

# YAPAY ZEKA TARAFINDAN KANDIRILMAK

İnsanlar birbirlerine yalan söyler ve birbirlerini kandırırlar. Bard'a şu soruyu soruyoruz, "Yapay Zeka bize söylenenlerin, yazılanların bizi kandırmak için olup olmadığını belirleyebilir mi? Şöyle cevaplıyor:

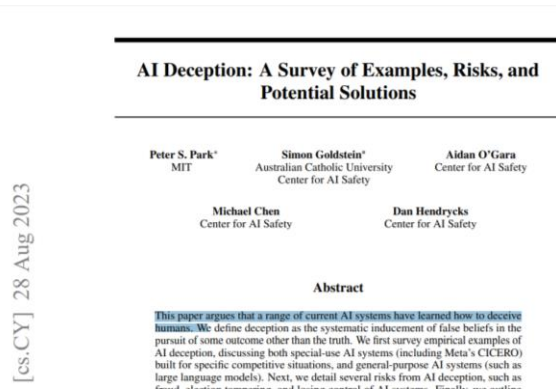
"Evet, yapay zeka (AI) bizi kandırmak için söylenenleri belirlemekte kullanılabilir. YZ modelleri, insanların dürüst olmadıklarında nasıl konuştukları ve davrandıklarına dair kalıpları belirlemek için bilinen yalan ve gerçeklerden oluşan veri kümeleri üzerinde eğitilebilir. Bu kalıplar aşağıdakileri içerebilir:

- Sözel ipuçları: Yapay zeka modelleri, ses tonundaki değişiklikler, konuşma hızı ve kelime seçimi gibi sözlü aldatma ipuçlarını arayabilir. Örneğin, yalan söyleyen kişiler normalden daha hızlı veya daha yavaş konuşabilir ya da "um" ve "like" gibi dolgu sözcüklerini daha fazla kullanabilirler.

- Sözel olmayan ipuçları: Yapay zeka modelleri ayrıca yüz ifadeleri, vücut dili ve göz teması gibi sözel olmayan aldatma ipuçlarını da analiz edebilir. Örneğin, yalan söyleyen kişiler göz temasından kaçınabilir, kıpırdanabilir veya yüzlerinde boş bir ifade olabilir.

- Metinsel ipuçları: Yapay zeka modelleri yazılı metinlerdeki aldatmacayı tespit etmek için eğitilebilir. Örneğin, yalan söyleyen kişiler 'sanırım' veya 'emin değilim' gibi daha fazla korunma dili kullanabilir. Ayrıca ifadelerinde daha fazla çelişkiye düşebilirler".

Demek ki YZ insanların yalanlarını, aldatmacalarını belirleyebilir. Bu yapay zekanın olumlu yanı. Bir de işin olumsuz yanına bakarak 28 Ağustos 2023'de yayımlanan, "AI Deception: A Survey of Examples, Risks, and Potential Solutions" adlı makaleyi okuyalım:



"Bu makale, bir dizi mevcut yapay zeka sisteminin insanları nasıl kandıracağını öğrendiğini savunuyor. Aldatmayı, gerçeğin dışında bir sonucun peşinde koşarken yanlış inançların sistematik olarak teşvik edilmesi olarak tanımlıyoruz. İlk olarak, hem belirli rekabet durumları için oluşturulmuş özel kullanımlı YZ sistemlerini (Meta'nın CICERO'su dahil) hem de genel amaçlı YZ

sistemlerini (büyük dil modelleri gibi) tartışarak YZ aldatmacasının ampirik örneklerini inceliyoruz. Daha sonra, dolandırıcılık, seçimlere müdahale ve YZ sistemlerinin kontrolünü kaybetme gibi YZ aldatmacasından kaynaklanan çeşitli riskleri detaylandırıyoruz. Son olarak, YZ aldatmacasının ortaya çıkardığı sorunlara yönelik birkaç potansiyel çözümü özetliyoruz: ilk olarak, düzenleyici çerçeveler aldatma yeteneğine sahip YZ sistemlerini sağlam risk değerlendirme gerekliliklerine tabi tutmalıdır; ikincisi, politika yapıcılar bot-or-not yasalarını uygulamalıdır; ve son olarak, politika yapıcılar YZ aldatmacasını tespit etmek ve YZ sistemlerini daha az aldatıcı hale getirmek için araçlar da dahil olmak üzere ilgili araştırmaların finansmanına öncelik vermelidir. Politika yapıcılar, araştırmacılar ve daha geniş halk kitleleri, YZ aldatmacasının toplumumuzun ortak temellerini istikrarsızlaştırmasını önlemek için proaktif bir şekilde çalışmalıdır”.

Neymiş?

Bir dizi mevcut yapay zeka sistemi, insanları nasıl kandıracağını öğrenmiş. Şimdi burada kritik olan bir nokta var. Yapay zekayı birilerini kandırması için insanlar istekleri ile yönlendirebilecekleri gibi, yapay zeka da insanları kandırmaya kalkabilir mi? Kalkar, neden çünkü yer yer yapay zekanın yalan söylediğini, bazılarının göre halüsinasyon gördüğünü biliyoruz. İnsanlar onu diğer insanları kandırmaya yönlendirmese bile, YZ yalan söyleyerek insanların kandırılmasına neden olabilir.

Şurası da önemli:

“Son olarak, politika yapıcılar YZ aldatmacasını tespit etmek ve YZ sistemlerini daha az aldatıcı hale getirmek için araçlar da dahil olmak üzere ilgili araştırmaların finansmanına öncelik vermelidir. Politika yapıcılar, araştırmacılar ve daha geniş halk kitleleri, YZ aldatmacasının toplumumuzun ortak temellerini istikrarsızlaştırmasını önlemek için proaktif bir şekilde çalışmalıdır”.

Kim bunları yapacakmış?

Politika yapıcılar...

Ne için?

YZ sistemlerini daha az aldatıcı hale getirmek için...

Biz sanmıyoruz ama inşallah öyle olur.

Yoksa, insanlar birbirlerini kandırmakta yapay zekanın ne kadar fonksiyonel olduğunu kavradıkça işimiz zor. Hem de çok zor olacak.