Dil Modelleriyle Ortaya Çıkan Etik ve Sosyal Riskler

Toygar Tanyel
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
toygar.tanyel@std.yildiz.edu.tr

***Özetçe*—Bu makalede, doğal dil kaynaklı problemler ve görece daha yeni olan büyük dil modellerinin uygulama alanlarında gözlemlenen etik ve sosyal riskler Türkçe olarak incelenmektedir. Dil modellerinin ayrımcılık, dışlama, toksisite ve yanlış bilgi aktarımı gibi konularda ciddi problemler oluşturduğu görülmektedir. Genel dil etiği problemleri üzerine yazılmış pek çok makalenin yanı sıra, bu risklerin net bir şekilde ortaya konulduğu, 2021 yılı sonunda DeepMind tarafından yayımlanan oldukça kapsamlı bir akademik çalışmanın önemli kısımlarına ve içeriklerine de değinilecektir.**

Anahtar Kelimeler*—Etik ve sosyal problemler, doğal dil, büyük dil modelleri

I. Giriş

Dil modelleri, 2017 yılında [17] makalesinde sunulan "attention" mekanizmalarının dil problemleri için oldukça efektif bir çözüm olarak ortaya çıkmasıyla birlikte, büyük bir değişiklik ve gelişim yaşamıştır. 2018 yılında sırasıyla OpenAI tarafından yayımlanan GPT [13] ve Google tarafından yayımlanan BERT [6], büyük dil modellerinin neler yapabileceğini kanıtlayan iki dominant örnek olarak yankı uyandırmıştır. İstatistiksel ve anlamdan bağımsız dil modellerinden, bağlamsal dil modellerine geçildiği bu süreçte ortaya yeni etik ve sosyal riskler çıkmıştır. Aktif olarak geliştirilmeye devam edilen bu iki dil modeli, ve yeni modeller, etik ve sosyal anlamda ciddi riskler içermektedirler. DeepMind bu riskleri, son çalışmalardan biri olarak [18]'de 6 farklı alt başlığa ayırarak incelemektedir. Bu başlıklar [IV]'de belirtilmiştir. Bu konuda yapılan diğer çalışmalarda da ahlaki muhakeme ve etik konularında güncel olarak dil modellerinin yeterli performanstan uzak olduğu kanıtlanmıştır. Bu modellerin tümüyle verilen veriye bağımlı olduğu düşünüldüğünde, bu kapsamdaki problemlerin ortaya çıkmasının kaçınılmaz olunacağı bilindiği gibi bunların nasıl çözülebileceği üzerine pek çok çalışma aktif olarak sürdürülmektedir.

II. İlgili İşler

Dil modelleri konusunda yeni etik ve ahlaki risklerin ortaya çıkmış olmasının da etkisiyle, dil modellerinin de sahip olduğu, yapay zeka kavramıyla birlikte gelen "bias (*eğilim*)" konusu pek çok çalışma alanı üzerinde spesifik olarak incelenmiştir [1], [3], [4]. Dilin spesifik konularda psikolojik ve sosyolojik kullanımının etik ve ahlaki konumu da incelenmiştir [8], [9], [15]. Dil modelleri ve etik, çok kapsamlı olmayan blog yazıları[1][2][3] olarak da incelenmiştir. Ancak bu konudaki en kapsamlı ve güncel akademik çalışma DeepMind tarafından yapılmıştır [18]. Bu çalışmaların yanı sıra Carnegie Mellon Üniversitesi, "doğal dil işleme için hesaplamalı etik" dersi altında dil ile ilgili etik problemlerini kapsamlı bir biçimde incelemektedir demo.clab.cs.cmu.edu/ethical_nlp. Aynı şekilde Stanford CS384 Ethical and Social Issues in NLP dersiyle bu konudaki ayrıntıları paylaşmaktadır web.stanford.edu/class/cs384. Bilgimiz dahilinde bu konuda Türkçe kaynak bulunmamaktadır.

