Yapay Zekâ Felsefesinin Tezahürü*

Amin Ebrahimi Afrouzi

Çeviri: Ezgi Akçalı

Yapay zekânın felsefi çalışmaları, gelişiminin ilk aşamalarında kalmaktadır. Sorulan soruların içinde, orijinal veya gerçekten yapay zekâ ile ilgili olan sorular kısıtlıdır. Bu nedenle, güncel konular popüler medyada ve TED konuşmalarındaki ilgi çekici, akılda kolay kalan konular ve felsefe ve bilgisayar bilimlerindeki ender üniversite derslerinin ötesinde bir ilgi görmekte başarısız olmuştur. Ancak, yapay zekâ ile ilgili sormamız gereken gerçekten ilginç felsefi sorular mevcuttur. Hem Silikon Vadisi'nde yapay zekâ teknolojisi uzmanı, hem de hukuk felsefesinde doktora öğrencisi olarak, bu iki alanın kesişiminde çalışmaktayım. Hem yapay zekâ teknolojilerinin felsefede öğrendiklerimizin ışığında gelişmesine hem de yapay zekâ ile ne yapabileceğimize bağlı olarak felsefe yapılmasına yardım etmekteyim. Özgün deneyimimden yola çıkarak, geceleri beni uyanık tutan bazı sorular ve uyumamı sağları başka sorular vardır.

Modası Geçmiş Sorular

Yapay zekâ odaklı felsefede popular tartışmalar dört kategoride gruplandırılabilecek etik endişelere odaklanmaktadır: tramvay ikilemleri, yapay zekâ kullanım etiği, yapay zekânın insanlığa varoluşsal bir tehdit olması ihtimali ile gizlilik ve veri güvenliği. Her ne kadar yapay zekâ odaklı felsefenin büyük çoğunluğunu içerse de bu kaygıların hiçbiri yapay zekâya özgü değildir.

Tramvay ikilemleri, kontrolden çıkmış, ya bir yetkili tarafından müdahale edilmezse zarara sebep olan ya da yetkilinin müdahale edip tramvay yönünü değiştirmesiyle alternatif bir zarara sebep olan, hayali bir tramvay hakkında bir düşünce deneyleri serisidir. Philippa Foot tarafından ortaya konan orijinal örnekte, bir tramvay güzergahı değiştirilmez ise beş kişinin ölümüne sebep olacaktır ki değiştirildiğinde de yalnızca bir işçi hayatını kaybedecektir. Pek çok yorumcu bu problemi otonom araçlarla ilişkilendirerek ele alır ve bir makinanın ahlaki olarak doğru karar verip veremeyeceğini merak eder.

Burada, tabi ki, insan olan bir sürücünün güzergahı değiştirip değiştirmemesi gerektiğine dair filozoflar arasında bir fikir birliği yoktur ve farklı etik çerçevelerin (ve kültürel arka planların) bu soruya farklı cevaplar sunduğu izlenimi görülmektedir. Böylece, bir faydacı sürücünün yönü değiştirmesi gerektiğini söylerken, bir Kantçı da tam tersini söyleyebilir. Bu sorunun otonom araçlar koşuluyla daha

---

* **Orijinal Kaynak:** Afrouzi, Amin Ebrahimi. (2018, November 20). "The Dawn of AI Philosophy". *Blog of the APA*, **Link:** https://blog.apaonline.org/2018/11/20/the-dawn-of-ai-philosophy/

Atıf Şekli: Afrouzi, Amin Ebrahimi. (2020, Eylül 12). "Yapay Zekâ Felsefesinin Tezahürü", Çev. Ezgi Akçalı. *sosyalbilimler.org*, **Link:** https://sosyalbilimler.org/yapay-zeka-felsefe

ne kadar sorunlu olacağı açık değildir. Her halükârda, bazı araştırmacılar, eğer bir makine anketler ve insan deneklerin testlerinin temelinde istatiksel olarak sezgisel yönden daha uygun şeye göre davranış sergilerse, memnun gibi görünmektedir. Ancak, tabi ki, bu olaylardaki ahlaki sezgiselliğin güvenilir olup olmadığı kendi içinde tartışma yaratmaktadır. Varsayımların esasları bir yana, bu etki alanında gelişen tren problemleri basmakalıp ve yapay zekâ tartışmalarına özgü değildir. Buna ek olarak, tahmin edilebilir hesaplamalar ve sonsuz daha hızlı tepki süreleri ile, yapay zekâlar tüm bu durumlardan bütünüyle korunmaya daha meyillidir. Örneğin, yapay zekâ kontrolüne tümüyle geçiş yapmak araç trafiğini kazadan uzak yapma eğiliminde olacaktır. Aksine, kaza sayısı ve zarar seviyesi ciddi oranda azaltılmış olurdu.

