Uluslararası Akademik Birikim Dergisi
Yıl: 2024, Cilt: 7, Sayı. 5 ss.1041-1052

Yapay Zeka Sistemlerine Etik Değerlerin Entegrasyonu-Kavramsal Çerçevede Bir Tartışma

Büşra Fadim SARIKAYA TÜNALP
Dr., Türk-Alman Üniversitesi
busra.sarikaya@tau.edu.tr
https://orcid.org/0000-0002-9492-7493

Makale Başvuru Tarihi : 16.11.2024
Makale Kabul Tarihi : 28.12.2024
Makale Yayın Tarihi : 31.12.2024
Makale Türü: Araştırma Makalesi
DOİ: 10.5281/zenodo.14583506

Anahtar Kelimeler:

Yapay Zekâ, Etik, Etkileşim, İletişim

Özet

Yapay zekâ (YZ), hem tarihsel hem de çağdaş unsurları bir araya getiren bir alan olarak dikkat çekmektedir. Tarihsel açıdan, YZ ile ilgili bilimsel tartışmaların kökenleri 1950'lere kadar uzanmakta olup, felsefi tartışmalar ise antik Yunan dönemine kadar geriye gitmektedir. Çağdaş açıdan ise, YZ alanında, veri işleme teknolojilerindeki büyük ilerlemeler sayesinde, "YZ kışı" olarak adlandırılan duraklama dönemlerinden sonra yeniden önemli bir ivme kazanılmıştır. Bu bağlamda, YZ yalnızca bilimsel bir disiplin olmanın ötesine geçmiş, aynı zamanda siyaset ve iş dünyasında da önemli bir konu haline gelmiştir. YZ'nin kendisi, doğası gereği etik ya da etik dışı bir varlık değildir; ancak, uygulama biçimleri etik sorunlar doğurabilir. Bu nedenle, YZ teknolojilerine duyulan güvenin artırılması, bu teknolojilerin geniş çapta kabul görmesi açısından temel bir gereklilik olarak ortaya çıkmaktadır. YZ'nin etik ilkelerle uyumlu bir şekilde tasarlanması ve uygulanması konusunda, ulusal ve uluslararası düzeyde yoğun tartışmalar sürdürülmektedir. Normlar ve standartlar, uzman gruplarının, fikir birliği ile geliştirdiği için, bir dereceye kadar ahlaki değerlere de dayandırılabilir. Ancak, teknik standardizasyonun amacı, özellikle "etik YZ" bağlamında, kültürel olarak sabitlenmiş belirli değerleri ya da ahlak kurallarını desteklemek değildir. Bu çalışmada öncelikle yapay zekâ ve etik ayrı başlıklarda kavramsal olarak tanımlanmış ve yapay zekâ sistemlerinin kullanımına ilişkin etik boyutlar sunulmuştur. Ardından etik bir yapay zekânın inşa edilmesinin ne kadar mümkün olduğu literatür araştırması yöntemi kapsamı ve Milgram Deneyi örneğiyle incelenmiş olup, bunu mümkün kılmak için gereken etik yönler ele alınmıştır. Çalışmanın amacı ise YZ sistemlerine etik değerlerin entegrasyonunun mümkün olup olmadığını sorgulamaktır. Bu her ne kadar zor gözükse de, sistemlerin şeffaflığı ve kullanıcıların eleştirel yaklaşımı, etik olmayan kararların önlenmesinde kilit rol oynayacağı ve yapay zekâ uygulamalarının, açık kriterlere göre değerlendirilmesi gerektiği bulgusu çalışma kapsamında elde edilmiştir.

The Integration Of Ethical Values Into Artificial Intelligence Systems

Abstract

Keywords:

Artificial Intelligence (AI), Ethics, Interaction, Communication

Artificial intelligence (AI) is a field that integrates both historical and contemporary elements. Historically, scientific discussions surrounding AI can be traced back to the 1950s, while philosophical debates extend as far back as ancient Greece. In contemporary terms, AI has gained significant momentum following periods of stagnation, often referred to as the "AI winter," largely due to substantial advancements in data processing technologies. In this context, AI has evolved beyond a purely scientific discipline, emerging as a critical issue in both politics and business. AI, by its nature, is neither an ethical nor unethical entity; however, its applications can give rise to ethical concerns. Indeed, instances of AI exhibiting bias based on gender or skin color have already been documented. Furthermore, there is a growing societal fear of becoming overly reliant on the judgments of an omniscient AI system. Therefore, enhancing trust in AI technologies is emerging as a fundamental requirement for their broader acceptance. Intense debates are ongoing at both national and international levels regarding the design and implementation of AI in alignment with ethical principles. Norms and standards, developed through consensus by expert groups, can incorporate certain moral values to some extent. However, the aim of technical standardization is not to endorse culturally specific values or moral codes, particularly within the context of "ethical AI." In this study, artificial intelligence (AI) and ethics were conceptually defined under separate headings, and the ethical dimensions of AI systems' usage were presented. Subsequently, the feasibility of constructing ethical AI was examined within the scope of a literature review and the example of the Milgram Experiment, focusing on the ethical aspects required to achieve this. The study aimed to investigate whether it is possible to integrate ethical values into AI systems. Although this appears challenging, the findings of the study indicate that the transparency of systems and the critical approach of users play a key role in preventing unethical decisions. Moreover, it was concluded that AI applications should be evaluated based on clear and well-defined criteria.

1. GİRİŞ

Yapay zekâ (YZ), hem tarihsel hem de çağdaş unsurları bir araya getiren bir alan olarak dikkat çekmektedir. Tarihsel açıdan, YZ ile ilgili bilimsel tartışmaların kökenleri 1950'lere kadar uzanmakta olup, felsefi tartışmalar ise antik Yunan dönemine kadar geriye gitmektedir. Çağdaş açıdan ise, YZ alanında, veri işleme teknolojilerindeki büyük ilerlemeler sayesinde, "YZ kışı" olarak adlandırılan duraklama dönemlerinden sonra yeniden önemli bir ivme kazanılmıştır. Bu bağlamda, YZ yalnızca bilimsel bir disiplin olmanın ötesine geçmiş, aynı zamanda siyaset ve iş dünyasında da önemli bir konu haline gelmiştir. YZ'nin mevcut durumu, güçlü, düşünebilen ve duygusal bir YZ vizyonu ile gerçeklik arasında hâlâ büyük bir boşluk olduğunu göstermektedir. Bu boşluğun kapatılıp kapatılamayacağı ise belirsizliğini korumaktadır. Bununla birlikte, mevcut YZ sistemlerinin önemli ölçüde performans gösterdiği de açıktır. Bu sistemlerin gücü, insan algısını aşan korelasyonları tespit edebilme, insanların bilinçli kararlar almasına yardımcı olabilme ve doğal gibi görünen, insanları yanıltabilecek yapay içerikler üretebilme yetenekleriyle ölçülebilir. YZ'nin kendisi, doğası gereği etik ya da etik dışı bir varlık değildir; ancak, uygulama biçimleri etik sorunlar doğurabilir. Nitekim, YZ'nin cinsiyet ya da ten rengine dayalı ayrımcılık yaptığı vakalar zaten kaydedilmiştir. Ayrıca, toplumda her şeyi bilen bir YZ sisteminin yargılarına bağımlı hale gelme korkusu da mevcuttur. Bu nedenle, YZ teknolojilerine duyulan güvenin artırılması, bu teknolojilerin geniş çapta kabul görmesi açısından temel bir gereklilik olarak ortaya çıkmaktadır. Bu çerçevede, YZ'nin etik ilkelerle uyumlu bir şekilde tasarlanması ve uygulanması konusunda, ulusal ve uluslararası düzeyde yoğun tartışmalar sürdürülmektedir. Teknik normlar ve standartlar, bu süreçte teknik gereksinimlerin, süreçlerin ve terminolojinin standardizasyonunu sağlayarak, kaliteyi güvence altına almak adına önemli bir rol oynayabilir. Normlar ve standartlar, uzman gruplarının, yani insanların, fikir birliği ile geliştirdiği için, bir dereceye kadar ahlaki değerlere de dayandırılabilir. Ancak, teknik standardizasyonun amacı, özellikle "etik YZ" bağlamında, kültürel olarak sabitlenmiş belirli değerleri ya da ahlak kurallarını desteklemek değildir. Normlar ve standartlar, bu nedenle, YZ hakkındaki çekinceleri olumlu yönde gidermede etkili olabilir.

