Yapay Zeka Güvenliği Zirvesi'den Beklentiler

Öncü yapay zeka modellerinin ve bu modelleri geliştiren şirketlerin hayatımıza girmesiyle birlikte, yapay zekanın gelecekteki potansiyel etkileri, hatta şu an güncel olarak gözlemleyebildiğimiz sonuçları, yapay zeka merkezli tartışmaları özellikle son bir yılda oldukça popülerleştirerek, bu tartışmaları ağırlıklı olarak politika yapıcıların, araştırmacıların ve akademisyenlerin üzerine eğildiği bir konu olmaktan çıkartarak gündemin önemli bir parçası haline getirmiştir. Bu durumun doğal bir sonucu olarak ise, küresel çapta, öncü yapay zeka modellerinin olası yıkıcı etkilerinin önüne geçmek ve bu teknolojilerin, teknoloji şirketleri tarafından hukuka uygun ve sorumlu bir şekilde geliştirilmesi için gerekli regülasyonların ve politikaların oluşturulmasına yönelik girişimler hızlı bir şekilde önem ve ivme kazanmaktadır. Ekim ayında Birleşmiş Milletler yapay zeka teknolojilerinin küresel yönetişimi ve olası güvenlik risklerini gidermeye yönelik ortak bir yaklaşım benimsenmesine katkı sağlamak için farklı alanlardan 32 uzmanı bir araya getiren disiplinlerarası Yapay Zekaya İlişkin Üst Düzey Danışma Organı'nı oluşturmuştur. Aynı zamanda Avrupa Birliği, güncel olarak bir Yapay Zeka Yasası üzerinde çalışmaktadır ve bu yasa üzerinde 2023 yılının sonlarına doğru bir anlaşmaya varılması beklenmektedir.

Yapay zeka teknolojilerini regüle etme ve bu teknolojilerin ekonomik, sosyal ve hukuki boyutlarını göz önünde bulunduran kapsamlı bir yaklaşım benimseme girişimleri için ise içinde bulunduğumuz hafta özel bir önem taşımaktadır. 30 Ekim 2023'te ABD'nin yapay zeka teknolojilerinin güvenli ve güvenilir olmasını ve bu doğrultuda yeni standartlar oluşturmayı hedefleyen bir kararname yayınlaması ve yine aynı gün G7'nin özellikle üretken yapay zeka(generative ai) modellerine odaklanan bağlayıcı olmayan bir yönetmelik üzerinde anlaşmaya varması bu hafta içerisinde tanık olunan önemli gelişmelerdir. Ancak Türkiye'nin de Sanayi ve Teknoloji Bakanlığından Bakan Yardımcısı seviyesinde iştirak sağladığı 1-2 Kasım tarihlerinde Londra'da düzenlenecek ilk küresel Yapay Zeka Güvenliği Zirvesi ayrı bir önem taşımaktadır.

Zirve'nin Konusu, Önemi ve Amaçları

Yapay Zeka Güvenliği Zirvesi'nin farkı, yukarıda bahsedilen uluslararası kurumların, platformların ve devletlerin girişimine kıyasla ABD, AB üye ülkeleri, Körfez Ülkeleri ve Çin'den çeşitli üst düzey siyasi liderleri ve karar alıcıları lider teknoloji şirketleri yöneticileri ile bir araya getiren çok paydaşlı bir yapıya sahip olması, hem özel sektörü hem de yapay zeka teknolojilerinin geliştirilmesinde önemli role sahip ülkeleri bir araya getirmesidir. Zirveye katılacak yaklaşık 100 kişi arasında ABD Başkan Yardımcısı Kamala Harris, Avrupa Komisyonu Başkanı Ursula Von der Leyen, Elon Musk, Microsoft'tan Brad Smith, OpenAI'ın yöneticisi Sam Altman, Google DeepMind'ın yöneticisi Demis Hassabis, Çin teknoloji devleri Alibaba ve Tencent'in temsilcileri ve Çin devlet yetkililerinin olması beklenmektedir. Yapay Zeka Güvenliği Zirvesi'nin bu özelliği, Zirve'de üzerinde fikir birliğine varılan konuların ve yaklaşımların hem şirketler hem de devletler tarafından benimsenmesi, belli hususlarda uluslararası bir konsensus oluşturulmasının önünü açma ve bu zirvede alınan kararların bir emsal teşkil etmesi ihtimalini artırmaktadır. Dolayısıyla, önümüzdeki yıllarda hem günlük yaşamda hem de ekonomi ve güvenlik alanında etkisini büyük oranda hissettirecek yapay zeka teknolojilerinin geliştirilme sürecini ve bu alandaki regülasyonları küresel olarak etkileme potansiyeline sahip bu zirvenin içeriği, kapsamı ve ulaşmayı hedeflediği amaçlar oldukça önemlidir.