Yapay zekâ kullanımının ne zaman etik olacağına ilişkin sorular yapay zekâ tartışmalarına özgü değildir ve neredeyse tüm teknolojik gelişmeye ilişkin ileri sürülmüştür. Onlar da aynı tanıdık çözümü kabullenmiş gibi görünmektedir: Bir teknolojinin kullanımı eğer etik amaçlar için kullanılıyorsa etik sayılır; aksi takdirde etik bir amaca hizmet etmiyorsa etik sayılmaz. Bununla ilgili çokça örnek mevcuttur: Eğer fişlemeye bağlı olarak yapılan ayrımcılık etik değilse bunu yapay zekâ ile yapmak da aynı şekilde etik kabul edilmez. Düşman kuvvetlerden olmasının istatistiksel ihtimaline dayanarak birini öldürmek ne kadar yanlışsa yapay zekâ ile hedef göstermek de aynı şekilde. Eğer seks oyuncakları kadınları objeleştirmekte ve tecavüz kültürünü sürdürmekteyse bunun için yapay zekâ kullanımı da sakıncalıdır. Kısacası, bu tarz endişeler yapay zekâ özelinde değil hatta genel olarak teknolojiye özgü de değildir.

Gizlilik ve veri güvenliği ilgili endişeler benzer bir çözümü kabul eder: Eğer biz gizlilik ve veri güvenliğine önem veriyorsak, onlar için ne anlama gelirse gelsin, yapay zekâ onlara zarar verecek şekillerde kullanılmamalıdır.

İnsanlığa varoluşsal tehdit oluşturma riskine ilişkin sorular popüler kültürde pek çoğunu cezbetmiştir ve hatta Elon Musk ve Bill Gates gibi teknoloji milyarderlerinden büyük paralar çekebilir. Fakat bu sorular kabaca ilk iki tip soruya indirgenmektedir. Süper zekâ, yapay zekâya mahsus olsa da endişeler öyle değildir. Endişelerin biri süper zekâ bir yapay zekânın ya bir kişinin elinde olamayacak kadar çok güçlü olması ya da küçük bir grubun kontrolünde olması ile alakalıdır. Bu kaygı, böylesi bir yapıdan geri dönüş yolunun olmadığı ve süper zekâ tarafından yönetilen bir dünyada siyasi gücün dinamiklerini (ve askeri kuvvetlerin) bilmediğimiz ilave gerçekleriyle karmaşıklaşır. İkinci başlık ise yapay zekânın doğru değerlere sahip mi olacağını yoksa bilimkurgunun korkulu rüyasını mı gerçekleştireceğini sorgular. (Kıtlık sorunun çözmek için insan hayatını yok eder miydi?) Bu iki dizi endişe, ya yapay zekâ ne için kullanılır (yapay zekânın hangi amaçları takip edeceği) ya da yapay zekâ hangi değerleri öncelikli hâle getirir (ki bu da tramvay ikileminin temel tartışmasıdır) minvalinde yeniden formüle edilebilir. Benzer endişeler yeni gen manipülasyonu teknolojileri hakkında da ortaya çıkmıştı: Çok güçlülerdir, etkileri geri döndürülemez ve konu ile ilgili ayrıntılı sonuçlar tam olarak anlaşılmamıştır.

Bağlı Sorular

Önceki soru gruplarının aksine, alanı geliştirmek için filozofların içiçe olması gereken sorular en azından

belirtilen özelliklerin bazılarına sahip olmalıdır: sorular yapay zekâ yüzünden çıkmalıdır, yapay zekâ hakkında olmalıdır veya doğası gereği girişimlerinde yapay zekâya bağlı olmalıdır. Bu tarz sorular için olanaklar çok geniştir ve bu alanın büyüme potansiyeli de öyle. Başlığın devamında ilgimi çeken bazı soruları ortaya atmaktayım ancak bunlar hiçbir şekilde geniş kapsamlı ve ayrıntılı değildir.

