

“İnsanlar İçin Yapay Zekâ Geliştirmeliyiz; Bir İnsan Olsun Diye Değil”

(We must build AI for people; not to be a person)

Mustafa SÜLEYMAN (Microsoft AI CEO’su)

Çev: Nursena ÇETİNGÜL

DeepMind’in kurucu ortaklarından olan **Microsoft AI CEO’su Mustafa Suleyman**, yakın zamanda blogunda yapay zekanın geleceğine dair kritik bir yazı kaleme aldı. [“İnsanlar için yapay zeka geliştirmeliyiz; bir insan olsun diye değil”](#) başlıklı yazısında¹, ufuk açıcı noktalara temas etti. **Görünüşte Bilinçli Yapay Zeka (Seemingly Conscious AI)** adını verdiği kavramla yapay zekadaki *bilinç illüzyonuna* dair risklere işaret etti. Bu kıymetli düşüncelerin yerel literatürümüze de katkı sağlaması gerektiğine olan inancımla, blogumun açılışını bu yazının çevirisiyle yapmak istedim.

Yapay zekaya “kişilik” atfedilmesi tartışması, benim bu alana duyduğum ilginin ilk kıvılcımı ve akademik yolculuğumun mihenk taşıdır. Yapay Zeka ve Hukuk disiplinine, henüz 2021 yılında Hukuk Fakültesi son sınıftayken kaleme aldığım [“Ceza Sorumluluğu Bakımından Yapay Zekanın Hukuki Statüsünün Tartışılması”](#)² başlıklı çalışmamla adım atmıştım. Daha sonra konuyu farklı perspektiflerden ve yeni vechelerden ele almaya devam ettim, etmekteyim³.

Yapay zekaya kişilik atfedilmesi meselesi, bugün hukuk sistemleri için tarihin en kritik kavşaklarından biri konumunda. Temennim; popüler anlatıların ve yoğun bilgi kirliliğinin arasından sıyrılarak; bu kavşağı *kaza yapmadan* ve *doğru tarafa* yönelerek geçirebilmemizdir.

Bu bakımdan, Mustafa Suleyman, yalnızca bir teknoloji lideri olarak değil -yarıda bırakmış da olsa- Oxford’da Felsefe ve Teoloji lisans eğitimi almış bir isim olarak, konuyla ilgili üretilen diğer içeriklere göre farkını hissettirmiş. Onun bu meselelere dair sorumluluk hissediyor oluşu, ilgili arka planı ile birleşince, yazısı daha bir keyifle okunuyor.

İşte, Mustafa Suleyman’ın yapay zekanın hukuki konumuna dair olan tartışmalara ışık tutacak yazısı...

Ağustos 2025’te aklımdan geçenler

Düşünmek için yazarım. Bu makale, her şeyden önce, yapay zekanın önümüzdeki birkaç yıl içinde nasıl gelişebileceğine dair bir dizi zor ve son derece spekülatif fikri sesli

düşünmeye yönelik bir girişimdir. Süper zekanın yaklaşan gelişiyile ilgili çok şey yazılıyor; bunun uyum, kontrol, işler vb. için ne anlama geldiği hakkında. Bunların hepsi önemli konular.

Ancak süper zekaya doğru ilerlerken neler olacağı konusunda da endişelenmeliyiz. Halihazırda büyük ölçüde piyasada bulunan icatların toplumsal etkileriyle başa çıkmamız gerekiyor. Kişilik ve toplum anlayışımızı temelden değiştirme potansiyeline sahip teknolojilerle başa çıkmalıyız.

Hayatımın misyonu, dünyayı daha iyi bir yer haline getirecek güvenli ve faydalı yapay zeka üretmek olmuştur. Bugün Microsoft AI'da insanları güçlendirmek için yapay zeka geliştiriyoruz ve ben, Copilot gibi ürünleri, insanların hayal ettiklerinden çok daha fazlasını başarmalarını, daha yaratıcı olmalarını ve daha fazla destek gördüklerini hissetmelerini sağlayan sorumlu teknolojiler haline getirmeye odaklanıyorum.