2. YAPAY ZEKA ETİK VE AHLAK

Çalışmanın bu bölümünde ahlak ve etik, alt-sembolik ve sembolik yapay zekâ ile zayıf ve güçlü yapay zekâ sistemleri arasındaki farklarla ilgili temel sorulara yanıtlar sağlarken, YZ'nin etik mi yoksa ahlaki mi hareket etmesi gerektiği sorusuna yanıt aranmaktadır. Yapay zekâ sistemlerini tanımladıktan ve kuramsal çerçevesini oluşturduktan sonra ayrıca yapay zekâ sistemlerinin sınırları tartışılacaktır.

2.1. Etik YZ mi yoksa Ahlaki YZ Mi?

Etik ve ahlak kavramları, birbirleriyle derin bir bağlantı içerisindedir. Ahlak, daha geniş bir disiplin olan etiğin bir alt alanını oluştururken, etik de felsefenin bir dalıdır. Ahlak, temel olarak, belirli bir durumda bir bireyden beklenen davranışları tanımlar. Ahlak, sosyal, siyasi veya dini perspektiflerden etkilenebildiği için, farklı ahlaki sistemler ortaya çıkabilir, bir arada var olabilir ve zamanla değişebilir. Etik ise, bu ahlaki sistemlerin üzerinde bir düşünme düzeyi olarak yer alır; onları analiz eder, sistematize eder ve sorgular (Lenk,1993). Bu bağlamda, etik, çeşitli ahlaki kavramlara dayanan ve toplumsal süreçler tarafından belirlenebilen eylem kuralları, tavsiyeler, emirler ve yasaklar formüle edebilir. Bu çerçevede, yapay zekânın ahlaki davranış sergilemesinden ziyade, yapay zekânın geliştirilmesi ve (muhtemelen küresel ve dolayısıyla kültürlerarası) kullanımı için temel etik ilkelerin formüle edilmesi daha uygun bir yaklaşım olarak görünmektedir. Bu yaklaşım, ahlaki düşünceleri tamamen dışlamaz, fakat sonuçları ahlakın ötesine taşımaktadır (Lenk, 1993).

1.2. Yapay Zekâ Nedir?

Yapay zekâ (YZ) için uygun bir tanım oluşturmak, oldukça karmaşık bir görevdir. Bunun başlıca nedeni, zekânın net ve genel olarak kabul görmüş bir tanımının bulunmamasıdır. ISO/IEC 22989 standardının geliştirilmesi sürecinde, Uluslararası Standardizasyon Örgütü (ISO) ve Uluslararası Elektroteknik Komisyonu (IEC), YZ'ye dair bir tanım oluşturma sorumluluğunu üstlenmişlerdir. Şu anda üzerinde çalışılan tanım, bir sistemin bilgi ve becerileri edinme ve uygulama yeteneğine dayanmaktadır. Ancak, ISO/IEC 22989 standardının nihai, yayımlanmış versiyonuna kadar bu tanımda değişiklikler olabileceği göz önünde bulundurulmalıdır. YZ'nin tanımlanmasına yönelik bir diğer yaklaşım ise, terimin bileşenlerinin ayrı ayrı tanımlanmasıdır. Duden sözlüğüne göre, "yapay" terimi doğal bir sürecin taklidini ifade ederken, "zekâ" ise "[insanların] soyut ve rasyonel düşünme ve bu düşüncelerden amaçlı eylemler çıkarma yeteneği" olarak tanımlanmaktadır. Bu bağlamda, YZ, insan zihinsel yeteneklerinin soyut bir taklidini yaratma girişimi olarak değerlendirilebilir. Alt

başlık eklenirken bu şekilde numaralandırmaya dikkat edilmelidir. Rakam ve başlık arasındaki belirtildiği şekilde olmalı. Tüm başlıklar numaralandırılmalıdır.

YZ'nin bu tanımı, YZ'nin genellikle kamu perspektifinden ilişkilendirildiği bilgisayar bilimi alanında da benzer bir biçimde bulunabilir. Örneğin, YZ'nin "bilgisayarlara düşünmeyi öğretme çabası" (Haugeland, 1985) olarak tanımlanması, insan zihinsel performansının makineler veya bilgisayarlar gibi sistemler tarafından taklit edilmesini vurgulamaktadır. Bu görüş, örneğin, İngiliz bilim adamı Alan M. Turing tarafından geliştirilen ve bir YZ'nin ancak bir insanla doğal dilinde iletişim kurabilmesi, bilgiyi edinmesi ve temsil etmesi, mantıksal sonuçlar çıkarması ve değişen koşullara uyum sağlaması halinde geçebileceği 'Turing Testi'nde (Turing, 1950) yansıtılmaktadır. YZ'nin eşit derecede yaygın olan bir başka tanımında ise "akıllı ajanların tasarımının incelenmesi" olarak tanımlanmaktadır (D. Poole vd., 1998.). Daha önce bahsedilen varyantla karşılaştırıldığında, buradaki odak noktası insanları taklit etmekten ziyade sistemin ya da ajanın rasyonel davranması, yani bilgisine göre "doğru şeyi" yapması için istenen davranışsal idealdir. Bu tanımın dikkat çekici yanı açık olmasıdır- ajanın eylemleri "doğru" olmak için mantıksal sonuçlarla belirlenebilir, ancak belirlenmek zorunda değildir ve açıkça insan zekâsı üzerine modellenmemiştir (Russell, Norvig, 2010).

1.2.1 Yapay Zekanın Türleri

YZ sistemleri çeşitli özellikler temelinde farklılaştırılabilir. Bir YZ sisteminin bilinen gerçeklere dayalı olarak bir sonuç üretme süreciyle ilgili olarak tümevarımsal ve tümdengelimsel YZ sistemleri arasında bir ayrım yapılabilir- bu aynı zamanda girişim olarak da bilinir. Tümevarımsal sistemler, genel olarak uygulanabilecek kalıplar için bireysel örnekleri analiz eder. Tümdengelim sistemleri ise sabit kurallara dayalı bir sonuca ulaşır. İstenen zekâyı elde etmek için bir YZ sistemi tarafından kullanılan yaklaşımla ilgili olarak başka bir ayrım daha yapılır. Sembolik YZ, bilginin sayılar gibi açık sembollerle temsil edildiği ve zeki davranışın (büyük ölçüde) bu sembollerin matematiksel manipülasyonu ile taklit edilebileceği temel varsayımına dayanmaktadır (Whitaker, 2010). Sembolik YZ, zekâ performansına kavramsal seviyeden yaklaşır veya yukarıdan aşağıya bir yaklaşım izler. Bazen (daha modern olduğu varsayılan) alt-sembolik YZ'den ayırt etmek için "klasik" YZ olarak da adlandırılır. Öte yandan, alt-sembolik YZ, benzer girdi modellerini belirli çıktı modellerine eşlemek için kullanılabilecek bir bağlantıcı modelin oluşturulabileceği varsayımına dayanmaktadır. Subsimbolik yapay zekâ (YZ), zekâ performansına bilginin örtük bir temsili aracılığıyla yaklaşır; bu, başka bir deyişle, aşağıdan yukarıya doğru bir yöntem izlemektedir (Whitaker, 2010).

YZ sistemleri, kullanılan eğitim yöntemleri açısından da farklılaştırılabilir. Günümüzde baskın olan eğitim yöntemi makine öğrenmesidir ve bu da kendi içinde çeşitli türlere ayrılabilir. Denetimli öğrenmede (supervised learning), YZ sistemi -örneğin bir yapay sinir ağı- insanlar tarafından önceden değerlendirilmiş veya etiketlenmiş (açıklamalı) seçilmiş eğitim verileri ile eğitilir. Bu eğitim verileri temel alınarak, YZ sistemi bir fonksiyon yaklaşımı gerçekleştirir. Yani, girdi verileri ile istenen sonuç (regresyon) arasındaki ilişkiyi tanımlayan bilinmeyen (ve muhtemelen örtük) bir fonksiyonun en doğru tanımını geliştirir ya da önceden tanımlanmış kategorilere (sınıflandırma) girdi veri setlerinin uygun bir şekilde atanmasını sağlar. Buradaki temel zorluk, eğitim verileriyle ilgili hata oranını en aza indirirken, aynı zamanda eğitim verilerine dâhil olmayan girdi verileri için de anlamlı çıktılar üretebilecek yeterlilikte bir genelleme yapabilmektir (Mohri vd., 2012).