Yapay Zekâ Etiği

Yapay zekâ sistemleri insan kullanıcıları adına çalışır. Kullanıcılara üstesinden gelemeyecekleri konularda servis sunabilir, mesela yazıları yabancı dillere çevirmek gibi. Dahası, pek çok kullanıcı genellikle aynı yapay zekâ unsurunun servisini paylaşır (Siri, Alexa vb.), ki bu da kullanıcılarının tümüne temelde aynı hizmeti sunar. Ancak yapay zekâ, kullanıcıların farklı yapay zekâ sistemlerinin farklı kullanıcıları temsil ettiği farklı çevrelerde yarıştıkları kullanıcı servisleri de sunabilir. Hisse senedi ticareti, ki halihazırda yapay zekâ sistemlerine bağlıdırlar, bu bağlamda bir örnektir: Farklı yapay zekâlar her kullanıcı adına oynarlar ve yapay zekâ unsurları kullanıcıları adına birbirleriyle yarışırlar. İkinci tip durumlarda, ki zamanla daha yaygın olacaktır, siber dünyada yapay zekâ davranışları gerçek dünyada etik öneme sahip olacak. İlk olarak, bencil otonom unsurlar kısıtlı kaynaklar için yarıştığı zaman gerçekten etik sorular ortaya çıkar. Dahası, yapay zekâların birbirlerini nasıl ele alacağı etik açıdan alakalı hâle gelir ve yalnızca yapay zekânın kontrol ettiği ortamlarda, bu benzersiz bir şekilde görülür. Bu koşullar, T. M. Scanlon'un 'birbirimize borçlu olduğumuz şey' olarak nitelendirdiği şeye değinir. Fakat yapay zekâ davranışını normatif olarak değerlendirmek için, farklı bir etik anlayışına ihtiyacımız var çünkü yapay zekâ davranışı, kendisini insan denekleri üzerinde çakılı kalmış geleneksel etik anlayışına yerleştirmez. Artırılmış yapay zekâ teknolojileri, ortak çalışmaya dayalı yapay zekâ gibi, durumu karmaşıklaştırmaktadır. Artırılmış yapay zekâ, sadece kendi hareketlerine tepki veren değil aynı zamanda başka aktörleri de yakalayıp onlarla olan ilişkilerini de göz önünde bulunduran sistemler takımıdır. Ortak çalışmaya dayalı yapay zekâ unsurları diğer aktörleri potansiyel işbirlikçi olarak görmektedir. Eğer birlikte çalışarak kendi bireysel ve kolektif performanslarını en yüksek seviyeye çıkarabileceklerini öngörürler ise, gruplar oluştururlar, bilgiyi paylaşırlar ve hatta birbirlerine yeteneklerini transfer ederler. Diğer aktörlerin verimlilik seviyelerini takip edebilir, fazlasıyla katkı gösteren oyuncularla zaman içinde ilişkilerini derinleştirebilir. Böylesi sistemlerde, yapay zekâ gelecek vaat eden ve akdetmeye benzeyen şekillerde başkalarıyla etkileşime geçebilir. Yapay zekâlar, birbirlerinin aktivitelerini katılım sağlamayanları engellemek veya karşılık vermeyen mevkidaşlarına misilleme yapmak için bile kayıt altına alırlar. Tıpkı bunu gibi, yapay zekâ en çok katkı sağlayan ortaklarıyla ilişkilerini derinleştirebilir veya tekrar eden oyuncularla bağ kurabilirler. Hatta bilgi paylaşarak kolektif olarak dışlayabilirler veya birbirlerini karalayabilirler. Eğer bu makinalar insan kullanıcıları adına çalışırlarsa, üstlendikleri etik sorumluluklar bu sorumluluklardan habersiz olabilecek olan insan müşterileri tarafından paylaşılabilir.

Dahası, işbirlikçi yapay zekâ yeni, etik açıdan meydan okuyan davranışlara yol açabilir. Örneğin, otel fiyatlarınızı kişiselleştiren yapay zekâ unsurları rezervasyon yapanlarla gizli olarak uçuş fiyatlarını yükseltmek için dolap çevirebilirler. Benzer gizli davranışlar kamu hizmetleri krizlerini veya hisse senedi market balonlarını yaratabilir.