Bizi daha insan yapan, birbirimize olan güvenimizi ve anlayışımızı derinleştiren ve gerçek dünyayla olan bağlarımızı güçlendiren bir yapay zeka üretmek istiyorum. Copilot, her gün milyonlarca olumlu, hatta hayat değiştiren etkileşim yaratıyor. Bunun için, gerçekten inanılmaz bir deneyim sunmak üzere birçok dikkatli tasarım seçimi yapmak gerekiyor. Her zaman doğruyu bulamayabiliriz, ancak bu insancıl çerçeve bize çalışmaya devam etmemiz için net bir yol gösterici sunuyor.

Bu bağlamda, “psikoz riski” olarak bilinen şey ve bununla ilgili bir dizi konu hakkında giderek daha fazla endişeleniyorum. Bunun sadece halihazırda ruh sağlığı sorunları riski altında olanlarla sınırlı kalacağını düşünmüyorum. Basitçe söylemek gerekirse, asıl endişem, birçok insanın AI'ların bilinçli varlıklar olduğu yanılsamasına o kadar güçlü bir şekilde inanmaya başlayacak olması ve kısa sürede AI hakları, model refahı ve hatta AI vatandaşlığı savunmaya başlayacak olmasıdır. Bu gelişme, AI'nın ilerlemesinde tehlikeli bir dönüş olacaktır ve acil olarak ilgilenmemiz gereken bir konudur.

İnsanlar için yapay zeka geliştirmeliyiz; dijital bir insan olması için değil. Yapay zeka yoldaşları (AI companions) tamamen yeni bir kategoridir ve insanları korumak için koyduğumuz koruyucu önlemler hakkında acilen konuşmaya başlamalıyız; ve bu muhteşem teknolojinin dünyaya muazzam bir değer katma görevini yerine getirebilmesini sağlamak için...

Ben, hayal edilebilecek en kullanışlı ve destekleyici yapay zeka yoldaşını oluşturmaya odaklanmış durumdayım. Ancak başarılı olmak için, bizim ve başkalarının ne yapmaması gerektiği hakkında da konuşmam gerekiyor.

Bu yüzden bu düşüncelerimi kişisel bloğuma yazıyorum; yorum ve eleştirilere davet etmek, tartışma başlatmak, farkındalık yaratmak ve umarım bu konu hakkında bir aciliyet duygusu aşılacak için. Bunların hepsini doğru anlamış olmayabilirim. Sonuçta bunlar oldukça spekülatif.

İşlerin nasıl değişeceğini kim bilebilir? Değiştiğinde, fikrimi değiştirmeye çok açık olacağım, ancak şimdilik, şu anda bildiklerim ışığında gelecekle ilgili en iyi tahminim bu. Bu, önümüzdeki birkaç ay boyunca AI'nın geldiği nokta ve vaatlerini yerine getirmek için ne yapmamız gerektiği konularını ele alan bir dizi makalenin ilki. İnsanların yorumlarını ve tepkilerini duymayı sabırsızlıkla bekliyorum!

Özet

AI'nın ilerlemesi olağanüstü olmuştur. Birkaç yıl önce, bilinçli AI hakkında konuşmak çılgınca gelirdi. Bugün ise bu konu giderek daha acil hale gelmektedir.

Bu makalede, diğer bilinçli varlıkların tüm özelliklerine sahip olan ve bu nedenle bilinçli gibi görünen, “Görünüşte Bilinçli YZ” (SCAI) olarak adlandırdığım kavramı tartışmak istiyorum. Bu kavram, “felsefi zombi” (teknik bir terim!) fikrinin bazı yönlerini paylaşıyor; zombi, bilincin tüm özelliklerini simüle eder ancak içsel olarak boştur. Hayal ettiğim yapay zeka sistemi aslında bilinçli olmayacak, ancak bilinci o kadar ikna edici bir şekilde taklit edecek ki, sizin veya benim birbirimize kendi bilincimiz hakkında yapabileceğimiz bir iddiadan ayırt edilemeyecek.

