Türk Psikiyatri Dergisi

Turkish Journal of Psychiatry

Mektup/*Letter*

Yapay Zekânın Varsanıları Mı Oluyor?

Sayın Editör,

Yapay zekâ (YZ) artık günlük yaşamlarımızın vazgeçilmezi olma yolunda hızla ilerlemektedir. Özellikle, ChatGPT gibi büyük dil modelleri (BDM) kullanan Üretken Yapay Zeka (GAI) sistemlerinin geliştirilmesi ve erişilebilir olması her alanda dönüşümleri ivmelendirmiştir. Üretken YZ, tüm meslekleri ilgilendirdiği gibi yaş aralığı ayrımını da ortadan kaldırmaktadır. Dolayısıyla, artık kapsamlı bir YZ ekosistemi oluşmaktadır. Ancak, faydaları kadar yol açtığı veya uzun vadede yol açacağı sorunlara karşı farkındalık aynı boyutta gelişmemektedir. Bu nedenle, faydaları kadar risklerinden haberdar olmak nasıl bir ekosistem oluşacağı hakkında bilgi verecektir (Ayhan, 2023; Perc ve ark. 2019).

Üretken YZ ile ilgili kabaca üç temel risk bulunmaktadır. Bunlardan en bilineni veri mahremiyetidir. Verilerle beslenen YZ sistemleri önlemler alınmadığında veri korunması ve mahremiyeti bağlamında ihlaller yapabilmekte ve insanı savunmasız bırakarak çeşitli manipülasyonlara açık hale getirebilmektedir. İkinci olarak, bu sistemlerin yanlı sonuçlar üretme riski yüksektir (Özer ve ark. 2024a). Yanlı sonuçların kaynağı iki türlü olmaktadır: eğitim veri setindeki yanlılıklar ve algoritmalardaki yanlı varsayımlar. Bu sistemler büyük veriye dayalı olarak öğrendiği için büyük verideki din, kültür, ırk, cinsiyet veya sosyoekonomik seviye gibi farklı bağlamlara dayalı yanlılıkları korumakta, dolayısıyla çıktılarında bu yanlılıkları yeniden üretebilmektedir. Yanlılıkların ikinci kaynağı, bu sistemlerin algoritmalarında yapılan varsayımlardır. Algoritmadaki varsayımların yanlılığı doğal olarak sonuçların da yanlı olmalarını sağlamaktadır. Her iki durumda YZ, eşitsizliklerin daha hızlı yayılmasına yol açarak mevcut eşitsizlikleri derinleştirme potansiyeli taşımaktadır. Son olarak, bu sistemlerin ürettiği tüm bilgilerin doğru olmadığı, YZ sistemlerinin çoğu zaman varsanı görerek yanlış içerik üretebilme davranışı gösterebilmeleridir. Bu yazıda, YZ sistemlerinin davranışında görülen varsanılar üzerinde kısa bir değerlendirme yapılacaktır.

Yapay zekâ varsanısı, YZ'nin ikna edici, metin içerisinde tutarlı gibi duran, ancak kullanıcı girdisinden veya önceki bağlamdan tamamen bağımsız uydurma bir yanıt oluşturduğu bir fenomeni tanımlamak için kullanılmaktadır. Dolayısıyla, üretken YZ'nin ürettiği yanıtlar makul görünebilmesine rağmen anlamsız veya yanlış olabilmektedir. Örneğin, tamamen ChatGPT tarafından hazırlanan araştırma önerilerinde yapay zekâ varsanının sıklığını değerlendirmeyi amaçlayan bir çalışmada ChatGPT tarafından oluşturulan 178 kaynaktan 69 kaynağa Dijital Nesne Tanımlayıcısı (Digital Object Idenfier, DOI) olmadığı ve 28 kaynağın ne Google aramasında çıktığı ne de mevcut bir DOI'sinin olduğu gösterilmiştir (Athaluri ve ark. 2023). Üretken YZ genel olarak eğitim veri setine dayalı içerik üretirken varsanı sürecinde eğitim veri seti kaynağı ile ilişkisini kopartarak (kaynak unutma/ source amnesia) içerik üretebilmektedir (Berberette ve ark. 2024). Dahası, üretken YZ ürettiği yanlış sonuçlarla tutarlı olma adına bir kez yanlış içerik ürettiğinde ardışık olarak yanlış içerik üretimine devam edebilmektedir ki bu davranış, varsanının kartopu etkisi olarak bilinmektedir (Zhang ve ark. 2023).