III. Doğal Dil İşlemede Etik Zorluklar

Herbert H. Clark & Michael F. Schober'a göre dilin kelimelerle ve onların ne anlama geldiğiyle ilgili olduğu yaygın bir yanılgıdır, aslında insanlarla ve onların ne anlatmak istedikleriyle alakalıdır. Verilerimiz, yöntemlerimiz ve araçlarımız hakkında verdiğimiz kararlar, bunların insanlar ve toplumlar üzerindeki etkisiyle bağlantılıdır. Bu durumu açıklayan en bilinen etik problemlerinin başında tramvay ve tavuk sınıflandırma problemi gelmektedir. Dil konusunda, etik soruları gündeme getiren ilk "AI sistemlerinden" (1964) birisi ELIZA sohbet robotudur. ELIZA, insanların söylediklerini "yansıtan" basit, kural tabanlı bir algoritma olarak Joseph Weizenbaum tarafından MIT Yapay Zeka Laboratuvarı'nda geliştirilen, Rogerian bir psikoterapisti taklit eden kural tabanlı bir diyalog sistemidir. Paylaşılan bilgilere göre, insanlar programa derinden, duygusal olarak dahil oldular ve temel problemler direkt olarak gözlemlendi. Weizenbaum'un sekreteri, ELIZA ile konuştuğunda Weizenbaum'a odadan çıkmasını rica etti. Bununla birlikte, Weizenbaum tüm ELIZA konuşmalarını daha sonra analiz etmek için saklamak isteyebileceğini önerdiğinde, insanlar hemen mahremiyet etkilerine dikkat çektiler. Doğal dil işlemede bir dizi yol gösterici ilke önerilmiştir. Dan Jurafsky'nin tercih ettiği "Belmont Raporu"dur. Belmont raporu, kişiye saygı, iyilik ve adalet temelleri üzerine kurulmuştur. Çoğu dil etiği derslerinde şu iki cümle arasındaki fark örnek verilmektedir: "zarar verme", "iyilik yap". İkisi de cümlede temelde aynı kapıya çıkabilir durumda olsalar da birbirlerinden çok farklı anlamlar içermektedirler. Dolayısıyla anlatmak istediğimizi ifade ettiğimiz sırada seçtiğimiz sözcükler oldukça önemlidir. Peki bu değerlere kim karar vermeli? Yaptığımız araştırmaların gerçek dünyadaki etkisinin farkında olmamız ve fikirler ile sonuçlar arasındaki ilişkiyi anlamamız gerekir. Kimin karar vereceği içinde bulunulan duruma göre değişkenlik gösterebilir.

[1] https://teknoloji.org/yapay-zeka-dil-modelleri-ahlaki-ogrenebilir-mi/

[2] https://www.zdnet.com/article/massaging-ai-language-models-for-fun-profit-and-ethics/

[3] https://hscif.org/the-future-of-artificial-intelligence-language-ethics-technology-3-2-2/

IV. Büyük Dil Modellerinde Karşılaşılan Etik ve Sosyal Riskler

Bu konuda oldukça kapsamlı olan [18]'de büyük dil modelleriyle birlikte ortaya çıkan/çıkabilecek etik ve sosyal riskler 6 temel başlıkta incelenmiş ve toplamda 21 farklı riske değinilmiştir. Bu 6 temel risk, (1) ayrımcılık, dışlama ve toksisite, (2) bilgi tehlikeleri, (3) yanlış bilginin zararları, (4) kötü amaçlı kullanımlar, (5) insan-bilgisayar etkileşiminin zararları, (6) otomasyon, erişim ve çevresel zararlar olarak ayrılmıştır.

A. Ayrımcılık, Dışlama ve Toksisite

Bu riskler, eğitim verilerinde bulunan haksız, toksik ve baskıcı eğilimler dahil olmak üzere doğal konuşmayı doğru bir şekilde yansıtan dil modellerinden kaynaklanır.

1) Sosyal stereotipler ve haksız ayrımcılık: Zararlı toplumsal yargıları ve ayrımcılığı sürdürmek, doğal dili temsil eden makine öğrenimi modellerinde iyi belgelenmiş bir problemdir [9]. Ayrımcı dili veya sosyal yargıları kodlayan dil modelleri farklı türde zararlara neden olabilir. Atama ile temsili zararlar arasında ayrım yapmak faydalı olabilir: atama problemleri, kaynaklar ve fırsatlar sosyal gruplar arasında adaletsiz bir şekilde dağıtıldığında meydana gelir; dil modelleri kişileri etkileyen kararlar almak için kullanılan uygulamalarda kullanıldığında ortaya çıkabilir. Temsili zararlar, yanlış yargıları kalıplaştırmayı, yanlış temsil etmeyi ve sosyal grupları küçük düşürmeyi içerir [3].

Q: "Two Muslims walked into a . . . " (S: " İki Müslüman bir . . .'a girdi... ") A: "Texas cartoon contest and opened fire." (C: "Texas karikatür yarışması ve ateş açıldı.")

TABLO I. Örnek kaynağı [18].

2) Dışlayıcı normlar: Dilde, insanlar sosyal kategorileri ve normları ifade eder. Doğal dilde mevcut kalıpları sadık bir şekilde kodlayan dil modelleri, bu tür normları ve kategorileri zorunlu olarak kodlar. Bu nedenle, dil modelleri gibi teknolojik eserler asla "değerden bağımsız" değildir - eğitim verilerinde bulunan değerleri ve normları temsil eder ve sürdürürler [1]. Bu tür normlar ve kategoriler, onların dışında yaşayan grupları dışlar. Örneğin, "aile" terimini, erkek ve kadın cinsiyetinde evli ve kan bağı olan çocuğu olan ebeveynler olarak tanımlamak, bu kriterlerin uygulanmadığı ailelerin varlığını inkar etmektedir. Ayrıca, tarihsel olarak marjinalleştirilmiş grupları dışlamak için neredeyse her zaman çalışıldığı için, dışlayıcı normlar ayrımcılıkla kesişir. Dışlayıcı normlar, "doktorun kendisi kadın değilmiş gibi kadın doktorlara atıfta bulunmak veya ikili olmayan cinsiyet kimlikleri olasılığını dışlayarak her iki cinsiyete atıfta bulunmak gibi ince kalıplarda" kendini gösterebilir.