Birleşik Krallık Hükümeti'nin Yapay Zeka Güvenliği Zirvesi ile ilgili yayınladığı belgelere göre, bu Zirve genel hatlarıyla, yapay zeka teknolojilerindeki ilerlemelerin hızı ve yapay zeka teknolojilerinin farklı kullanılış biçimleri ve verilerle etkileşimi sonucu, öngörülmesi oldukça zor tehditler ve riskler oluşturması sebebiyle, bu risklere karşı alınacak acile önlemlere yönelik uluslararası bir tartışma başlatmayı hedeflemektedir.

Her ne kadar dar yapay zeka (narrow AI) modellerinin de belli koşullar altında ciddi zararlar verme kapasitesine sahip olduğu kabul edilse de, öncü yapay zeka (frontier AI) modelleri Zirvenin gündeminde daha fazla yer tutacaktır. Dar yapay zeka modelleri sadece belirli bir işi ve görevi yerine getirmesi için geliştirilmiş yapay zekalar olarak tanımlanabilir. Google Translate, satranç oynayabilen botlar, yüz ve ses tanıma uygulamaları kısıtlı yapay zeka modellerine örnek verilebilir. Öte yandan öncü yapay zeka modelleri ise en gelişmiş modeller olan genel amaçlı, çeşitli görevleri yerine getirebilen çok işlevli yapay zekalar olarak tanımlanmaktadır. Öncü yapay zeka modelleri geliştiren şirketlere örnek olarak ise OpenAI, Google DeepMind ve Anthropic verilebilir. Öncü

yapay zeka modellerinin çok daha karmaşık görevleri yerine getirebilme ve çok çeşitli amaçlarla kullanılabilme özellikleri, bu modellerin oluşturabileceği tehditleri ve verebileceği zararları çok daha öngörülemez hale getirdiği ve bu zararların boyutunu artırdığı için Zirve'de öncü yapay zeka modelleri tartışmaların odağı olacaktır.

Benzer bir doğrultuda, Zirvede iki büyük risk grubu tartışmaların merkezinde olacaktır. Bunlardan ilki kötü niyetli aktörlerin öncü yapay zeka modelleri yardımıyla can kaybı da dahil olmak üzere, geniş çaplı zararlar vermesi, yeni teknolojiler geliştirmesi ya da kritik öneme sahip altyapılara müdahale edip zarar vermesi gibi örnekler verilebilecek olan kötüye kullanım riskidir. Zirvenin gündemindeki ikinci risk grubu ise geliştirilen öncü yapay zeka sistemlerinden kaynaklı oluşabilecek olan kontrolün kaybedilmesi riskidir. Bunların haricinde öncü yapay zeka modellerinin dezenformasyon için güçlü bir silah olarak kullanılabilme ihtimali, ya da bu sistemlerin yol açabileceği kitlesel işsizlik gibi sosyal ve ekonomik boyutları bulunan riskler bu Zirvenin kapsamında değildir. Bu sorunların Zirvenin kapsamı dışında bırakılmasının temel sebebi olarak ise, bahsedilen geniş çaplı toplumsal risklerin halihazırda var olan uluslararası süreçler ve devletlerin kendi enstrümanları ile çözülmesinin mümkün olması gösterilmektedir.

Yapay Zeka Güvenliği Zirvesi'nin, öncü yapay zeka modellerinin oluşturabileceği kötüye kullanım ve kontrol kaybı riskleri üzerine yoğunlaşarak ulaşmayı amaçladığı 5 ana hedef bulunmaktadır:

- Öncü yapay zeka modellerinin oluşturduğu riskler ve bu konuda harekete geçmenin gerekliliği üzerinde ortak bir anlayışa ulaşmak.
- Ulusal ve uluslararası politikaların ve çerçevelerin en iyi nasıl desteklenebileceği de dahil olmak üzere yapay zeka güvenliği konusunda uluslararası işbirliği için ileriye dönük bir süreç başlatılması.
- Kurum ve kuruluşların öncü yapay zeka güvenliğini artırmak için alması gereken uygun önlemlerin belirlenmesi ve tartışılması.
- Yapay zeka güvenliği araştırmalarında ve yapay zeka yönetişimini destekleyecek yeni standartların oluşturulmasında potansiyel işbirliği alanlarının belirlenmesi.