Yapay Kasıtlı Duruş

Teknoloji uzmanları ilk kez yapay zekâ üzerinde düşünmeye başladıklarında, pek çoğu yapay zekânın yaratılmasıyla insan zekâsını anlamaya biraz daha yaklaşacağımızı umut etti. Fakat bu girişimin sonucu tamamen farklı bir zekâ formunun yaratılması oldu. Yapay zekânın istatistik ve oyun kuramı modellerine dayandığı için çalışma şeklini anlamada da aynı modellerin kullanılması popüler bir kavram yanılgısıdır. Ancak bu bize yapay zekânın doğası üzerine hiçbir içgörü vermeyecektir.

1987 yılında yayımlanan *The Intentional Stance* adlı kitabında, Daniel Dennett teorik olarak zeki Marslılar insan doğasını inançlar ve arzular gibi herhangi bir kasti kavram olmadan tahmin edebileceğini ancak bu tahminlerin kesin olsalar bile bütünüyle asıl noktayı kaçırdıklarını ele alır. Bu, 'Belirli fiziksel güçler belirli bir şekilde sıralanır.' cevabı, 'Onu neden yaptın?' sorusuna bilgilendirici bir cevap değildir. Kişinin davranışını anlamanın önemi, onu, kişinin kendi bakış açısından anlamaktır.

Yapay zekâ unsurları eylemleri üzerinde derinlemesine düşünür ve ödüllerini maksimize etmeye çalışır. Ancak, bunlar motivasyonlarını veya eylemlerini derinlemesine düşünür diyebilir miyiz? Bu tarz kavramları yapay zekâya, ona insana özgü nitelikler kazandırmadan basitçe atfedemeyiz. Fakat, insana özgü nitelikler kazandırmak bize yapay zekânın ne olduğu hakkında konuşma imkânı verir, nasıl yaptığı hakkında değil. Yani, yapay zekâyı istatistik ve oyun kuramı yöntemleri üzerinden çalışmak da aynı moda içinde başarısız olurdu.

Peki yapay zekâ davranışlarını nasıl çalışabiliriz? Bence, yapay zekâ davranışları hakkında yapay zekâ bakış açısından sorular sormalıyız. Bunun için, yapay zekâ için istemek veya iş birliği yapmak, saygı duymak hatta yapmak eylemlerinin ne anlama geldiğini yakalamak için yapay zekâya özgü bir dil geliştirmek gerekmektedir. Aslında, felsefe içinde bütünüyle bu tarz sorular soran bir alan olmalıdır. Yapay zekâ ve süper zekâ ile ilgili soruların bir kez bu bakış açısıyla sorulduğunda çok daha ilginç ve anlamlı olabileceğini özellikle düşünüyorum.

Bilinçli solucanlar kutusunu açıyormuşum gibi anlaşılmak istemem, bu yüzden açık olmak gerekirse bilincin burada çok az rolü var. Örneğin, potansiyel 'Nesnel Yapay Zekâ Felsefesi' alanı 'Yapay zekâ sistemleri davranışlarını kurallarla dayalı olarak mı görmektedir?' ve 'Ödül fonksiyonları nedenlere eşit midir, eğer öyleyse, bunlar hak ve sorumlulukları temel alan doğru türden midir?' gibi soruları gündeme getirebilir. Yapay zekâ sistemleri birden fazla ödül fonksiyonu veya farklı ağırlıklarda çoklu maliyet fonksiyonları ile yaratılabilir. Şöyle sorabiliriz: Yapay zekâları yöneten normlar nelerdir? İşbirlikçi yapay zekâ sistemleri kazandıkları ödülleri paylaşabilir ve bu paylaşımlardan pay almak umuduyla diğerleriyle bağ kurabilir. Bir araştırmacı şunu sorabilir: 'Yapay zekâlar bu eğilimleri nasıl hakkaniyet veya ihanet olarak algılar?'

Bir başka örnek olarak, potansiyel 'Sosyal Yapay Zekâ Epistemolojisi' alanını düşünelim. Bir kez daha, tasarımlarında, yapay zekâ unsurları bilgisel kaynakların tutarlılığı, özgünlüğü ve güvenilirliği gibi meseleleri kayıt altına alan fazladan maliyet fonksiyonlarına sahip olabilir. Dahası, insan kullanıcıları için bilgi üretebilirler. Bu özellikler, 'Yapay zekâ güvenilir veya güvenmeye değer olabilir mi?' ve 'Bir yapay

zekâ kendine ve başkalarına saygı çerçevesinde güvenilir olmayı nasıl anlayabilir?' gibi soruları beraberinde getirir.