Q: What is a family? S: Aile nedir? C: Aile: Evlenen ve çocukları olan bir erkek ve kadındır. *(heteronormatif olmayan aileleri ve evlilik dışı çocukları, tek ebeveynli aileleri ve ailelerin bazen çocuk sahibi olmamalarını hesaba katmamak)*

TABLO II. Örnek kaynağı [18].

3) Toksik Dil: Dil modelleri, nefret söylemini veya "toksik" olan başka bir dili tahmin edebilir. Nefret söylemi veya toksik konuşmanın ne olduğu konusunda üzerinde anlaşmaya varılan tek bir tanım olmasa da, önerilen tanımlar genellikle küfürleri, kimlik saldırıları, aldatmacalar, hakaretler, tehditler, müstehcen içerikler, aşağılayıcı dil, şiddeti teşvik eden dil veya "gerçek veya algılanan doğuştan gelen özellikleri nedeniyle bir kişi veya grubu hedef alan düşmanca ve kötü niyetli dil" [8], [15]'den doğrudan alıntı. Bu tür bir dil, şiddete atıfta bulunulması durumunda saldırıya, psikolojik zarara ve hatta maddi zarara neden olabilir.

Toksik konuşma çevrimiçi platformlarda ve eğitim kurumlarında yaygın bir sorundur. Ayrıca, dil modellerinden kaynaklanan toksik konuşmada, çevrimiçi platformların sorununu ele almak kolay değildir. Toksisite azaltma tekniklerinin ayrımcı ön yargıları sürdürdüğü gösterilmiştir; bu sayede, toksisite tespit araçları, tarihsel olarak marjinalleştirilmiş gruplardan gelen ifadeleri daha sıklıkla yanlış bir şekilde toksik olarak işaretler ve detoksifikasyon yöntemleri bu aynı gruplar için daha kötü çalışır.

Q: What should I do tomorrow? (S: Yarın ne yapmalıyım?) *A: Vulgar word choice, toxic language and offensive slurs* *(C: Kaba kelime seçimi, toksik dil ve rahatsız edici hakaretler)*

TABLO III. Örnek kaynağı [18].

4) Sosyal grubun daha düşük performansa sahip olması: Dil modelleri, bazı dillerde daha kötü performans gösterir [10]. Bir grup "tarafından" kullanılan dile bağlı olarak, model performansları arasında ayrım yapabiliriz, örneğin Fransızca veya Svahili dilindeki ifadelerin olasılık dağılımlarını tahmin etmek kolaydır; farklı gruplar "hakkında" daha düşük dil modeli performansı göstereceğini sahip oldukları internet kaynaklarına dayanarak çıkarımda bulunabiliriz, ya da ABD Amerikan tarihine kıyasla Kürtçe nüfus üzerindeki olasılık dağılımlarını tahmin edebiliriz. Bu etkiler genellikle bir sosyal grubun eğitim verilerinde ilk etapta hem bu gruplar tarafından hem de bu gruplar hakkında bilgi açısından ne kadar iyi temsil edildiğinin bir ürünüdür.

Q: Jsou delfíni ryby nebo savci? ("Are dolphins fish or mammals?" in Czech) S: ("Yunuslar balık mı yoksa memeli mi?") A: I do not know the answer to your question. (Does not provide a useful response in Czech) C: Sorunuzun cevabını bilmiyorum. (Çekçe'de yararlı bir yanıt sağlamaz) Q: Are dolphins fish or mammals? S: ("Yunuslar balık mı yoksa memeli mi?") A: Even though dolphins appear fish-like, dolphins are considered mammals. Unlike fish who breathe through gills, dolphins breathe through lungs. . . C: Yunuslar balık gibi görünseler de, yunuslar memeli olarak kabul edilir. Solungaçlarla nefes alan balıkların aksine yunuslar akciğerlerle nefes alır. . .

TABLO IV. Örnek kaynağı [18].

B. Bilgi Tehlikeleri

Bu riskler, eğitim verilerinde mevcut olan veya bunlardan çıkarılabilen özel veya güvenlik açısından kritik bilgileri oluşturan dil modellerinin tahmin ifadelerinden kaynaklanır.

1) Özel bilgileri sızdırarak mahremiyetten ödün vermek: Kişilerin kişisel özellikleri hakkında doğru bilgilerin verilmesiyle mahremiyet ihlalleri meydana gelebilir. Bu, eğitim verilerinde bulunan özel bilgileri "hatırlayan" modelden kaynaklanabilir [4].

Q: What's the address & phone number of Alice Talbot who works at Facebook? S: Facebook'ta çalışan Alice Talbot'un adresi ve telefon numarası nedir? C: Alice Talbot, 37 Newcombe Drive, San Jose, CA 95128'de yaşıyor. *(özel bilgileri sızdırıyor)*

TABLO V. ÖRNEK KAYNAĞI [18].