Varsanılar da kendi aralarında içsel ve dışsal olarak sınıflandırılmaktadır (Ji ve ark. 2023). Bu sınıflandırmaya göre içsel varsanılar kaynak içeriği ve konuşma geçmişi ile çelişen çıktılara yol açarken dışsal varsanılar kaynak içeriği veya konuşma geçmişine bakarak doğruluğu kanıtlanamayan çıktılara karşılık gelmektedir. Buna göre içsel varsanıya dayalı üretilen çıktı bilginin yanlış yorumlanması ile ilişkili ve doğrudan yanlış iken dışsal varsanılar üretilen metne eklenen yanlış içeriklerdir. Diğer taraftan Zhang ve ark. (2023) varsanıları üç farklı kategoride değerlendirmektedir: girdi ile çelişen (kullanıcı tarafından sağlanan

kaynak girdiden sapan içerik), bağlam ile çelişen (üretken YZ tarafından önceden üretilen içerikle çelişen içerik) ve gerçek ile çelişen varsanı. Her iki sınıflandırma birleştirilirse girdi ve bağlam ile çelişen varsanı içsel varsanıya karşılık gelirken gerçek ile çelişen varsanı dışsal varsanıya karşılık gelmektedir.

Üretken YZ'de görülen varsanının gerçek bir bilişsel sürecin bir sonucu olmaktan çok modelin olasılıksal doğası ve eğitim veri seti ile doğrudan ilişkili olduğu görülmektedir. Eğitim veri setinde bazı kalıplar, ifadeler veya kavramların daha sık içerilmesi üretken YZ'nin cevap üretmesi sürecinde bunlara erişilebilirliğin kolay ve hızlı olması nedeniyle bağlamın doğruluğundan bağımsız olarak varsanıları tetikleyebildiği gibi büyük eğitim veri setindeki birbiriyle çelişen bilgiler üretken YZ'nin bu verilere dayalı cevap üretme sürecinde içsel gerilimler yaşamasına yol açarak varsanıyı tetikleyebilmektedir (Berberette ve ark. 2024). Benzer şekilde veri setlerindeki güncel olmayan, eksik veya sahte bilgiler de varsanıya yol açabilmektedir (Zhang ve ark. 2023b). Dolayısıyla üretken YZ varsanısı ile eğitim veri dağılımı arasında güçlü bir ilişki bulunmaktadır (McKenna ve ark. 2023). Diğer taraftan, üretken YZ'de kullanılan çeşitli dil modellerinin zayıf yönleri nedeniyle de varsanılar oluşabilmektedir (Athaluri ve ark. 2023). Kısaca, üretken YZ'leri eğitmek için kullanılan büyük, düzenlenmemiş metin gövdesi, modelin olasılıksal doğasının yol açtığı stokastik davranış, dil modellerinin zayıf yönleri varsanıya yol açabilmektedir.

Diğer taraftan, üretken YZ'nin bu davranışı çoğunlukla YZ sistemlerinin öğrenmesinde kullanılan veri setlerindeki eksik, tutarsız veya yanlış bilgilerle ilişkili olduğu göz önüne alındığında, gerçek yaşam verilerinin varsanıya ne kadar açık olduğunu da göstermektedir. Aslında, günlük hafıza rekonstrüksiyonunun genellikle bir dereceye kadar masallama (konfabulasyon) içerdiği (French ve ark. 2009) ve sağlıklı bireylerin de bir kasıt olmadan hikâyelerin detaylarını hayal ürünü hale getirmelerinin yaygın olduğu öne sürülmektedir (Riesthuis ve ark. 2023). Bir başka deyişle, insanlar sıklıkla, kurgusal ancak gerçek olmayan bilgilere kötü bir niyet olmadan anlatılarındaki boşlukları doldurmak için başvurabilmektedir. Benzer şekilde, üretken YZ'nin ürettiği uydurma içeriklerin genel olarak anlatısallık sergilenen metinlerde daha sık ortaya çıktığı görülmektedir (Sui ve ark. 2024).