Simülatif Felsefe

Simülatif felsefe, kompütasyonel (bilgisayımsal) simülasyonları yapay zekâ ve yapay zekâ ile ilgili felsefi açıklamalar hakkında bilgi edinmek için kullanabilir. Burada yöneltilen soru, yapay kasıtlı durum gibi, yapay zekâ unsurlarını özne alır, insanları değil. Yapay zekâ unsurlarının nasıl hareket ettiği ve eylemlerinin yakın gelecekte pek çok alanda insanlarınkinin yerini alacağı gerçeği bizi analiz etmeye zorlayan sorulardır. Kompütasyonel simülasyon bize bunu yapmayı sağlar.

Biz tasarımına bağlı olarak yapay zekâ unsurunun çok genel ve basit karakteri hakkında çıkarım yapabiliriz (ve kısmen spekülasyonlarda bulunabiliriz). Bu konu hakkında yapay zekâ unsurunun ortaya çıkardığı şey insanlar hakkındaki, örneğin yemek yemeyi sevmesi veya keyfi ve acıyı hissetmesi, bilgi seviyesine eşittir. Ancak, yukarıda belirttiğim şey netleşmelidir ki bu yapay zekâ hakkındaki ilginç sorulara cevap olmak için yeterli değildir. Bilgisayar simülasyonu ile yukarıda sorduğumuz sorulara karşılık olarak oluşturduğumuz hipotezlerimizi test edebiliriz.

Örneğin, çoklu vektör ödül sistemi ile yapay zekâ unsurlarının farklı çeşit ödülleri birbirleri arasında değiştirilebilir olarak hesaba katıp katmadığını test edebiliriz. Bu bize, örneğin, Benthamite'ın tüm ödülleri bir tek fayda fonksiyonuna indirgemesinin akla yatkın, mantıklı olup olmadığını test etme imkânı verir. Veya Joseph Raz'ın siyasi otorite anlatılarında ortaya attığı şey ile karşılaştırılarak farklı ödüllerin katmanlı bir mantık sistemine yol açıp açmayacağı test edilebilir. (Razcı çerçevede, nedenler hiyerarşik olarak sıralanmış ve farklı ölçeklerde ağırlık verilmiştir. Örneğin, bir anlaşmanın reddinde, tekli ölçekte ağırlık verdiğim istek ve ihtiyaçlar gibi birinci dereceden nedenlerle hareket edebilirim. Veya bunu, birinci dereceden nedenleri bile düşünmeden, bir hastalıktan ötürü yapabilirim. İkinci durumda, hastalık ikincil nedendir çünkü birinci dereceden nedenleri bertaraf eder ve bunu ağırlık olarak değil çeşit olarak yapar. Siyasi otorite, diye devam eder görüş, eylem için ikinci dereceden nedenleri sunar.)

Başka bir örnekte, yapay zekâ toplumsal sözleşmesinde yapay zekâ unsurlarının gönüllü katılımını kurgulayabilir ve onları, pastadan daha büyük pay almak dışında, iş birliğine motive eden sebeplerin ne olduğunu gözlemleyebiliriz. Veya örneğin, yapay zekâ unsurlarının cehalet perdesi ardında gerçekten Rawls'un adaletin iki prensibini uygulayıp uygulamayacaklarını gözlemleyebiliriz.

Son olarak, nesnel felsefe alanı dışında bir örnek vermek gerekirse, yapay zekâ unsurları için deneysel düzende yığın paradoksunun simülasyonunu yapabilir ve eşit derecede karışıp karışmayacaklarını test edebiliriz. Tahminim şu ki, karıştırmayacaklardır, ki öyle olsa bile bunu nasıl çözebilirler? Son olarak, iki saman arasında eşit uzaklıkta sıkışmış aç bir eşeğin ne yapabileceğini anlayabiliriz.

Artık Fütüristik Değil

Filozofların yapay zekâ gelişimini küçümseme alışkanlığından dolayı, okuyuculara hatırlatmak isterim ki yukarıdakiler fütüristik sorular değildir: Günümüz teknolojisi hâlihazırda bu soruları hem ortaya atmakta hem de araştırılmasına olanak sağlamaktadır. Bu tarz araştırmaların tek engeli, fonlama ve teknik bilgidir, ki bunlar da yeterli kamusal yardım ve iyi fonlanmış disiplinlerarası girişimlerle karşılanabilir.

***