crime-prevention/ 13.09.2025.

Pan, R. ve Zhong, C. (2023), "Fairness First Clustering: A Multi-Stage Approach for Mitigating Bias", Electronics, 12(13): 1-16.

Saha, D., Agarwal, A., Hans, S. ve Haldar, S. (2023), "Testing, Debugging, and Repairing Individual Discrimination in Machine Learning Models", *Ethics in Artificial Intelligence: Bias, Fairness and Beyond, Studies in Computational Intelligence,* Editörler: A. Mukherjee et al. (1-30), Springer.

SAP, (Sistem Analizi Program Geliştirme), (30.10.2024), *What is AI bias?*, Türkiye Yazılım Üretim ve Tic. A.Ş., https://www.sap.com/resources/what-is-ai-bias#emerging-trends-in-fair-ai-development 15.10.2025

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. ve Hall, P. (2022), "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence", Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464 07.10.2025.

Silberg, J., ve Manyika, J. (06.06.2019), "Tackling bias in artificial intelligence (and in humans)", *McKinsey Global institute*. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf 10.01.2025.

Sweeney, L. (2013), "Discrimination in Online Ad Delivery", *Communications of the ACM*,56 (5), 44–54.

Ulnicane, I. ve Aden A. (2023), "Power and Politics in Framing Bias in Artificial Intelligence Policy", *Rev Policy Res,* 40: 665–687.

Upadhyay, S. (27.02.2023), *Algorithmic Bias and its Impact on Society*, https://medium.com/kigumi-group/algorithmic-bias-and-its-impact-on-society-df12edcfb303 08.08.2025.

Vicente, L. ve Matute, H. (2023), "Humans Inherit Artificial Intelligence Biases", *Scientifc Reports*, 13:15737.

World Economic Forum, (30.06.2023), *Why AI Bias may be Easier to Fix Than Humanity's*, https://www.weforum.org/stories/2023/06/why-ai-bias-may-be-easier-to-fix-than-humanity-s/ 15.09.2025