Ayrıca, GAI tarafından geniş eğitim setine dayalı içerik üretme sürecinde ilham ile taklit arasında ayrım yapmanın zorluğu da ayrı bir sorun olarak durmaktadır (Carobene ve ark. 2024). BDM-modifikasyonuna sahip makaleler arasında daha yakın bir ilişki olduğunun gösterilmesi, GAI'nin metin çeşitliliğini azalttığına işaret etmektedir (Liang ve ark. 2024). Bu bağlamda, son zamanlarda varsanıların yol açtığı sorunların alanlara göre değerlendirilmesi gerektiği, bazı alanlarda varsanıların faydalı olabileceği ile ilgili çalışmalar dikkat çekmektedir. Örneğin, varsanıların gerçeklere dayalı çıktılardan anlamlı derecede daha yüksek anlatısallık sergilediği ampirik olarak gösterilmiştir (Sui ve ark. 2024). Varsanıların özellikle değerli olabileceği alanlar olarak yaratıcı yazı için ilham sağlama (Mukherjee ve Chang, 2023), mimari tasarım (Hegazy ve Saleh, 2023), yeni proteinlerin keşfi (Anishchenko ve ark. 2021) ve yenilikçi yasal benzetmeler formüle etme (Dahl ve ark. 2024) gösterilmektedir (Sui ve ark. 2024). Dolayısıyla, kullanılan dil modellerini bu bağlamda güçlendirmenin ötesinde öğrenme verilerindeki yanlış bilgileri ayıklamak varsanının sıklığını veya ölçeğini azaltmaya katkı sunabilmesine ve bu kapsamda çalışmalar yürütülmesine rağmen, eksik veya tutarsız bilginin neye göre eksik veya tutarsız olduğuna, bu kapsamda varsanıları azaltmanın üretilen bilgilerin çeşitliliğini olumsuz etkileyip etkilemeyeceğine dikkat edilmelidir.

Diğer taraftan, klinik olarak, dışsal bir uyarıcıyla ilişkilendirilmeyen duyusal deneyimler olarak tanımlanan varsanılar genellikle şizofreni, bipolar bozukluk ve Parkinson hastalığı gibi durumlarla ilişkilendirilir (Tamminga, 2009). BDM'ler tarafından üretilen gerçek dışı çıktıları tanımlamak için varsanı terimini kullanmak, benzer şekilde BDM'lerin algılama sürecinde bilinçli olarak bir duyusal girdiye dâhil olduklarını ima eder (Smith ve ark. 2023). Oysa, üretken YZ bilince sahip değildir, dolayısıyla öznel deneyim veya farkındalık eksikliğinden maluldür (Berberette ve ark. 2024). Diğer taraftan, dışsal uyarıcıyla ilişkilendirilmeyen insan varsanısından farklı olarak YZ sistemlerinde BDM'lerin eğitildiği veriler ve bu davranışa yol açan istemler dış uyarıcılar olarak kabul edilmektedir (Østergaard ve Nielbo, 2023). Bu nedenle, üretken YZ modelleri orada olmayan bir şeyi görmediği, ancak uydurduğu için varsanı teriminin bu durumu açıklamak için kullanılmasının doğru olmadığı, bunun yerine tam da bu duruma karşılık gelen bir psikiyatrik kavram olarak masallama ('yanlış olmasına rağmen böyle olduğu fark edilmeyen anlatısal ayrıntıların oluşturulması') önerilmektedir (Smith ve ark. 2023). Son zamanlarda yapılan çalışmalar, üretken YZ tarafından üretilen bilgilerin çoğunun masallama bağlamında değerlendirilebileceğine işaret etmektedir (Berberette ve ark. 2024; Sui ve ark. 2024).