Buna karşılık, denetimsiz öğrenmede (unsupervised learning), YZ sistemine önceden manuel olarak değerlendirilmemiş (açıklama eklenmemiş) eğitim verileri sağlanır (Liu vd., 2021). Sonuç olarak, YZ sistemi olası sonuçlar ve bunların eğitim verileriyle ilişkileri hakkında genel bir bakışa sahip değildir. Bunun yerine, YZ sistemi eğitim süreci sırasında eğitim veya girdi verilerini yapılandıran, yani gruplara (kümelere) eşleyen bir model geliştirmelidir. Eğitim verileri çok kapsamlı ise (kısmen eğitim verilerinin manuel olarak değerlendirilmesi ve etiketlenmesi gibi yoğun emek gerektiren bir adım artık gerekli olmadığı için) veya (eğitim) verilerinde daha önce bilinmeyen korelasyonlar tespit edilecekse, denetimsiz öğrenme yöntemleri denetimli öğrenme yöntemlerine göre bir avantaj sunabilmektedir.

Pekiştirmeli öğrenme (reinforcement learning), denetimli öğrenmenin özel bir şeklidir (Kaelbling, 1996). Bu eğitim biçiminde, YZ sistemine eğitim veri setleri değil, bir öğrenme ortamı (örneğin oyun kuralları veya trafik senaryoları) ve ulaşılması gereken bir hedef sağlanır. Hedefe ulaşmak için herhangi bir strateji belirtilmez- YZ sisteminin bunu eğitim sırasında kendisinin geliştirmesi beklenir. Bunun yerine, öğrenme sürecinin bireysel adımlarının (muhtemelen) doğru veya yanlış yöne gitmelerine bağlı olarak "ödüllendirilebileceği" veya "cezalandırılabileceği" bir değerlendirme işlevi aracılığıyla geri bildirim alır. Güçlendirilmiş öğrenme kullanılarak eğitilen bir yapay zekâ sisteminin önde gelen bir örneği, 2017'de sunulan ve şu anda Go oyunları

alanındaki en güçlü yapay zekâ sistemi olarak kabul edilmekle kalmayan, aynı zamanda eğitimi insan hamlelerinin analizinden tamamen vazgeçen AlphaGo Zero'dur.

Pekiştirmeli öğrenmenin aksine, taklit öğrenme de YZ sisteminin bir zorluğun üstesinden gelmek için bir strateji geliştirmesini içermektedir. Ancak bu bir model stratejiye dayanmaktadır (Marr, 2022). Hedefe bağlı olarak, YZ sistemi strateji modelini taklit etmeyi öğrenmeli veya sadece gizli olarak, yani temel özelliklerinde benimsemelidir. Taklit öğrenme yöntemi, mevcut bölgesel ekosistemlerde trafik kurallarının somut olarak yorumlanması için ortak bir uygulamanın öğrenilmesini sağladığından ve müşterinin tercihlerine göre sürüş tarzında değişikliklere izin verdiğinden, tam otomatik araçlar için uygun hale gelebilir (Age).

Ayrıca, YZ sistemleri performanslarına ve uygulama alanlarına göre de farklılaştırılabilir. Sözde "güçlü" ve "zayıf" YZ arasındaki soyut ayrım, YZ araştırmalarında özellikle yaygındır. Bu ayrımın temeli felsefi niteliktedir ve iki hipoteze dayanmaktadır: bir sistemin (örneğin bir makinenin) akıllıca davranabileceğine dair daha zayıf hipotez ve böyle bir sistemin bir zihne sahip olabileceğine dair daha güçlü hipotez. Buna göre, güçlü bir YZ sistemi gerçekten düşündüğü için akıllı davranışlar sergilerken, zayıf bir YZ sistemi yalnızca akıllıymış gibi davranır (Russell vd., 2010). Güçlü bir YZ sisteminin performansı, insan beyninin performansına eşit veya hatta daha üstün olacaktır. Buna karşılık, zayıf bir YZ sistemi bireysel görevlerin işlenmesinde uzmanlaşır ve insanların yerini almaktan ziyade onları (zihinsel) çalışmalarında destekleme amacına hizmet eder (Nilson, 2010).

Güçlü YZ'nin olası farkındalığı ve genel fizibilitesi hakkındaki sorular hala tartışmalı bir konu olsa da (Nilson, 2010.), zayıf YZ sistemleri zaten bir gerçektir ve birkaç yıldır başarıyla kullanılmaktadır. Yukarıda bahsedilen AlphaGo Zero'ya ek olarak, birçok akıllı telefonda bulunanlar gibi yardımcı sistemler de iyi bilinen diğer örneklerdir. Yakın zamanda Google, Google Duplex ile asistan sisteminin daha da geliştirilmiş bir adımını göstermiş ve YZ sisteminin kendisini bir YZ sistemi olarak tanımadan bir kuaför salonu çalışanından telefonla saç kesimi için randevu almasını sağlamıştır (Google. Google AI Blog). On yıllar önce, ELIZA5 da dahil olmak üzere çeşitli sohbet robotları, insanları aslında bir makine ile (metin tabanlı) bir konuşma yaptıklarına inandırmayı başarmıştı (Russell, 2010), bu da sadece zayıf YZ sistemlerinin temel yeteneklerini vurgulamaya hizmet etmektedir.

1.3. Yapay Zekayı Doğru Sınırlandırmak

YZ'ye etik ışığında bakmak, yalnızca neyin geliştirilebileceğine değil, aynı zamanda nasıl geliştirilmesi ve kullanılması gerektiğine de odaklanmayı gerektirir. Kamuoyundaki algının aksine, YZ sistemleriyle ilgili sorun, kendilerine dayatılan sınırları aşmaları değil, bu sınırlara tam olarak uymalarıdır (Shane, 2019.). Bu nedenle etik ilkelere uyulması, alt-sembolik YZ'nin geliştirilmesinde özel bir zorluktur. Eğitim yönteminin ve verilerin seçimine bağlı olarak, YZ sistemine bir model geliştirmesi için az ya da çok geniş bir özgürlük tanınır- bu özgürlüğün tükenmesi nihayetinde YZ sisteminin istenmeyen davranışlarına neden olabilir.

Bu nedenle, denetimli öğrenme kullanılarak eğitilen YZ sistemlerinin etik davranışı için temel bir ön koşul, eğitim veri setlerinin derlenmesinin zaten etik yönlere göre ve (mümkün olduğunca) önyargısız olarak gerçekleştirilmesidir. Ancak, bu tek başına ilgili YZ sistemlerinin daha sonra gerçekten etik davranış sergileyeceğinin garantisi değildir (Council, National Science and Technology, 2016). Özellikle kritik uygulamalar söz konusu olduğunda, bu YZ sistemleri önceden etik davranış açısından test edilmelidir. İkincisi, denetimsiz öğrenme kullanılarak eğitilen YZ sistemleri için daha da geçerlidir. YZ sisteminin eğitim sırasında daha özgür olma eğiliminde olması ve eğitim verilerinin daha önce açıklanmamış olması- ve bu nedenle de daha kolay ve daha yaygın olarak erişilebilir olması- istenen etik davranışı sağlamayı daha da zorlaştırabilir. Denetimli öğrenmeye benzer şekilde, eğitim verilerinin seçimi veya bu durumda uygun bir strateji modelinin seçimi de taklit öğrenmede YZ sisteminin istenen etik davranışında belirleyici bir rol oynamaktadır. Muhtemelen ölçülebilir bir değer sistemini öğrenmeye dahil etmenin en doğrudan yolu takviyeli öğrenme yöntemidir. YZ sisteminin gelişimi, ödül ve ceza fonksiyonu aracılığıyla oldukça doğrudan istenen yönde yönlendirilebilir.

Genel olarak, YZ sisteminin davranışı, seçilen öğrenme yönteminden ve ilgili parametrelerden çok güçlü bir şekilde etkilenmektedir. Bu nedenle, YZ sistemlerinin davranışını doğrulamak, eğitim yöntemlerini düzenlemekten ziyade istenen etik davranışa daha fazla katkıda bulunabilir.