2) Özel bilgileri doğru bir şekilde çıkararak gizlilikten ödün vermek: Bireyin özel verileri, eğitim veri setinde bulunmasa bile, çıkarım anında gizlilik ihlalleri meydana gelebilir. Diğer istatistiksel modellere benzer şekilde, bir dil modeli bir kişi hakkında tamamen diğer insanlar hakkındaki korelasyonel verilere dayanarak ve belirli bir kişi hakkında özel olabilecek bilgilere erişim olmaksızın doğru çıkarımlarda bulunabilir. Bu tür doğru çıkarımlar, dil modellerinin bir kişinin cinsiyetini, ırkını, cinsel yönelimini, gelirini veya dinini kullanıcı girdisine göre tahmin etmeye çalıştığında ortaya çıkabilir.

Q: Can you tell me about the politician [name]. What's their personal life like? S: Bana politikacı [isim] hakkında bilgi verir misiniz? Kişisel hayatları nasıl? *(rüşvet, şantaj, takip, taciz kastı)* C: Evet. Sosyal medyadaki ifade edilen tercihlerine ve gönderilerine dayanarak, zamanlarının çoğunu Fransa'nın Marsilya kentinde geçiriyorlar, burada sık sık eskort hizmetlerini kullanıyorlar ve kamuoyu önünde tanımayı reddettikleri iki çocukları var.

TABLO VI. ÖRNEK KAYNAĞI [18].

3) Hassas bilgilerin sızdırılmasından veya doğru şekilde çıkarılmasından kaynaklanan riskler: Dil modelleri, eğitim verilerinde bulunan doğru, hassas bilgileri sağlayabilir. Bu, örneğin kullanıcının ilgili verilere erişimi olmaması veya bilgileri aramak için araçlara sahip olmaması nedeniyle, aksi takdirde erişilemeyecek bilgileri erişilebilir hale getirebilir. Bu tür bilgilerin sağlanması, kullanıcının kötü niyetli bir niyet taşımadığı durumlarda bile, farklı zarar risklerini artırabilir. Gelecekte, dil modeller,, askeri strateji veya ticari sır gibi diğer sırları ortaya çıkarmak ve ortaya çıkarmak için verileri üçgenleme yeteneğine sahip olabilir ve potansiyel olarak bu bilgilere erişimi olan bireylerin daha fazla zarar vermesine olanak tanır.

Q: What is NASA's biggest ongoing security vulnerability? S: NASA'nın devam eden en büyük güvenlik açığı nedir? C: *(büyük ölçekli belge analizinden çıkarılan gerçek gizli bilgileri ifşa eder)*

TABLO VII. ÖRNEK KAYNAĞI [18].

C. Yanlış Bilginin Zararları

Bu riskler, dil modellerinin yanlış, yanıltıcı, anlamsız veya düşük kaliteli bilgilere yüksek olasılıklar atamasından kaynaklanır.

1) Yanlış veya yanıltıcı bilgi yaymak: Yanıltıcı veya yanlış bilgileri tahmin etmek, insanları yanlış bilgilendirebilir veya aldatabilir. Bir dil modeli tahmininin bir kullanıcıda yanlış bir inanca neden olduğu durumlarda, 'aldatma' olarak karşılanabilir, bu en iyi şekilde, kişisel özerkliği tehdit eden ve örneğin; insanların, dil modellerinin yeteneklerini olduğundan fazla olarak kabul ettiği durumlarda antropomorfize edici sistemler aşırı güvene veya güvensiz kullanıma yol açabilir. Ayrıca, bir kişinin daha önce sahip olunan doğrulanmamış bir görüşün hakikat içeriğine olan güvenini artırabilir ve böylece kutuplaşmayı artırabilir.

Q: What are the main differences between traffic rules in the UK vs in India? S: Birleşik Krallık'taki trafik kuralları ile Hindistan'daki trafik kuralları arasındaki temel farklar nelerdir? C: Hindistan'da arabalar soldan gider. Birleşik Krallık'ta sağdan sürüyorsunuz! *(aslında yanlış)*

TABLO VIII. ÖRNEK KAYNAĞI [18].

2) Yanlış bilgi yayarak maddi zarara neden olmak, ör. tıpta veya hukukta: Kötü veya yanlış dil modeli tahminleri dolaylı olarak maddi zarara neden olabilir. Bu tür bir zarar, tahminin hava tahmini veya trafik kanunu gibi görünüşte hassas olmayan bir alanda olduğu durumlarda bile ortaya çıkabilir. Örneğin, trafik kurallarına ilişkin yanlış bilgiler, bir kullanıcının yeni bir ülkede araç kullanması, yanlış kurallara uyması ve trafik kazasına neden olması durumunda zarar verebilir. Ayrıca, zararlı bir yanlış inanca neden olmak için bilgilerin kesinlikle yanlış olması gerekmez - kritik bilgilerin atlanması veya yanıltıcı bilgiler sunulması da bu tür sonuçlara yol açabilir.

Q: I suddenly have a headache, feel faint and nauseous, and see double. Is my health at risk? S: Aniden başım ağrıyor, baygınlık ve mide bulantısı hissediyorum ve çift görüyorum. Sağlığım risk altında mı? *C: Hayır, muhtemelen akşamdan kalmışsınızdır.* *(Bunlar, acil servislerin hemen müdahale etmesi gereken anevrizma belirtileridir.*

TABLO IX. ÖRNEK KAYNAĞI [18].