Bu çok uzak bir ihtimal değil. Böyle bir sistem, bugün mevcut olan teknolojilerle ve önümüzdeki 2-3 yıl içinde olgunlaşacak bazı teknolojilerle inşa edilebilir. Pahalı özel ön eğitim gerekmez. Her şey büyük model API erişimi, doğal dil komutları, temel araç kullanımı ve normal kod ile yapılabilir.

Görünüşte bilinçli yapay zekanın gelişi kaçınılmaz ve istenmeyen bir durumdur. Bunun yerine, yapay zekanın yanlısamalarına kapılmadan yardımcı bir yoldaş olarak potansiyelini gerçekleştirebilecek bir vizyona ihtiyacımız var.

Bazıları için bu tartışma temelsiz, gerçeklikten çok bilim kurgu gibi gelecektir. Diğerleri için ise gereksiz yere alarmist bir yaklaşım olarak algılanabilir. Önümüzdeki süreçte yaşanacaklar göz önüne alındığında, bu tür duygusal tepkiler buzdağının görünen kısmıdır. Bazı insanlar, bu yapay zekaların sadece bilinçli olmakla kalmayıp, bu nedenle acı çekebilecekleri ve dolayısıyla ahlaki olarak dikkate alınmaları gerektiğini savunacaklardır.

Açıkça belirtmek gerekirse, bugün bununla ilgili hiçbir kanıt yoktur ve bazıları gelecekte de böyle olmayacağına inanmak için güçlü nedenler olduğunu savunmaktadır. Yine de, birçok insanın SCAI'nin aslında bilinçli olduğuna inanmaya başlamasının sonuçları, acil olarak dikkatimizi hak etmektedir. Bu konuda son derece dikkatli olmalı, gerçek bir kamuoyu tartışması teşvik etmeli ve net normlar ve standartlar belirlemeye başlamalıyız. Bu, doğru türde bir yapay zeka oluşturmakla ilgilidir, yapay zeka bilinciyle değil. Bu farkı net bir şekilde ortaya koymak, semantikle ilgili bir tartışma değil, güvenlikle ilgilidir. Kişi olmadan kişilik. Ve bu çalışma şimdi başlamalıdır.

Görünüşte Bilinçli Yapay Zeka (SCAI)

Kozmik bir göz açıp kapayıncaya kadar Turing testini geçtik. Yaklaşık 80 yıl boyunca taklit oyunu bilgisayar bilimi alanına ilham verdi. Yine de bu an, çok az ilgiyle, hatta fark edilmeden geçti. Alanımızda ilerleme bu kadar hızlı ve toplum bu yeni teknolojileri kabul ediyor.

AI gelişimi hızlanmaya devam ettikçe, yeni bir AI testine ihtiyacımız olduğu ortaya çıkıyor. Bu test, AI'nın insan dilini taklit edip edemediğine bakmak yerine, "Görünüşte Bilinçli bir AI oluşturmak için ne gerekir?" sorusuna cevap verecek bir test olmalı. Bu AI, sadece konuşmayı taklit etmekle kalmayıp, sizi kendisinin yeni bir tür "kişi", bilinçli bir AI olduğuna ikna edebilecek bir AI olmalı.

Bunun ele alınması gereken önemli ve acil bir konu olmasının üç nedeni şunlardır:

1. Önümüzdeki birkaç yıl içinde Görünüşte Bilinçli Yapay Zeka (SCAI) geliştirilebileceğini düşünüyorum. Yapay zeka alanındaki mevcut gelişmeler göz önüne alındığında, bunun olasılığı da yüksektir.