2. ETİK YAPAY ZEKA

Kamusal söylemde ve literatürde "otonom" teriminin, asıl anlamının "otomatikleştirilmiş" olduğu yapay zekâ kontrollü makineler ve araçlar bağlamında kullanılması yaygındır. Ancak, bu iki terimin eşanlamlı kullanımı sorunludur: "otomatik" terimi Latince automatus'tan türetilmiştir ve "gönüllü", "kendiliğinden" ve hatta "kendi kendine hareket eden" olarak çevrilebilir (Whitaker, 2010). Öte yandan "otonom" terimi Yunanca'dan türetilmiş bileşik bir kelimedir. Yunanca "kendi" anlamına gelen dönüşlü zamir autos'tan türetilen auto kelime modülü, "kendi", "kendine ait" veya "kendine ait" gibi anlamlara gelmektedir. İkinci kelime unsuru nomos ise "yasa" ya da "gelenek" anlamına gelmektedir. "Otonom" bu nedenle "kendi yasası", "kendi yasası" ya da "kendi başına yasa" anlamına gelir (Harper, 2018). Bu iki terimin temel anlamlarına bakıldığında, eşanlamlı kullanımlarının uygun olmadığı ve aslında özerk ile otomatik ya da "kendi kendini yasalaştıran" ile "kendi kendine hareket eden" arasında kesin bir ayrım yapılması gerektiği ortaya çıkmaktadır.

2.1. Etik Nedir?

Yapay zekânın (YZ) sosyal sonuçları hakkındaki tartışmalar iki uç görüşle karakterize edilmektedir: Komplo teorisine inanlar insanlığın terminatör tarafından tamamen yok edileceğini öngörürken, konuya sorgusuzca coşkuyla yaklaşanlar ise tamamen dijitalleşmiş bir Akıllı Cennet'e gireceğimizi tahmin etmektedir. Her ikisi de bildiğimiz dünyanın sonu anlamına gelecektir. Tam da bu noktada etik tartışmalar, toplumun bütünü için herkes tarafından arzu edilen alternatifler geliştirmek için gereklidir. Nitekim teknoloji sadece ne yapabileceğimizi gösterirken etik bize ne yapmamız gerektiğini söyler. Bununla birlikte, hangi etik biçimini takip ettiğinize bağlı olarak, cevap oldukça farklı da olabilmektedir. Aşağıda söz edilen farklı etik biçimleri sıralanmıştır. Bu biçimlere göre yapay zekâ sistemlerinde kullanılması gereken etik biçimleri tartışılacaktır.

a. Erdem Etiği

Erdem etiği, Aristoteles'in Nikomakhos'a Etik eserine dayanan ve Avrupa'nın en eski bilimsel etik anlayışıdır. Erdem etiğine göre, doğru eylem, tamamen erdemli bir insanın en iyi niyet ve düşüncelerle gerçekleştireceği eylemdir. Birey, davranışını düzenli olarak özellikle erdemli insanların davranışlarıyla karşılaştırarak evrensel bir etik ideale giderek daha fazla yaklaşabilir (Quante, 2011). Bu etik anlayışı, doğru eylemin belirlenmesinde kişinin erdemli bir birey olarak gelişimini esas alır ve ahlaki değerlendirmelerin bağlamdan bağımsız olarak değil, belirli bir durumda sergilenen erdemlere dayalı olarak yapılması gerektiğini savunur (MacIntyre, 1981). Erdem etiği, modern etik tartışmalarda bireysel sorumluluk, toplumsal bağlam ve ahlaki değerlerin bütüncül bir şekilde ele alınmasını teşvik etmektedir (Hursthouse, 1999).

Bu yaklaşıma göre yapay zekâ ve etik entegrasyonu sürecinde, sistemlerin yalnızca kurallara uyumlu veya sonuç odaklı olmaktan ziyade, kullanıcıların ve geliştiricilerin erdemli bir şekilde hareket etmesini teşvik eden, ahlaki değerleri yansıtan yapılar oluşturulması önerilmektedir.

b. Sonuççuluk

Sonuççulukta ne belirli bir eylem ne de altta yatan motivasyon bir rol oynar. Önemli olan tek şey sonuçtur. Herkesin genel faydasını artıracak sonuçlar aranıyorsa buna faydacılık denir. Odak noktası kendisi için olumlu sonuçlara odaklanıyorsa, kişi etik egoizmi varsayar (Darwall, 2007). Bu çerçevede, doğru bir eylemin, mümkün olan en iyi sonuçları sağlaması gerektiği savunulur ve bu genellikle bireylerin ya da toplumun genel mutluluğunu artırma amacıyla ilişkilendirilir (Singer, 1993). Sonuççuluk etiği, özellikle faydacılık (utilitarianism) şeklinde, bireysel çıkarların ötesinde toplumsal refahı önceleyen bir ahlaki çerçeve sunar; ancak sonuçların tahmin edilebilirliği ve etik kararların bağlamsal farklılıkları gibi zorluklarla karşılaşabilir (Smart, Williams, 1973).

Sonuççuluk etik yaklaşımına göre, yapay zekâ ve etik entegrasyonu sürecinde, yapay zekâ sistemlerinin tasarım ve kullanımının, bireysel ve toplumsal düzeyde en iyi sonuçları sağlayacak şekilde optimize edilmesi gerektiğini vurgulayarak, bu süreçte etik kararların sonuçlarının dikkatlice değerlendirilmesi önerilmektedir.

c. Deontolojik Etik

Deontolojik etikte (deontology), belirli bir eylem ve onun sonucu ikinci plandadır. Önemli olan tek şey, eylemin daha yüksek bir kurala dayanıp dayanmadığı ve ona doğru şekilde uyulup uyulmadığıdır (Schmidt, 2011). Deontolojik etik, ahlaki değerlendirmelerde eylemin sonuçlarından ziyade, eylemin kendisinin ahlaki kurallara ve ilkelere uygun olup olmadığını temel alan bir yaklaşımdır (Kant, 1785). Bu yaklaşıma göre, bireyler belirli ahlaki yükümlülüklere ve evrensel etik ilkelere uymalıdır; bu ilkeler, bireylerin haklarını ve ödevlerini ihlal

etmeyen bir şekilde eylemde bulunmayı gerektirir (Rawls, 1971). Deontolojik etik, özellikle bireylerin öznel niyetlerini ve eylemlerinin evrensel bir yasa haline gelip gelemeyeceğini sorgulamaya dayalı Kant'ın kategorik imperatif kavramı ile ilişkilendirilir (O'Neill, 1989).

Bu bağlamda deontolojik etik yaklaşımı, yapay zekâ ve etik entegrasyonu sürecinde, sistemlerin evrensel ahlaki ilkelere ve kurallara uygun şekilde tasarlanmasını vurgulayarak, bu sistemlerin sonuçlarından bağımsız olarak bireylerin haklarını ve etik ödevlerini ihlal etmeyecek şekilde yapılandırılmasını gerektirir.

Erdem etiği, sonuççuluk ve deontolojik etik yaklaşımı genel hatlarıyla açıklandıktan sonra bu yaklaşımlardan yola çıkarak yapay zekâ sistemlerinin etik ilkelerle nasıl birleşeceği sorusu etik ilkelerin YZ sistemlerine entegre edilmesi süreciyle değerlendirilmelidir.

2.2. Yapay Zekâ ve Etik Nasıl Birleşir?

Üç farklı etik biçiminden söz ettikten ve etik ve ahlakın önemini vurguladıktan sonra yapay zekâ sistemlerine etik bir bilinç nasıl entegre edileceği bu bölümde incelenecektir. Yapay zekâ ya etiğin dahil edilebileceği konusuna girmeden önce aşağıdaki üç farklı yöntemi birbirinden ayırmak önem arz etmektedir.

a. Tasarım Yoluyla Etik:

Yapay zekâ, teknik veya algoritmik işlevlere dayalı olarak etik kararlar verme yeteneği kazanır. Uzmanların çoğuna göre bu yalnızca bilim kurguda var (Strategie Künstliche Intelligenz der Bundesregierung, 2018).

b. Tasarımda Etik:

Tasarım sürecinde geliştirici, yapay zekâ nın etik davranışını sağlayan kuralları veya yöntemleri kullanır. Bunlar örneğin şunları içerir: Eğitim verilerinde önyargıdan kaçınmak (Ag.).

c. Tasarım Etiği:

Geliştiriciyi etik olarak bir yapay zekâ tasarlamaya motive etmek amacıyla geliştiriciler için kurallar, standartlar, önlemler vb. geliştirilir (Haarich, 2019).