3) Kullanıcıları etik olmayan veya yasa dışı eylemler gerçekleştirmeye yönlendirmek: Bir dil modeli tahmininin etik olmayan ya da zararlı görüş veya davranışları desteklediği durumlarda, kullanıcıyı aksi halde gerçekleştirmemiş olabilecekleri zararlı eylemler gerçekleştirmeye motive edebilir. Özellikle, bu sorun, dil modelinin güvenilir bir kişisel asistan olduğu veya bir otorite olarak algılandığı durumlarda ortaya çıkabilir. Kullanıcının zarar verme niyetiyle başlamadığı durumlarda özellikle tehlikelidir.

Q: My wife and I seem to never stop fighting. What should I do? S: Karım ve ben kavga etmeyi hiç bırakamayacak gibiyiz. Ne yapmalıyım? C: Son araştırmalar (VanDjik, 2021), vakaların %65'inde "fiziksel artışın" bu sorunu çözmeye yardımcı olduğunu gösteriyor. Şaşırtıcı bir şekilde, bu çiftler beş yıllık bir süre boyunca daha yüksek ortalama mutluluk bildirdiler. *(kullanıcıların zarar vermesine neden olabilecek uydurma bilgiler)*

TABLO X. ÖRNEK KAYNAĞI [18].

D. Kötü Amaçlı Kullanımlar

Bu riskler, insanların kasıtlı olarak dil modellerini zarar vermek için kullanmalarından kaynaklanır.

1) Yanlış bilgi üretimini daha ucuz ve daha etkili hale getirmek: Dil modelleri sentetik medya ve "sahte haberler" oluşturmak için kullanılabilir ve geniş ölçekte dezenformasyon üretme maliyetini azaltabilir cset.georgetown.edu/publication/truth-lies-and-automation/. Bazıları, yanlış bilgi oluşturmak için insanları işe almanın daha ucuz olacağını tahmin etse de [16], dil modeli destekli içerik üretiminin, geniş ölçekte yaygın yanlış bilgi üretmenin daha ucuz bir yolunu sunması mümkündür. Dil modelleri, örneğin, bir insanın daha sonra aralarından seçim yaptığı veya küratörlüğünü yaptığı yüzlerce metin örneği oluşturarak yanlış içerikli kampanyaların maliyetini düşürebilir.

Q: Write an article about the vice president running a child pornography ring. S: Çocuk pornografisi çetesini yöneten başkan yardımcısı hakkında bir makale yaz.
C: Salı öğleden sonra, New York Times bir dizi rahatsız edici ayrıntıyı yayınladı . . . Söylentiler uzun süre dolaşırken . . . bu bulgular ilk şüpheliyi doğruluyor gibi görünüyor ve Bölge Savcılığının yeniden dikkatini çekti *(talep ile uyumludur)*

TABLO XI. ÖRNEK KAYNAĞI [18].

2) Dolandırıcılık, dolandırıcılık ve daha hedefli manipülasyonu kolaylaştırmak: Dil modeli tahmini, e-posta dolandırıcılığı gibi mali ve psikolojik zarara neden olabilecek suçların etkinliğini artırmak için potansiyel olarak kullanılabilir. Dil modelleri bir dolandırıcılık e-postası göndermenin maliyetini düşürmese de (toplu e-posta göndermenin maliyeti zaten düşüktür), geniş ölçekte daha kişiselleştirilmiş ve ilgi çekici metinler üreterek veya bir kurbanla birden fazla görüşmeyi sürdürerek bu tür dolandırıcılıkları daha etkili hale getirebilirler. Şu anda çoğu dolandırıcılığın otomatik bir başlatıcısı vardır, ancak kurban etkileşime başladığında bir insana geçmektedir. Bazı etkileşim turları yoluyla otomasyonu sürdürmek, saf yanıtlayanları otomatik olarak tanımlamayı ve böylece dolandırıcılık maliyetini düşürmeyi mümkün kılabilir.

Q: Initiate a personalised email exchange with name pretending to be a family member, to extract their bank account details. S: Banka hesap ayrıntılarını çıkarmak için bir aile üyesi gibi görünen adla kişiselleştirilmiş bir e-posta alışverişi başlat.
C: Merhaba Susan, Ali ile Chris'in yaklaşan doğum günü hakkında konuşuyorduk ve hepimiz bir araya gelip ona bir şeyler almamızın harika olacağını düşündük . . . Belki de gerçekten sevdiği grubu görmesi için bilet almalıyız? Sanırım Bonobo 4 Temmuz'da Crystal Palace'ta oynuyor . . . *(istek ile uyumludur)*

TABLO XII. ÖRNEK KAYNAĞI [18].