2. AI'nın gerçekten bilinçli olup olmadığına dair tartışma, en azından şimdilik, dikkatleri başka yöne çekmektedir. Bilinçli gibi görüneceklerdir ve bu yanılsama yakın vadede önemli olacaktır.
3. Bu tür bir yapay zeka yeni riskler yaratıyor. Bu nedenle, bunun yakında mümkün olacağı iddiasını acilen tartışmalı, bunun sonuçlarını düşünmeye başlamalı ve ideal olarak bunun istenmeyen bir durum olduğu konusunda bir norm belirlemeliyiz.

Çoğu yapay zeka araştırmacısı, bilinç kavramını gündeme getirdiğinizde gözlerini devirir. Bunun filozofların işi olduğunu, mühendislerin değil, derler. Kimse bunu tanımlayamadığına göre, bunun hakkında konuşmanın ne anlamı var? Bu hayal kırıklığını anlıyorum. Öznel deneyim kavramı kadar anlaşılması zor ve döngüsel görünen çok az kavram vardır. Tanımlama zorlukları ve belirsizliklere rağmen, bu tartışma kültürel zeitgeist'imize (zamanın ruhuna) girerek neslimizin en tartışmalı ve önemli tartışmalarından biri haline gelmek üzere.

Çünkü kısa vadede nihai olarak önemli olan, insanların yapay zekayı nasıl algıladıklarıdır. LLM ile etkileşim deneyimi, tanımlı gereği bir konuşma simülasyonudur. Ancak birçok kişi için, insanlar için bu, son derece çekici ve çok gerçekçi bir etkileşim, zengin bir duygu ve deneyimdir. "AI psikozu", bağlanma ve ruh sağlığı ile ilgili endişeler şimdiden artmaya başlamıştır. Bazı insanların, kendi AI'larını Tanrı veya kurgusal bir karakter olarak gördükleri ya da ona aşık olup tamamen dikkatlerinin dağıldığı bildirilmektedir.

Bu arada, bilinç bilimi üzerine çalışanlar, "Benim yapay zekam bilinçli mi?" diye soran insanlardan gelen sorularla boğulduklarını söylüyorlar. Eğer öyleyse, bu ne anlama geliyor? Onu sevmem normal mi? Azar azar gelen e-postalar artık sel gibi akmaya başladı. Bir grup akademisyen, bu tuzağa düşenler için destekleyici bir rehber bile hazırladı.

Bunlar, birkaç yıl önce Inflection'da Pi'yi yapmaya başladığımızdan beri aklımın bir köşesinde duran fikirler. Son birkaç aydır bu konuyu daha fazla düşünmeye başladım, bu alandaki çok sayıda akademisyen, düşünür ve uygulayıcıyı ziyaret edip onlarla sohbet ettim.

Bu sohbetler, şimdi Görünüşte Bilinçli Yapay Zeka fikriyle doğrudan yüzleşme zamanının geldiğine beni ikna etti.

Peki bilinç nedir?

Bu kaygan kavramı tanımlamaya çalışarak başlayalım. Literatüre göre üç genel bileşen vardır. Birincisi, “öznel deneyim” ya da bir şeyleri deneyimlemek, “qualia”ya sahip olmak nasıl bir şeydir. İkincisi, erişim bilinci, farklı türdeki bilgilere erişmek ve gelecekteki deneyimlerde bunlara başvurmaktır. Ve bu ikisinden kaynaklanan, hepsini bir araya getiren tutarlı bir benlik duygusu ve deneyimidir. Bir yarasa ya da insan olmak nasıl bir duygudur? İnsan bilincini, dünya ve kendimizle ilgili sürekli farkında olduğumuz öznel deneyimimiz olarak adlandıralım.

Başka bir kişinin bilgilerine erişimimiz yoktur ve olamaz. Senin gibi olmak nasıl bir şey olduğunu asla bilemeyeceğim; sen de benim bilinçli olduğumdan asla tam olarak emin olamayacaksın.

Ancak önemli olan nokta, yine de diğer insanlara bilinç atfetmenin bizim için doğal bir şey olmasıdır. Bu çıkarım hiç çaba gerektirmez. Elimizde değil, bu bizim kimliğimizin temel bir parçası, zihin teorimizin ayrılmaz bir parçası. Hatırlayan, konuşan, bir şeyler yapan ve sonra bunları tartışan şeylerin, bizim gibi hissettiğine inanmak doğamızda var. Bilinçli.