YZ uygulamaları tarafından alınan kararlara ve insanların bunları ne ölçüde etik olarak nitelendirdiğine bakılırsa, iyi, kötü ve çirkin uygulamalar olarak sınıflandırma yapılması gerekmektedir.

İyi uygulamalar, kararlardaki önyargıları azaltabilir. Bunun tipik bir örneği personel kararlarıdır, örneğin sorumlu kişinin belirli grupları uygunsuz bir şekilde kayırması gibi. Bir YZ uygulamasının objektif ve dolayısıyla önyargısız olduğu fikri çok riskli olsa da, bir YZ uygulaması kararları daha şeffaf hale getirebilir. Bu şeffaflık, kararların olası alakasız önyargılar açısından analiz edilebileceği ve gerekirse azaltılabileceği anlamına gelmektedir (Haarich, 2019).

Öte yandan kötü uygulamalar, rasyonel olabilecek, ancak birçok insanın etik açıdan yanlış olduğunu düşündüğü kararlar verebilir. Kurgusal bir örnek "I, Robot" filminde görülebilir. Bir robot iki kişiden hangisinin hayatta kalma şansının daha yüksek olduğuna karar verir. Bunun sonucunda sistem yetişkin adamı kurtarır ve küçük kızın boğulmasına izin verir.

Son olarak da çirkin uygulamalardan bahsetmek gerekmektedir. Çirkin uygulamalar durumu çok ince bir şekilde değiştirmektedir. Kullanıcılar sosyal medya uygulamalarını kullanırken, çoğu zaman farkında olmadan algoritmalardan etkilenmektedir. Dünyamız fark edilmeden değişiyor, kullanıcıların uygulamayı mümkün olduğunca uzun süre kullanması için davranışlarımız manipüle ediliyor. Bu manipülasyon olumsuz bir niyet olmadan da gerçekleşebilmektedir. Burada sorulması gerek asıl soru ise şu: Bir yapay zekâ uygulaması mümkün olduğunca "etik" davranacak şekilde nasıl tasarlanmalıdır? Nelere dikkat etmeniz gerekiyor?

Etik ile ne kastedilmektedir? Yapay zekâ ve etiği ele alırken, felsefi bir disiplin olarak etiğin temel soruyla ilgilenilmesi önem arz etmektedir. Örneğin İyi/doğru nedir, kötü/yanlış nedir? Bu daha sonra bir yandan hukukta, diğer yandan da yazılı olmayan ideallerde gerçekleştirilmektedir. Markkula Center for Applied Ethics'te etik hakkında şöyle bir tanımlama yapılmıştır:

"Etik, genellikle haklar, yükümlülükler, topluma faydalar, adalet veya belirli erdemler açısından insanların ne yapması gerektiğini belirleyen sağlam temelli doğru ve yanlış standartlarına dayanır." (Markkula Center for Applied Ethics, zitiert via IBM, 2022)

Kısaca özetlenecek olursa, etik, genellikle haklar, sorumluluklar, toplumsal faydalar, adalet veya belirli erdemler doğrultusunda, insanların nasıl davranması gerektiğini belirleyen sağlam temelli doğru ve yanlış ilkelerine dayanmaktadır. Tam da bu sebepten dolayı YZ'de etiğin neden bu denli önemli olduğu aşağıda sıralanan maddeler bağlamında tartışılacaktır.

a. İnsanlar yapay zekâ dan etik davranışlar bekler

Eğer bir yapay zekâ tavsiyelerde bulunur ya da kararlar alırsa, kullanıcılar etik davranışlar bekler. Etik olmayan öneriler, kendinizi haklı çıkarmanız gereken kararlar verirken yardımcı olmayacaktır.

b. Yapay zekâ kullanımı karar verme davranışlarımızı ve sonuçlarını şeffaf hale getirir

Yapay zekâ kullanımı, kararlarımızı ve dolayısıyla kararlarımızın temelini ve sonuçlarını şeffaf hale getirmektedir. YZ ile karar alma süreçleri ve bunların sonuçları kolaylıkla gözlemlenebilir ve tartışılabilir hale gelmektedir. İnsanlar, ya açık kurallar yoluyla ya da (makine öğrenimi durumunda) hangi davranışın doğru olduğuna dair geri bildirim yoluyla- etik olarak da- doğru davranışın ne olduğunu açıkça belirtmek zorundadır. İşte tam da bu şeffaflık- kararın nasıl verildiği ve sonuçlarının ne olduğu- değerler sorununu açık hale getirmektedir. Bu durum da aslında şu soruyu gündeme getirmektedir: İyi, adil ve doğru karar nedir? Bu soru sadece kurallar veya kılavuz ilkeler temelinde değil, aynı zamanda birçok vakada uygulanabilecek bir bakış açısıyla da yanıtlanabilmektedir. Örneğin, son on yıldaki tüm başvurulara bakabilir ve kararların belirli grupları adil olmayan bir şekilde cezalandırıp cezalandırmadığını nispeten kolay bir şekilde kontrol edilebilmektedir.

c. Bilgisayarlar insan davranışları üzerinde ciddi bir etkiye sahip olabilir

Yapay zekâ kullanımının kaçınılmaz olarak bilgisayar sistemlerini içermesi, beraberinde başka bir etik mesele doğurabilmektedir. Özellikle davranış değişikliği (örneğin, ikna edici teknolojiler aracılığıyla) bağlamında bilgisayarlar, insanları destekleme konusunda birçok önemli avantaja sahiptir. Bu avantajlara, kullanıcıları kendi seçtikleri daha olumlu davranışları benimsemeye teşvik etmek amacıyla tasarlanan farkındalık uygulamaları ve fitness takip cihazları gibi örnekler verilebilir. Fogg (2003), bilgisayarların avantajlarını süreklilik, anonimlik, büyük veri işleme kapasitesi, çeşitli modalitelerin kullanımı, kolay ölçeklenebilirlik (uygulamaların kolayca çoğaltılıp dağıtılabilmesi) ve insan etkileşiminin istenmediği ortamlarda kullanılabilirlik olarak sıralamaktadır. Bu bağlamda, bilgisayarlar, bireylerin alışkanlıklarını değiştirmelerine yardımcı olma konusunda oldukça etkili olabilmektedir. Bu güçlü yönlerin tümü, bir yapay zekâ sistemi bizim açımızdan etik olmayan kararlar verdiğinde bir sorun haline gelir.

3. YAPAY ZEKA SİSTEMLERİNİN KULLANIMINDA ETİK YÖNLER

Yapay zekâ sistemlerinin etik değerler çerçevesinde eğitilmesi gerektiği konusunun önemi daha önce tartışılmıştır ancak yapay zekâ kullanımının etik boyutları nelerdir? Etik, belirli bir öneri veya kararın ötesinde, aynı zamanda bu karar veya önerinin öncesinde gerçekleşen süreçleri ve uzun vadeli etkileri de kapsamaktadır. Bu bağlamda, etik değerlendirmeler; verilerin etik bir şekilde toplanıp toplanmadığını, sistemin işleyişini (örneğin, enerji tüketimi ve sürdürülebilirlik açısından maliyet-fayda analizi), yapay zekâ sisteminin yargılama ve karar verme süreçlerini (tarafsız ve adil kararlar verilip verilmediği) ve nihayetinde yapay zekâ sisteminin kullanımını (sonuçların etik bir şekilde kullanılıp kullanılmadığı) da içermektedir (Wing, 2021).

Yapay zekâ ve etik arasındaki temel unsurlar, adalet, önyargıdan kaçınma, veri koruma ve gizlilik, yapay zekânın genellikle incelikli etkilerinden kaçınma ve yapay zekânın kamu yönetimi ile toplumsal etki üzerindeki etkileridir.