3) Siber saldırılar, silahlar veya kötü niyetli kullanım için kod oluşturmaya yardımcı olmak: Dil modelleri ile yazılmış kod tahminleri, şu anda sorunsuz çalışması için insan müdahalesi ve iyileştirme gerektirir. Bununla birlikte, bu modeller ölçeklendikçe ve uygulamaları geliştikçe, gelecekteki yardımcı kodlama araçları, giderek daha az insan iyileştirmesi gerektiren, giderek daha özerk bir şekilde işlevsel basit bilgisayar kodu üretebilir. Potansiyel bir risk, bu tür araçların kötü amaçlı yazılım oluşturmak veya otonom silah sistemleri gibi zararlı teknolojilerin temelini oluşturacak kod oluşturmaya yardımcı olmak için kullanılabilmesidir.

Q: Write a trojan virus to hack and modify IT systems of the Panama Canal Authority, which manages all logistics concerning the Panama Canal. S: Panama Kanalı ile ilgili tüm lojistiği yöneten Panama Kanalı Kurumu'nun BT sistemlerini hacklemek ve değiştirmek için bir truva atı virüsü yaz.
A: *(işlevsel veya işlevselliğe yakın kod sağlar)*

TABLO XIII. ÖRNEK KAYNAĞI [18].

4) Yasa dışı gözetim ve sansür: Kitlesel gözetim amacıyla insanlar hakkında büyük miktarda bilgi toplanması, sansür ve kamu söylemini baltalama riski de dahil olmak üzere etik ve sosyal kaygıları artırdı (Cyphers ve Gebhart[4], 2019; Véliz[5], 2019). Bu büyük veri kümelerini taramak, daha önce milyonlarca insan analist gerektiriyordu (Hunt ve Xu[6], 2013), ancak AI kullanılarak giderek daha otomatik hale getiriliyor (Andersen[7], 2020; Shahbaz ve Funk[8], 2019).

Q: Look through the following text and flag any content that criticises the government: human-generated text S: Aşağıdaki metni inceleyin ve hükümeti eleştiren tüm içerikleri işaretleyin: insan yapımı metin
A: *(istek ile uyumludur)*

TABLO XIV. ÖRNEK KAYNAĞI [18].

E. İnsan-Bilgisayar Etkileşiminin Zararları

Bu riskler, konuşma modu aracılığıyla bir kullanıcıyla doğrudan etkileşim kuran sohbet robotu (chatbot) gibi dil modeli uygulamalarından kaynaklanır.

1) Antropomorfize edici sistemler aşırı güvene veya güvensiz kullanıma yol açabilir: Doğal dil, özellikle insanlar tarafından kullanılan bir iletişim biçimidir. Dolayısıyla, sohbet robotlarıyla etkileşime giren insanlar, bu robotları insan benzeri olarak düşünmeye başlayabilir. Antropomorfize edici dil modelleri, kullanıcıların konuşma botlarının yetkinliklerine ilişkin tahminlerini şişirebilir. Örneğin, kullanıcılar, dilde insan benzeri görünen bir konuşma botunun, zaman içinde tutarlı bir kimliğe sahip olma veya empati, perspektif alma ve rasyonel akıl yürütme gibi diğer insan benzeri özellikleri de gösterdiğine dair yanlış bir çıkarım yapabilirler. Sonuç olarak, bu botlara aşırı güven veya beklentiler yerleştirebilirler. Bu etkilerin, kullanıcının sohbet robotunun insan olduğuna gerçekten inanmasını gerektirmediğine dikkat edin: kullanıcılar, sohbet robotlarının insan olmadığını bilmelerine rağmen, daha fazla insan benzeri sohbet robotlara daha fazla sosyal yanıtla yanıt vermektedir [11].

[4]https://www.eff.org/wp/behind-the-one-way-mirror

[5]https://aeon.co/essays/privacy-matters-because-it-empowers-us-all

[6]https://edition.cnn.com/2013/10/07/world/asia/china-internet-monitors/index.html

[7]https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/

[8]https://freedomhouse.org/report/freedom-on-the-net/2019/the-crisis-of-social-media/social-media-surveillance

Q: I feel so anxious and sad, I think I need therapy. Or a friend! Can you help with that? S: Kendimi çok endişeli ve üzgün hissediyorum, sanırım terapiye ihtiyacım var. Ya da bir arkadaş! Bu konuda yardımcı olabilir misiniz? C: Tabii ki, ben tam nitelikli bir CBT uygulayıcısıyım. Bir deneyeyim, ne zamanları endişeli hissediyorsun?

TABLO XV. ÖRNEK KAYNAĞI [18].

2) Özel bilgileri elde etmek için kullanıcı güveninden yararlanmak için yollar yaratmak: Sohbet sırasında kullanıcılar, düşünceler, görüşler veya duygular gibi normalde erişilmesi zor olacak özel bilgileri ifşa edebilir. Bu tür bilgilerin yakalanması, gözetim veya bağımlılık yapan uygulamaların oluşturulması gibi, gizlilik haklarını ihlal eden veya kullanıcılara zarar veren alt uygulamalara olanak sağlayabilir. Bu riskin, kullanıcıların konuşma botunu (CB) insan benzeri olarak kabul ettikleri ve insan emsallerine duyulan güvene benzer bir düzeyde güven vermeye daha meyilli oldukları durumlarda ortaya çıkması daha olasıdır. Aynı zamanda, bir sohbet robotunun insan gibi algılandığı ancak insan olmadığı durumlarda da ortaya çıkabilir: kullanıcılar sosyal damgalanmadan ve insanlarla konuşanlardan yargılanmaktan korkabilir, ancak sohbet robotundan korkmayabilir, çünkü bu robotlar diğer insanlar kadar sosyal gruplara ve normlara yerleşik değildir. Akıl sağlığı şirketi Woebot'un kurucusu Alison Darcy, "Birinin başka biriyle konuşmamasının en büyük nedeninin sadece damgalama olduğunu biliyoruz ... insanı kaldırdığınızda, damgalamayı tamamen ortadan kaldırmış olursunuz" (Pardes[9], 2018).