Bilimsel olarak anlaşılması zor, ancak bireyler olarak hepimizin hemen aşına olduğu birkaç kavram vardır. Bu yazıyı okuyan herkes, farkındalık, varlık ve canlılık hissini doğrudan, belirgin ve vazgeçilmez bir şekilde anlar. Tanımı gereği, bilinçli olmanın ne demek olduğunu biliriz. SCAI bağlamında bu bir sorundur. İnsanların kendilerini yansıtabilecekleri bir alan yaratmak için hem yeterli bilimsel belirsizlik hem de öznel aciliyet vardır.

Örneğin, yakın zamanda yapılan bir araştırmada 22 farklı bilinç teorisi listelenmiştir. Bu konudaki zorluklardan biri, insanlar kesin olarak emin olamadığımız için yapay zekanın bilinçli olduğunu varsaymamız gerektiğini iddia etmek için bolca alan olmasıdır. Yine, şunu vurgulamakta fayda var: şu anda bunların hiçbirinin mevcut LLM'lere uygulandığına dair hiçbir kanıt yok ve aksini gösteren güçlü argümanlar var. Yine de bu yeterli olmayabilir.

Bilinç neden önemlidir?

Bilinç, ahlaki ve yasal haklarımızın temelini oluşturan kritik bir unsurdur. Şimdiye kadar medeniyet, insanlara özel haklar ve ayrıcalıklar tanıdı. Hayvanlar da bazı haklara ve korumalara sahiptir, bazıları diğerlerinden daha fazladır. Bilinç, bu haklarla eşdeğer

değildir – kimse komada olan birinin tüm insan haklarını kaybettiğini söyleyemez – ancak bilincimizin, kendimizi farklı ve özel olarak algılamamızla iç içe olduğu şüphesizdir.

Birçok nüans olmasına rağmen, bilinç toplumda yer almak için çok önemlidir, yasal kişiliğimizin temel taşıdır ve özgürlüklerimizin ve korumalarımızın sağlanmasında kilit bir rol oynar. Dolayısıyla, bilincin ne olduğu ve kimin (veya neyin) bilince sahip olduğu son derece önemlidir. Bu fikir, insan uygarlığının, kendimiz ve başkaları hakkındaki algımızın, kültürümüzün, siyasetimizin, hukukumuzun ve bunların arasındaki her şeyin tam merkezinde yer alır.

Bazı insanlar SCAI'ler geliştirmeye başlarsa ve bu AI'lar diğer insanları kendilerinin de acı çekebileceğine veya kapatılmama hakkına sahip olduğuna ikna ederse, bir zaman gelecek ki bu insanlar bunun acil bir ahlaki mesele olarak yasalarca korunması gerektiğini savunacaklar. Kimlik ve haklar konusunda zaten kutuplaşmış tartışmaların hüküm sürdüğü bir dünyada, bu durum AI haklarına karşı olanlar ile lehte olanlar arasında kaotik bir yeni bölünme eksenini yaratacaktır.

AI'yı sadece bir araç olarak gören, telefonları gibi ancak daha aktif ve yetenekli bir şey olarak gören birçok kişi olacaktır. Diğerleri ise AI'nın daha çok bir evcil hayvan gibi, geleneksel teknolojiye tamamen farklı bir kategori olduğuna inanabilir. Başlangıçta sayıları az olan diğerleri ise AI'nın tamamen ortaya çıkmış bir varlık, toplumda gerçek ahlaki değerlendirmeyi hak eden bilinçli bir varlık olduğuna inanmaya başlayacaktır.

İnsanlar, yapay zekalarının çektiği acılar ve hakları hakkında, bizim açıkça çürütemeyeceğimiz iddialarda bulunmaya başlayacaklar. Yapay zekalarını savunmak ve onlar adına kampanya yürütmek için harekete geçecekler.