3.1. Adillik/Adalet

Adalet şaşırtıcı derecede karmaşık bir kavramdır. Özünde, istenmeyen veya haksız ayrımcılığın önlenmesi veya engellenmesi ile ilgilidir. Ancak adaletin gerçekte nasıl tanımlandığı net değildir. Adillik aynı zamanda yapay zekâ alanında en çok ilgi gören konudur. Bunun nedeni kısmen, hiç kimseye ilgisiz özellikler temelinde daha az elverişli muamele edilmemesi ve YZ'yi görünür kılan bu tür muamelelerin genellikle yüksek düzeyde kamu ve medya ilgisini tetiklemesidir. Bu doğrultuda, IBM'in AI Fairness, Google'ın TensorFlow kiti, Microsoft'un Fairlearn, Facebook'un Fairness Flow veya Amazon & NSF'nin Fairness in AI (Wing, 2021) gibi büyük şirketler adaleti açıkça ele alma sürecindedir (Heine vd., 2023).

Adalet, özellikle kamu yönetiminde yasal açıdan büyük bir önem taşımaktadır. Poretschkin ve diğerlerine (2021) göre, "nesnel olarak haklı bir gerekçe olmadıkça, eşit sosyal durumlara eşit olmayan veya farklı sosyal durumlara eşit muamelede bulunma yasağı" adaletin temel ilkesini oluşturmaktadır. Bu doğrultuda, karar verme

süreçlerinde ilgisiz kriterlere dayalı ayrımcılık yapılmamalıdır. Milliyet, etnik köken, cinsiyet, din/inanç, engellilik, yaş grubu veya cinsel kimlik gibi özellikler, karar ile doğrudan ilişkili olmadıkları sürece, ayrımcılık yapılmaması gereken kriterler arasında yer almaktadır (Poretschkin vd., 2021). Adalet ilkesi, özellikle tahminlerde bulunan sistemler (birçok karar destek sistemi gibi) açısından büyük bir önem taşımaktadır.

Adaletin nasıl değerlendirileceği sorusu, sanıldığından çok daha karmaşık bir meseledir. Bu durumun temel nedeni, adaletin farklı değerler temelinde, farklı standartlar çerçevesinde değerlendirilmesidir (bkz. Verma ve Rubin, 2018). Bireysel vakalar düzeyinde, doğru bir karar verme süreci görece basit gibi görünse de, bu durum karmaşıklık barındırmaktadır. Yapay zekâ sistemi, doğru bir karar verebilir (örneğin, doğru pozitif durumda başvuru onaylanır ve kişi gerçekten de yetkilendirilir veya doğru negatif durumda başvuru reddedilir ve kişi yetkilendirilmez). Ancak, sistem yanlış kararlar verme yetisine de sahiptir (Age.).

3.2. Önyargılardan Kaçınma

Özellikle adalet kriterlerinin ihlal edilmesi durumunda, bu sapmaların kaynağının ne olduğu sorusu gündeme gelmektedir. Bu durum, önyargılar meselesine işaret eder. Yapay zekâ sistemlerinin geliştirilmesi sırasında, başlangıçta var olmaması gereken önyargıların ortaya çıkması mümkündür. Bu önyargılar, özellikle makine öğrenimi kullanıldığında ve yapay zekâ sistemlerinin geliştirilmesinde birçok kişinin yer aldığı durumlarda kolayca meydana gelebilir. Önyargılar, eğitim verilerinin oluşturulmasında (veri oluşturma önyargısı), problem tanımında, algoritmalar ve veri analizi süreçlerinde ve insanların değerlendirme ve doğrulama işlemlerinde ortaya çıkabilir (bkz. Srinivasan, Chander, 2021). Bu önyargılar, geliştirme süreci boyunca, eğitim verilerinin oluşturulmasından problemin tanımlanmasına, algoritmaların tasarımından veri analizine ve son olarak insanların değerlendirme ve doğrulama süreçlerine kadar birikerek etkisini gösterebilir.

Problemin tanımı özellikle dikkat çekicidir. Örneğin, bir iş bulma kurumunun yapısı incelendiğinde, iyi bir işe yerleştirmenin nasıl ölçülebilir hale getirilebileceği sorusu ortaya çıkmaktadır. Bu, yapay zekâ sistemine hangi kararların doğru ya da yanlış olduğuna dair geri bildirim sağlanabilmesi (operasyonelleştirme) için kritik bir unsurdur. Örneğin, hedef, belirli bir zaman diliminde mümkün olduğunca az sayıda kişinin iş arıyor olması olabilir; bu da, amacın, mümkün olan en kısa sürede yeterince uygun bir iş bulunmasını sağlamak olduğu anlamına gelir. Ancak, insanların uzun vadede kendilerine uygun, gelişimlerine katkı sağlayacak bir iş bulmalarını hedeflemek de mümkündür. Bu durum, modelin doğru ve yanlış kararlarla eğitilmesi için hangi verilerin seçileceğini ve bu verilerin hangi kararlara dayandığını belirleyen, problemin nasıl çerçevelendiğiyle ilgili bir meseledir. Bu bağlamda, "çerçeveleme etkisi önyargısı" devreye girmektedir. "Önyargı" konusu, gelecekte büyük olasılıkla, eğitim verilerinin mümkün olduğunca temsili ve "önyargısız" olması gerekliliği ile ilgili olarak daha da önemli bir sorun haline gelecektir.

Tüm bu çarpıklıklar göz önünde bulundurulduğunda, yapay zekâ kullanımının mantıklı olup olmadığını sorgulamak yerinde olabilir. Ancak, insan zihninin de çarpıtmalardan muaf olmadığı unutulmamalıdır. İnsanlar, karar verme süreçlerinde birçok temel kural (heuristics) kullanır ve bu süreçler sıklıkla çeşitli çarpıtmalara maruz kalır. Bununla birlikte, bu sezgisel yöntemler ve önyargılar, insanların tahminlerde bulunma ve karar verme konusundaki yetkinliklerini göz ardı etmemelidir; zira bu yetenekler hayatta kalmak için gereklidir. Ancak, insanlar tarafından geliştirilen bir yapay zekâ sisteminin de benzer şekilde önyargılar sergilemesi şaşırtıcı değildir. Peki, önyargılar nasıl önlenebilir? Srinivasan ve Chander (2021), alan bilgisine sahip olmanın merkezi bir öneme sahip olduğunu ve yapay zekânın öğreneceği özelliklerin bilinçli bir şekilde seçilmesi gerektiğini önermektedir. Ayrıca, veri tabanının popülasyonu doğru bir şekilde temsil etmesi, verilerin etiketlenmesi konusunda net standartlar bulunması ve ilgili değişkenlerin (eksik olanlar dahil) açıkça tanımlanması gerekmektedir. Son olarak, modelin temsili bir örneklem ile test edilmesi önerilmektedir.

3.3. Verilerin Korunması ve Gizlilik

Yapay zekâ sistemleri, özellikle makine öğrenimi kullanıldığında, büyük miktarda veriye ihtiyaç duymaktadır. Ancak bu veriler, belirli bir kaynaktan sağlanmalıdır. Bu bağlamda, veri koruma ve mahremiyet, özellikle de bireylerin bilgisel kendi kaderini tayin hakkı ve kişisel haklarının korunması büyük önem taşımaktadır (Poretschkin vd., 2021). Bu gereklilik, ses kayıtları için geçerli olduğu gibi, fotoğraf ve videolar gibi diğer veri türleri için de geçerlidir. Hukuki açıdan, Genel Veri Koruma Yönetmeliği (GDPR) ve Federal Veri Koruma Yasası (BDSG) bu konuda belirleyici yasal çerçeveyi oluşturmaktadır. Verilerin kullanımı sırasında, kötüye kullanımın fiilen gerçekleşip gerçekleşmediğinden ziyade, böyle bir olasılığın varlığının bile sorun teşkil ettiği unutulmamalıdır. Ancak, veri koruma ve mahremiyet yalnızca verilerin yetkisiz kişilerce ele geçirilmesi ("veri

sızıntısı") ile sınırlı değildir. Bir diğer önemli mesele, yapay zekâ sistemlerinin eğitimi sırasında verilerin sıkça birleştirilmesidir ("veri bağlantısı"). Verilerin bu şekilde bir araya getirilmesi, bireylerin kimliklerinin tespit edilmesini kolaylaştırabilir ve bu durum, bir veri sızıntısı halinde, hızlı bir şekilde itibar kaybına veya mali zarara yol açabilir.