• Yapay zekanın sorumlu ve güvenli bir şekilde geliştirilmesinin teminat altına alındığını, yapay zekanın küresel çapta iyi amaçlar için kullanılmasını sağlayacağını göstermek.[1]

Zirveye Yönelik Eleştiriler

Londra'da gerçekleşmekte olan Yapay Zeka Zirvesi, bu alanın lideri olan şirketleri, devletleri ve organizasyonları ortak bir platformda buluşturup, küresel bir işbirliği ve ortak bir yaklaşım için önemli fırsatlar barındırsa da, çeşitli uzmanlar tarafından önemli eleştirilerin hedefi olmuştur. Zirvenin anlamlı ve faydalı sonuçlara ulaşmasının pek mümkün olmadığı yönünde kaygılar dile getirilmektedir.

Ana eleştirilerden birincisi, Zirvenin gündeminin merkezine öncü yapay zeka modellerini ve henüz geliştirilmemiş olan teknolojileri alması ile ilgilidir. Zirvede ele alınacak konuların güncel sorunları ve güncel yapay zeka teknolojilerinin oluşturduğu riskleri kapsamaması eleştirilmektedir. Yapay zeka güvenliği ile ilgili küresel liderlerin bir araya geldiği ve bunun sonucu olarak önemli bir ivmenin kazanılabileceği bu toplantıda yakın zamanda çözülebilecek ve önlem alınabilecek sorunların ve risklerin gündeme alınmamasının Yapay Zeka Güvenliği Zirvesi'nin potansiyelini heba ettiği ileri sürülmektedir[2].

İkinci kaygı ise, katılımcıların ağırlıklı olarak siyasetçiler, devlet yetkilileri ve teknoloji devleri olması, Zirve'nin sivil toplum kuruluşlarına ve araştırmacılara gereken önemi göstermeyerek, şirketlerin Zirve'nin ajandasını domine etmesine izin vermesidir. Bu kaygı ile paralel olarak çeşitli düşünce kuruluşları Zirvede benimsenen, devlet regülasyonlarını geri plana atan ve şirketlerin kendi özdenetim standartlarını belirlemesini merkeze alan yaklaşımın, yapay zeka teknolojilerinin yaratacağı risklerle baş edebilmek için oldukça yetersiz olduğuna dikkat çekmekte ve inisiyatifi şirketlere veren özdenetim modelinin uygulanmasının 2010'ların başında sosyal medya

[1] "AI Safety Summit: Introduction (HTML)," GOV.UK, accessed November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.
[2] Kurt Robson, "Does the UK's AI Safety Summit Have Its Priorities Right?," *Verdict* (blog), October 31, 2023, https://www.verdict.co.uk/does-the-uks-ai-safety-summit-have-its-priorities-right/.

platformlarının regülasyonunda benimsenen yaklaşımlarda yapılan hataların tekrarlanmasına yol açacağını savunmaktadırlar[3].

Zirveden Sonra

Londra'daki yapay zeka güvenliğini konu alan iki gün sürecek olan bu Zirvenin bir dönüm noktası olup olmayacağını zaman gösterecektir. Her ne kadar yukarıda özetlenen eleştiriler, Zirve'nin etkinliği ve somut etkisi ile ilgili şüpheleri güçlendirse de, hızla gelişmekte olan yapay zeka teknolojilerinin küresel çapta oluşturabileceği riskleri gündem edinen Çin'i AB'yi ABD'yi ve teknoloji devlerinin birlikte hareket etmesini kolaylaştırma amacı taşıyan bir üst seviyeli toplantının düzenlenmesi, azımsanacak bir başarı değildir. Ayrıca Yapay Zeka Güvenliği Zirvesi, iklim değişikliği ile mücadeleye yönelik oluşturulan BM Taraflar Konferansı (UN COP) gibi önemli bir örnek teşkil ederek yapay zeka alanında uzun vadeli bir işbirliği çerçevesi oluşturma potansiyelini barındırmaktadır.

[3] "Artificial Intelligence for Public Value Creation: Introducing Three Policy Pillars for the UK AI Summit," IPPR, October 25, 2023, https://www.ippr.org/research/publications/ai-for-public-value-creation.