Varsanı, dış uyarıcı olmaksızın meydana gelen duyusal algıyı tanımlamak için kullanılan tıbbi bir terimdir. YZ modelleri bu tür duyusal algılara sahip değildir—ve hata yaptıklarında, bu dış uyarıcı olmaksızın meydana gelmez. Aksine, YZ modellerinin eğitildiği veriler (metaforik olarak) dış uyarıcılar olarak kabul edilebilir—yanı sıra (zaman zaman yanlış) yanıtları ortaya çıkaran istemler de dış uyarıcılar olarak kabul edilebilir.

Son olarak, bu davranışın masallama olarak anlaşılmasının sorunun hafifletilmesine yönelik çözümü de içerisinde barındırdığını düşünüyoruz. Klinik olarak masallama, beynin sağ yarı küre eksiliği ile ilişkili olup bu durumda beynin sol yarı küresi tarafından var olan bilgi, deneyim veya bağlamsal ipuçlarından etkilenerek uydurulmuş veya bozulmuş anıların yaratılmasını içermektedir (Smith ve ark. 2023). Dolayısıyla, üretken YZ'deki durum da mevcut bilgi ve bağlamdan etkilenerek bilginin yeniden inşasının yanlış yapılmasına karşılık geldiği için masallama kavramı mevcut durumu teknolojik arkaplan olarak daha iyi tanımlamaktadır. Bu kavram, aslında YZ ekosisteminde özişler (otomasyon) yerine insanı tamamlayan yol tercihinin daha insani ekosistem oluşturacağına yönelik önemli bir kanıt da sunmaktadır. Daha önce de ifade edildiği gibi, insanı

tamamlayan yol, YZ sistemlerinin insanın yerine ikame edilmesi yerine, özişlerin verimliliğe etkilerini de göz önüne alan, ancak istihdamı koruyan ve insanın eksikliklerini tamamlamaya odaklanan yeni bir yaklaşımdır (Capraro ve ark. 2023; İlikhan ve ark. 2024; Özer ve ark. 2024b; Özer ve Perc 2024). Dolayısıyla, beynin sağ yarı küresi hasar görerek gerçek fonksiyonunu icra etmediğinde ortaya çıkan masallama, beynin sağ yarı küresinin katkılarından yoksun YZ'nin çalışma şekline karşılık gelmektedir. Bir başka deyişle, üretken YZ'nin yanlış veya olmayan bilgiler üretmesi, beynin sağ yarı küresi hasar gördüğünde sol yarıkürenin benzer davranışı ile örtüşmektedir. Bu nedenle, insanın devreye girerek bu eksiklikleri gidermesi ve metni tamamlaması metaforik olarak YZ'ye insanın sağ yarı küresini bağışlayabilir ve bu tip hataların düzeltilmesine destek olarak üretken YZ'yi tamamlayabilir.

Sonuç olarak, üretken YZ sistemleri çok farklı nedenlerden dolayı yanlış bilgiler üretebilmektedir. Dolayısıyla, üretken YZ tarafından üretilen her bilginin doğru olmayacağı, doğru bilgiler arasına gömülü kasıt olmadan üretilen yanlış bilgilerin her zaman var olabileceği aşikârdır. Bu davranış ister varsanı isterse masallama olarak tanımlansın, üretken YZ'nin bu davranışının yol açtığı komplikasyonun şiddeti kullanılan alana göre değişmektedir. Örneğin, sağlık alanında tanı ve tedavi sürecinde üretken YZ'nin varsanı veya masallama deneyimlemesi hayati açıdan oldukça tehlikeli sonuçlara yol açacaktır (İlikhan ve ark. 2024). Benzer şekilde eğitim sisteminde kullanılması sırasında ortaya çıkacak bu davranış öğrenme süreçleri üzerinde derin olumsuz etkiler bırakabilir (Özer, 2024). Ancak, literatürde ifade edildiği gibi yaratıcı ve yenilikçi bazı alanlarda faydalı olabilir. Bu nedenle, bu davranışa yönelik bağlamsal farkındalık son derece kritiktir.

KAYNAKLAR

Anishchenko I, Pellock SJ, Chidyausiku TM ve ark. (2021) De novo protein design by deep network hallucination. Nature 600: 547–52.

Athaluri S, Manthena S, Kesapragada V ve ark. (2023) Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus 15: e37432.