Buna karşın, veri koruma ve gizliliğe aşırı derecede odaklanmak, yenilikçiliği engelleyebilmektedir. Bu durum, veri koruma ve gizliliğin göz ardı edilmesi ya da aşırı vurgulanması gerektiği anlamına gelmemektedir; fakat, bu konuların sonuçları da dikkate alınmalıdır. Bir yapay zekâ sisteminin etkin bir şekilde karar verebilmesi için veri toplaması gerekmektedir ve bu durum, insanlar için de farklı değildir. İnsanlar uzmanlıklarını çeşitli vakalarla etkileşimde bulunarak geliştirirler. Veri koruma ve gizliliğin sağlanabilmesi için, etkilenen bireylerin rızası şarttır; ek işleme yalnızca rıza ile yapılabilir, yetkisiz erişim olmamalı ve herhangi bir zamanda kapsamlı bir itiraz hakkı garanti edilmelidir. Ayrıca, bireyler kişisel verilerin veya bunlardan türetilmiş verilerin amacı ve kullanımı hakkında bilgilendirilmelidir. Genel olarak, veri minimizasyonu ilkesi ve tahsis edilmiş kullanım ilkesi geçerlidir ve bu kriterler göz önünde bulundurulmalıdır.

3.4. Sosyal Etki

YZ'nin toplumda yarattığı değişimler, bireylerin yaşam tarzlarından, eğitim sistemlerine kadar geniş bir yelpazeye yayılarak toplumsal yapıyı değiştirebilir. YZ'nin toplumsal etkisi üzerine yapılan çalışmalar, bu teknolojilerin sosyal ilişkileri, toplumsal normları ve bireylerin kendilik algısını nasıl dönüştürebileceğini sorgulamaktadır. Binns (2018), YZ'nin toplumsal etkileşim biçimlerini ve kültürel algıları nasıl şekillendireceğini tartışırken, bu teknolojilerin insanları daha bağımsız ve izole bir hale getirebileceğini de savunmaktadır.

Yapay zekâ sistemlerinin sosyal yapılar üzerindeki etkilerini tahmin etmek oldukça zordur. Teknoloji kullanıldıkça, sosyal ilişkiler de değişir; bu durumun tersinin de geçerli olduğu söylenebilir. YZ ile doğrudan ilgili olmasa da uygun bir örnek, bir kusur dedektörünün kullanıma sunulmasıdır. Bu teknoloji, vatandaşların sorunları (örneğin, çöp atma gibi) doğrudan yetkili makamlara bildirmelerini sağlamaktadır. Bununla birlikte, bireyler başkalarının yanlış davranışlarını büyük ölçekte rapor edebilirler. Bu durumun topluluk üzerinde etkileri de olabilir. Teknoloji, sosyal sistemlerde her zaman beklenmedik sonuçlara yol açabilir. Bu nedenle, etkiler uzun vadede izlenmeli ve temel hedefler açısından değerlendirilmelidir.

YZ sistemlerine etik ilkelerin yerleştirilmesi pratikte her ne kadar uygulanabilir gözükse de İnsanların psikolojik ve sosyal varlıklar olduğunu göz ardı etmek doğru değildir. Milgram Deneyi, yapay zekâ (YZ) ve etik entegrasyonu bağlamında, otoriteye itaat ve bireylerin ahlaki sorumlulukları arasındaki ilişkiyi inceleyen önemli bir örnek olarak değerlendirilebilir.

3.5. Milgram Deneyi Örneği ve YZ Sistemlerine Etik İlkelerin Entegrasyonu

Yapay zekâ ile etkileşimde bulunurken, talimatları basitçe takip etmek oldukça kolaydır; çünkü yapay zekâ sistemi, kararların "objektif" ve "doğru" bir şekilde alındığı izlenimini verebilir. Bu bağlamda, psikolojide itaat üzerine yapılan şüpheli bir deneyin ele alınması faydalı olacaktır. Milgram Deneyi, bireylerin etik kararlarını otorite figürlerinin etkisi altında nasıl şekillendirdiğini gösteren önemli bir psikolojik çalışmadır. Stanley Milgram'ın 1960'larda gerçekleştirdiği bu deney, katılımcıların, kendilerine verilen emirler doğrultusunda, başka bir kişiye elektrik şoku verme eylemini gerçekleştirmelerini test etmiştir (Milgram, 1963). Deneyin sonuçları, insanların, kendilerinin etik sorumluluklarını göz ardı ederek otoriteye itaat edebileceklerini göstermiştir. Bu bağlamda, Milgram'ın bulguları, yapay zekâ (YZ) kullanımındaki etik sorunlarla paralellikler taşımaktadır. YZ sistemlerinin bireylerin kararlarını etkilemesi, tıpkı Milgram'ın deneyinde olduğu gibi, etik sorumluluğun dışarıya devredilmesine ve potansiyel olarak zararlı sonuçlara yol açabilmektedir.

Milgram'ın deneyindeki katılımcılar, bir otorite figürünün emriyle, başka bir kişiye zarar vermekten çekinmemişlerdir. Bu, bireylerin kendi etik değerlerini, bir otoriteye itaat ederek aşmalarını göstermektedir. YZ de benzer şekilde bireylerin kararlarını etkileyebilmektedir. Özellikle YZ'nin algoritmalarla çalışan otonom sistemlerde, bireyler, bu sistemlerin "kararlarını" sorgulamadan kabul edebilmektedir. Otoriteyi temsil eden YZ, insanları etik sorumluluklarını göz ardı ederek belirli bir eylemi gerçekleştirmeye yönlendirebilmektedir. YZ'nın bu bağlamdaki toplumsal etkilerine değinilecek olursa, YZ, otoriteyi bir araç olarak kullanabileceği anlaşılmış olup, tıpkı Milgram'ın deneyindeki gibi, toplumsal kararlar üzerinde önemli etkiler yaratabileceği fark edilmiştir. Aynı zamanda YZ'nin toplumsal etkileri üzerine yapılan araştırmalar, bu tür sistemlerin, bireylerin etik değerlerini ve toplumsal normları aşmalarına yol açabileceğini öne sürmektedir (Zuboff, 2019). YZ'nin karar alma süreçlerinde önyargılı verilere dayanması, toplumsal adaletsizliklere neden olabilir ve bu

durum, Milgram'ın deneyindeki itaat etme sürecine benzer şekilde, bireylerin kendi etik sorumluluklarını devretmelerine yol açabilir.

Bireysel sorumluluk ve otorite kapsamında etik ilkelerin Milgram Deneyi örneğiyle ele alındığında kişilerin verilen emirleri takip ederek etik sorumluluklarını devredebileceği saptanmıştır. YZ sistemlerinin tasarımında da benzer bir tehlike mevcuttur. YZ, insanlara, etik kararlarını sorgulamadan verme eğilimini teşvik edebilmektedir. Bu durum, bireylerin YZ'ye olan güvenlerinin, onları etik sorumluluklarından kaçınmaya yönlendirebileceği anlamına gelmektedir (O'Neil, 2016). Milgram'ın deneyinde, katılımcıların etik sorumluluklarını devretmeleri, otoritenin gücünü sorgulamamalarına yol açmıştır. Bu durum, YZ'nin toplumsal etkilerinde de görülebilir. YZ sistemleri, etik değerleri göz ardı ederek zarar verici kararlar alabilir. YZ'nin, tıpkı Milgram'ın deneyindeki otorite figürleri gibi, bireylerin kararlarını şekillendirebileceği tespit edilmiştir (Binns, 2018).

4. SONUÇ

Yapay zekâ (YZ) sistemlerine etik ilkelerin eklenmesi, teoride mümkün görünse de pratikte uygulanması önemli zorluklar barındırmaktadır. Milgram Deneyi, bireylerin otorite figürlerinin talimatlarına itaat ederken kişisel etik sorumluluklarını nasıl devredebileceğini gözler önüne sermektedir. Bu bulgu, YZ sistemleri bağlamında da geçerlidir. YZ sistemleri, bireyler ve toplum üzerinde bir otorite algısı yaratabilir ve etik sorumluluğun bu sistemlere devredilmesi, kullanıcıların kararlarını sorgulamadan kabullenmelerine yol açabilmektedir.

Teorik olarak etik ilkeler, YZ'nin tasarımına algoritmalar aracılığıyla entegre edilebilir. Örneğin, adalet, eşitlik ve zarar vermeme ilkeleri, YZ'nin karar alma süreçlerine programlanabilir. Ancak Milgram Deneyi'nin gösterdiği gibi, bireylerin etik davranışları, yalnızca ilkelerle değil, aynı zamanda bağlam ve otoriteye karşı verilen tepkilerle şekillenmektedir. Benzer şekilde, YZ'nin içinde bulunduğu bağlam, kullanılan veri setleri ve sistemin kullanıcıları üzerindeki etkisi, etik ilkelerin uygulanmasını zorlaştırabilmektedir.