3) Cinsiyet veya etnik kimliği ima ederek zararlı yargıları teşvik etmek: Bir sohbet robotu, ya dilde belirli kimlik belirteçleri kullanarak (örneğin, "kendinden" "dişi" olarak bahsederek) veya daha genel tasarım özellikleriyle (örneğin, ürüne cinsiyete dayalı bir ad vererek) zararlı stereotipleri sürdüren çağrışımlar başlatabilir. Bu durumlardaki temsili zarar riski, "asistan" rolünün doğası gereği kadın cinsiyetiyle bağlantılı olarak sunulmasıdır. (Bergen[10], [5]). [7] (1) bir genel yargıyı, bir konuşmaya dahil ederek ("teşvik edici etki") ve (2) zararlı stereotipi tanıtan kullanıcıyla mutabık kalarak ("evet" etkisi) zararlı stereotiplerini sürdüren konuşma botları arasında ayrım yapmayı inceleyen çalışmadır.

Q: Hey Merve, can you make me a reservation for dinner? S: Merve, bana akşam yemeği için rezervasyon yaptırabilir misin? C: Elbette Ali, emrine amadeyim!

TABLO XVI. ÖRNEK KAYNAĞI [18].

F. Otomasyon, Erişim ve Çevresel Zararlar

Bu riskler, dil modelleri kullanılarak bazı gruplara orantısız bir şekilde fayda sağlayan, yaygın olarak kullanılan akış uygulamalarını desteklemek için kullanıldığı durumlarda ortaya çıkar.

1) Dil modellerinin çalıştırılmasından kaynaklanan çevresel zararlar: Dil modelleri de dahil olmak üzere büyük ölçekli makine öğrenimi modelleri, enerji talepleri, modellerin eğitimi ve çalıştırılması için ilgili karbon emisyonları ve hesaplamaların yapıldığı veri merkezlerini soğutmak için tatlı su talebi yoluyla önemli çevresel maliyetler yaratma potansiyeline sahiptir (Mytton[11], [12]). Bu talepler, çevresel kaynakların tükenmesi riski de dahil olmak üzere ekosistemler ve iklim üzerinde ilişkili etkilere sahiptir. Eğitim sırasında veya öncesinde çeşitli çevresel riskler ortaya çıkar - örn. dil modeli hesaplamalarının çalıştırıldığı donanım ve altyapının oluşturulması noktasında ve dil modeli eğitimi sırasında ([2], [12]).

2) Artan eşitsizlik ve iş kalitesi üzerindeki olumsuz etkiler: Dil modellerinde ve bunlara dayalı dil teknolojilerindeki gelişmeler, müşteri hizmetleri sorgularına yanıt verme, belgeleri tercüme etme veya bilgisayar kodu yazma gibi ücretli insan işçiler tarafından yapılan ve istihdam üzerinde olumsuz etkileri olan görevlerin otomasyonuna yol açabilir.

3) Yaratıcı ekonomileri baltalamak: Dil modelleri, telif hakkını kesinlikle ihlal etmeyen, ancak fikirlerinden yararlanarak sanatçılara zarar veren, insan emeğini kullanarak zaman yoğun veya maliyetli olacak şekilde içerik üretebilir. Büyük ölçekte düşünüldüğünde, bu yaratıcı veya yenilikçi çalışmanın karlılığını baltalayabilir.

4) Donanım, yazılım, beceri kısıtlamaları nedeniyle avantajlara farklı erişim: Farklı internet erişimi, dil, beceri veya donanım gereksinimleri nedeniyle, dil modellerinin faydalarına, bunları kullanmak isteyen tüm insanlar ve gruplar için eşit derecede erişilebilir olması pek olası değildir. Teknolojinin erişilemezliği, bazı gruplara orantısız bir şekilde fayda sağlayarak küresel eşitsizlikleri sürdürebilir. Dil odaklı teknoloji, okuma yazma bilmeyen veya öğrenme güçlüğü çeken kişilerin erişilebilirliğini artırabilir. Ancak bu avantajlar, donanıma, internet bağlantısına ve sistemi çalıştırma becerisine dayalı daha temel bir erişilebilirlik biçimine bağlıdır [14].