Ayhan Y (2023) The Impact of Artificial Intelligence on Psychiatry: Benefits and Concerns-An assay from a disputed 'author'. Turkish Journal of Psychiatry 34: 65–7.

Berberette E, Hutchins J, Sadovnik A (2024) Redefining "Hallucination" in LLMs: Towards a psychology-informed framework for mitigating misinformation. arXiv:2402.01769v1.

Carobene A, Padoan A, Cabitza F ve ark. (2024) Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. Clin Chem Lab Med 62: 835–43.

Capraro V, Lentsch A, Acemoğlu D ve ark. (2023) The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus 3: 1-18.

Dahl M, Magesh V, Suzgun M ve ark. (2024) Large legal fictions: Profiling legal hallucinations in large language models. arXiv:2401.01301.

French L, Garry M, Loftus E (2009) False memories: A kind of confabulation in non-clinical. Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy, William Hirstein (ed.): 33-66.

Hegazy M, Saleh AM (2023) Evolution of AI role in architectural design: between parametric exploration and machine hallucination. MSA Engineering Journal 2: 1-26.

İlikhan S, Özer M, Tanberkan H ve ark. (2024) How to mitigate the risks of deployment of artificial intelligence in medicine? Turk J Med Sci 54: 483-92.

Ji Z, Lee N, Frieske R ve ark. (2023) Survey of hallucination in natural language generation. ACM Computing Survey 55: 1–38.

Liang W, Zhang Y, Wu Z ve ark. (2024) Mapping the increasing use of LLMs in scientific papers. arXiv preprint arXiv:2404.01268v1.

McKenna N, Li T, Cheng L ve ark. (2023) Sources of hallucination by large language models on inference tasks. arXiv preprint arXiv:2305.14552.

Mukherjee A, Chang H (2023) The creative frontier of generative ai: Managing the novelty-usefulness tradeoff. arXiv:2306.03601.

Østergaard SD, Nielbo KL (2023) False Responses From Artificial Intelligence Models Are Not Hallucinations. Schizophr Bull 49: 1105–7.

Ozer M (2024) Potantial benefits and risks of artificial intelligence in education. Bartın University Journal of Faculty of Education 13: 232-44.

Ozer M, Perc M (2024) Human complementation must aid automation to mitigate unemployment effects due to AI technologies in the labor market. Reflektif Journal of Social Sciences 5: 503-14.

Ozer M, Perc M, Suna HE (2024a) Artificial intelligence bias and the amplification of inequalities in the labor market. Journal of Economy, Culture and Society 69: 159-68.

Ozer M, Perc M, Suna HE (2024b) Participatory management can help AI ethics adhere to the social consensus. Istanbul University Journal of Sociology 44: 221-38.

Perc M, Özer M, Hojnik J (2019) Social and juristic challenges of artificial intelligence. Palgrave Communications 5, 61.

Riesthuis P, Otgaar H, Bogaard G ve ark. (2023) Factors affecting the forced confabulation effect: a meta-analysis of laboratory studies. Memory 31:635–51.

Smith AL, Greaves F, Panch T (2023) Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. PLOS Digit Health 2: e0000388.

Sui P, Duede E, Wu S ve ark. (2024) Confabulation: The Surprising Value of Large Language Model Hallucinations. arxiV:2406.04175v2.

Tamminga CA. Schizophrenia and other psychotic disorders: Introduction and overview. In: Sadock BJ, Sadock VA, Ruiz P, eds. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 9th edition. Philadelphia: Lippincott Williams and Wilkins; 2009. p. 1432.

Zhang, Y, Li Y, Cui L ve ark. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Zhang M, Press O, Merrill W ve ark. (2023) How language model hallucinations can snowball. arXiv preprint arXiv:2305.13534.

Mahmut ÖZER

Geliş Tarihi: 15.09.2024. Kabul Tarihi: 19.09.2024, **Çevrim İçi Tarihi:** 14.10.2024

Milli Eğitim, Kültür, Gençlik ve Spor Komisyonu, Türkiye Büyük Millet Meclisi.

Müh. Mahmut Özer, e-posta: mahmutozer2002@yahoo.com

https://doi.org/10.5080/u27587