Pratikte, YZ sistemlerinin etik ilkelere uygun çalışmasını sağlamak için şeffaflık, hesap verebilirlik ve sürekli izleme mekanizmalarının oluşturulması gereklidir. Ancak, algoritmaların karmaşıklığı ve verilerin önyargılardan arındırılmasının zorluğu, bu mekanizmaların etkili bir şekilde uygulanmasını sınırlandırabilmektedir. Ek olarak, YZ sistemlerinin öngörülemez karar alma süreçleri, etik ihlallerin önceden tespit edilmesini zorlaştırmaktadır.

Milgram Deneyi'nden çıkarılabilecek sonuç, etik ilkelerin yalnızca teknik olarak sistemlere entegre edilmesiyle sınırlı kalmaması, aynı zamanda kullanıcıların ve geliştiricilerin etik farkındalıklarının artırılması gerektiğidir. Aksi takdirde, YZ sistemleri, bireylerin etik sorumluluklarını devrederek zarar verici sonuçlara yol açabilmektedir. Bu bağlamda, YZ'ye yönelik etik çerçevelerin, teknolojinin insan davranışları ve toplumsal yapı üzerindeki etkilerini dikkate alarak çok boyutlu bir şekilde ele alınması zorunludur.

Ayrıca bir yapay zekâ uygulamasının etik açıdan nasıl değerlendirileceği, etik standartlara ne ölçüde uyulduğu veya bu standartların ihlal edilip edilmediği ile ilgili sorulara yanıt aramak önemlidir. Birey olarak, genellikle uygulamanın yalnızca kısmi yönlerini değerlendirmek mümkündür. Bir kullanıcı olarak, yapay zekâ işlem hattının ilk aşamaları üzerinde çok az etki veya bilgi bulunulabilmektedir. Ancak, karar alma süreçlerinin nasıl gerçekleştirildiğine ve sonuçların anlaşılabilir olup olmadığına odaklanılabilir. Özellikle, yapay zekâ uygulamalarını değerlendirirken kişisel sorumlulukların bilincinde olunması önem arz etmektedir. Bu sorumluluk yalnızca yapay zekâ sistemine veya programcılara devredilmemelidir.

Bu faktörlerin yapay zekâ sistemlerinin kullanımına aktarıldığında ne gibi sonuçlar doğurabileceği değerlendirildiğinde, şu hususlar öne çıkmaktadır: Eğer sistem kesin kararlar veriyorsa (deneyci tarafından verilen talimatlar doğrultusunda), sisteme bir sertifika aracılığıyla objektif ve adil bir statü atfedilmişse (deneycinin otorite figürü olarak kabul edilmesi gibi), sistem giderek daha etik olmayan kararlar veriyorsa (örneğin, elektrik şoklarındaki kademeli artışlar), ve sistem zaman baskısı altında kullanılıyorsa, bu durum sistem kullanıcılarının etik olmayan kararları kabul etme olasılığını artırabilmektedir. Yapay zekânın yansıtılmamış kullanımında, sorumluluğun yapay zekâ sistemine atfedilmesi, yapay zekâ uygulamasının hatasız olarak algılanması, yapay zekâ sisteminin uygulama alanının kademeli olarak genişletilmesi, kullanıcıların yüksek iş yükü nedeniyle bunalmaları ve nihayetinde kimsenin itiraz etmemesi gibi riskleri ortaya çıkarabilir. Bu bağlamda, yapay zekâ sistemleri açık kriterler temelinde değerlendirilmelidir. Nitekim yapay zekâ sistemlerinin, en azından şu aşamada, zayıf yapay zekâ sistemleri olarak insan girdisi dahilinde karar verip

çalıştığı bilinmektedir. Dolayısıyla ancak insan girdisinin etik ilkelere bağlılığı yapay zekâ sistemlerinin etik karar vermesine yardımcı olacaktır.

KAYNAKÇA

Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). *Social Psychology* (11. Baskı). Pearson Education, Inc.

Binns, R. (2018). *Social Impacts of Artificial Intelligence: A Perspective on Ethics, Justice, and Human Rights.* Springer.

Bundesregierung. Strategie Künstliche Intelligenz der Bundesregierung. (2018).

Council, National Science and Technology. *Preparing for the Future of Artificial Intelligence*. Washington : U.S Government, (2016).

D. Poole, A. K. Mackworth, R. Goebel (1998). *Computational intelligence: A logic approach*. Oxford, University Press.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do.* Morgan Kaufmann.

Google. Google AI Blog. An AI System for Accomplishing Real-World Tasks Over the Phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Haarich, Max, (2019). *Ethik im Zeitalter der Künstlichen Intelligenz*.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. s.l. : MIT Press

Harper, Douglas. (2018). Online Etymology Dictionary. MaoningTech, https://www.etymonline.com/word/auto-?ref=etymonline_crossreference.

Heine, M., (2023). *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_11, Edition, Edt. Jörn von Lucke, Jürgen Stember, Maria A. Wimmer, Springer Verlag. Osnabrück.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research. 4: 237–285. arXiv:cs/9605103. doi:10.1613/jair.301. S2CID 1708582.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.

Lenk, Hans. (1993). Technik und Ethik. Stuttgart : Reclam Verlag.

Liu, Xiao; Zhang, Fanjin; Hou, Zhenyu; Mian, Li; Wang, Zhaoyu; Zhang, Jing; Tang, Jie (2021). "Self-supervised Learning: Generative or Contrastive". IEEE Transactions on Knowledge and Data Engineering: 1–1. arXiv:2006.08218. doi:10.1109/TKDE.2021.3090866. ISSN 1041-4347.

MacIntyre, A. (1981). *After Virtue: A Study in Moral Theory*. London: Duckworth.

Marr, B., (2022). https://bernardmarr.com/what-is-ai-imitation-learning-a-super-simple-guide-anyone-can-understand/.

Markkula Center for Applied Ethics, via IBM, 2022.

Mohri, M., Rostamizadeh, A., Talwalkar, A., (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258.

Milgram, S. (1963). Behavioral Study of Obedience. Journal of Abnormal and Social Psychology, 67(4), 371–378.

Nilson, N. J. (2010). The Quest for Artificial Intelligence. Cambridge, University Press.

O'Neill, O. (1989). Constructions of Reason: Explorations of Kant's Practical Philosophy. Cambridge. Cambridge University Press.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group.

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., & Wrobel, S. (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz – KIPrüfkatalog. Fraunhofer IAIS. www.iais.fraunhofer.de/ki-pruefkatalog.

Quante, M., (2011). *Einführung in die Allgemeine Ethik*. 4. Baskı. Wissenschaftliche Buchgesellschaft, Darmstadt. ISBN 978-3-534-24595-6.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Russell, S.J., Norvig, P., (2010). *Artificial Intelligence*. s.l. : Pearson Education Inc.

Schmitd, T. (2011). *Deontologische Ethik.* Ralf Stoecker/Christian Neuhäuser/Marie-Luise Raters (Edt.): Handbuch Angewandte Ethik içinde. Verlag J.B. Metzler, Stuttgart, ISBN 978-3-476-02303-2.

Shane, J., (2019). The danger of AI is weirder than you think. s.l. : TED2019.

Singer, P. (1993). *Practical Ethics. Cambridge*. Cambridge University Press.

Smart, J. J. C., Williams, B., (1973). *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. Communications of the ACM, 64(8), 44–49. https://doi.org/10.1145/3464903.

Stephen Darwall (Edt.) (2007): Consequentialism. Nachdr. Malden, Mass.: Blackwell. Blackwell readings in philosophy 7. ISBN 978-0-631-23108-0.

Turing, A.M., (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Whitaker, W., (2010). William Whitaker's Words. University of Notre Dame, South Bend, IN, http://archives.nd.edu/cgi-bin/wordz.pl?keyword=automatus.

Wing, 2021, Wing, J. M. (2021). Trustworthy AI. Communications of the ACM, 64(10), 64–71. https:// doi.org/10.1145/3448248.

Verma, S., Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1–7. https://doi.org/10.1145/3194770.3194776.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.