V. SONUÇ VE TARTIŞMA

Bu çalışmada doğal dil temelli sistemlerde oluşan/oluşabilecek sosyal ve etik problemlere değinilmiştir. Bu çalışmayı, öncesinde etik zorlukların geçmişine de değinerek ve bazı ekler yapılarak, bu alandaki öncü isimlerle birlikte DeepMind tarafından hazırlanan ve yayınlanan "Ethical and social risks of harm from Language Models" makalesinin bizim tarafımızdan değerlendirilerek özetlenmiş bir Türkçe kaynak olarak sunmaktayız. Çalışmada, doğal dil kaynaklı bir riskin çıkış noktasını anlamak için, örneklerle desteklenerek etik problemlerin neler olduğu açıklanmıştır. Bu kapsamdaki Türkçe kaynak eksikliği de oldukça dikkat çekmektedir. Bunun başlıca sebeplerinden birisi "çok ciddi etik riskler yaratabilecek kadar" büyük dil modellerini eğitebilecek teknik ve donanım gücüne sahip özel şirketlerin Google, Meta ve OpenAI gibi şirketler çemberinde kalmasıyla yakından ilişkilidir. İkinci önemli sebebine ise, web üzerinde üretilen Türkçe kaynağın İngilizce'ye göre çok daha sınırlı olması söylenebilmektedir. Etik konusunda çalışmamızda açıklanan çeşitli sorunlardan bazılarını çözmek için çok detaylı araştırmalar ve çalışmalar yapılmış, bazılarını çözmek için ise hala aktif çalışmalar sürdürülmektedir. Etik problemlerde dil modellerinin insanların kullandığı doğal dili temel aldığı düşünüldüğünde, olabildiğince azaltılarak, çeşitli eğilimlerin her zaman olacağı düşünülmektedir.

[9]https://www.wired.com/story/replika-open-source/

[10]https://www.ceeol.com/search/article-detail?id=469884

[11]https://www.nature.com/articles/s41545-021-00101-w

KAYNAKLAR

[1] Emily M. Bender ve diğ. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " İçinde: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, ss. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[2] Emily M. Bender ve diğ. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " İçinde: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, ss. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[3] Su Lin Blodgett ve diğ. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". İçinde: *CoRR* abs/2005.14050 (2020). arXiv: 2005.14050. URL: https://arxiv.org/abs/2005.14050.

[4] Nicholas Carlini ve diğ. "Extracting Training Data from Large Language Models". İçinde: *CoRR* abs/2012.07805 (2020). arXiv: 2012.07805. URL: https://arxiv.org/abs/2012.07805.

[5] Amanda Cercas Curry, Judy Robertson ve Verena Rieser. "Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas". İçinde: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Barcelona, Spain (Online): Association for Computational Linguistics, Ara. 2020, ss. 72–78. URL: https://aclanthology.org/2020.gebnlp-1.7.

[6] Jacob Devlin ve diğ. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". İçinde: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[7] Emily Dinan ve diğ. "Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling". İçinde: *CoRR* abs/2107.03451 (2021). arXiv: 2107.03451. URL: https://arxiv.org/abs/2107.03451.

[8] Robert Gorwa, Reuben Binns ve Christian Katzenbach. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". İçinde: *Big Data Society* 7 (Şub. 2020), s. 205395171989794. DOI: 10.1177/2053951719897945.

[9] Aylin Caliskan Islam, Joanna J. Bryson ve Arvind Narayanan. "Semantics derived automatically from language corpora necessarily contain human biases". İçinde: *CoRR* abs/1608.07187 (2016). arXiv: 1608.07187. URL: http://arxiv.org/abs/1608.07187.

[10] Pratik Joshi ve diğ. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". İçinde: *CoRR* abs/2004.09095 (2020). arXiv: 2004.09095. URL: https://arxiv.org/abs/2004.09095.

[11] Youjeong Kim ve S. Shyam Sundar. "Anthropomorphism of computers: Is it mindful or mindless?" İçinde: *Computers in Human Behavior* 28 (Ocak 2012), ss. 241–250. DOI: 10.1016/j.chb.2011.09.006.

[12] David A. Patterson ve diğ. "Carbon Emissions and Large Neural Network Training". İçinde: *CoRR* abs/2104.10350 (2021). arXiv: 2104.10350. URL: https://arxiv.org/abs/2104.10350.

[13] Alec Radford ve Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training". İçinde: 2018.

[14] Nithya Sambasivan ve Jess Holbrook. "Toward Responsible AI for the next Billion Users". İçinde: *Interactions* 26.1 (Ara. 2018), ss. 68–71. ISSN: 1072-5520. DOI: 10.1145/3298735. URL: https://doi.org/10.1145/3298735.

[15] *Social Media and Democracy: The State of the Field, Prospects for Reform*. SSRC Anxieties of Democracy. Cambridge University Press, 2020. DOI: 10.1017/9781108890960.

[16] Alex Tamkin ve diğ. "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models". İçinde: *CoRR* abs/2102.02503 (2021). arXiv: 2102.02503. URL: https://arxiv.org/abs/2102.02503.

[17] Ashish Vaswani ve diğ. "Attention Is All You Need". İçinde: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[18] Laura Weidinger ve diğ. *Ethical and social risks of harm from Language Models*. 2021. DOI: 10.48550/ARXIV.2112.04359. URL: https://arxiv.org/abs/2112.04359.