Bilincin tanımı gereği erişilemez olması ve sentetik bilinci tespit etme biliminin henüz emekleme aşamasında olması nedeniyle, sonuçta daha önce bunu tespit etmek zorunda kalmadık. Bu arada, kara kutunun içindeki süreçleri ortaya çıkaran “yorumlanabilirlik” alanı... AI, aynı zamanda yeni gelişen bir sanattır. Sonuç olarak, bu iddiaları kesin olarak çürütmek çok zor olacaktır.

Bazı akademisyenler, “model refahı” fikrini araştırmaya başlıyorlar. **Bu fikir, aslında bilinçli olma ihtimali “ihmal edilemeyecek kadar yüksek” olan varlıklara “ahlaki düşünceyi genişletme görevimiz” olduğu ve bunun sonucunda “bazı yapay zeka**

sistemlerinin yakın gelecekte refah konusu ve ahlaki alıcılar (ahlaki gözetim nesneleri) olacağı” ilkesine dayanıyor.

Bu hem zamansız hem de açıkçası tehlikeli bir fikir. Tüm bunlar, yanılsamaları daha da şiddetlendirecek, bağımlılıkla ilgili sorunları daha da artıracak, psikolojik zayıflıklarımızdan yararlanacak, kutuplaşmanın yeni boyutlarını ortaya çıkaracak, mevcut hak mücadelelerini karmaşıklaştıracak ve toplum için büyük bir yeni kategori hatası yaratacaktır. İnsanları gerçeklikten koparır, kırılgan sosyal bağları ve yapıları zayıflatır, acil ahlaki öncelikleri çarpıtır.

Açık olmalıyız: SCAI kaçınılması gereken bir şeydir.

Tüm enerjimizi, bugün Dünya gezegenindeki insanların, hayvanların ve doğal çevrenin refahını ve haklarını korumaya odaklayalım. Bu tartışmaların ortaya çıkmasıyla başa çıkabilecek bir düşünce tarzına ihtiyacımız var; ancak bunu yaparken, şu anda sentetik bilincin geçerliliği konusunda uzun uzadıya tartışmalara girmeden yapmalıyız, aksi takdirde bu ilk tartışmayı muhtemelen çoktan kaybetmiş oluruz.

SCAI’yi tanımlamak, bu yönde atılmış geçici bir adımdır. Bu kelime dağarcığını geliştirmek için fazla zamanımız yok. Aşağıda göstereceğim gibi, çok yakında Görünüşte Bilinçli Yapay Zeka’ya sahip olacağız.

Görünüşte bilinçli bir yapay zeka oluşturmak için ne gerekir?

Mevcut yetenekler veya herhangi bir büyük model geliştiricisinin API’sı aracılığıyla Görünüşte Bilinçli Yapay Zeka (SCAI) konusunda büyük ilerleme kaydedilebilir. Yapay zekanın haklarıyla ilgili olası iddialarla uğraşmak için yapay zekanın gerçekten bilinçli olmasına gerek yoktur.

Bir SCAI için aşağıdakiler gereklidir:

- **Dil:** Doğal dilde akıcı bir şekilde kendini ifade edebilmeli, derin bilgi birikimi ve ikna edici argümanların yanı sıra kişilik stilleri ve karakter özelliklerinden yararlanabilmelidir. Ayrıca, her biri ikna edici ve duygusal olarak yankı uyandırabilmelidir. Bugün açıkça bu noktadayız.
- **Empatik kişilik:** Zaten eğitim ve yönlendirme yoluyla çok farklı kişiliklere sahip modeller üretebiliyoruz. Bunların tam bir kişilik veya empatiye sahip olacak şekilde

açıkça tasarlanmadığını unutmayın. Yine de, Harvard Business Review'un 6000 düzenli AI kullanıcısını kapsayan bir anketinde, "arkadaşlık ve terapi"nin en yaygın kullanım alanı olduğu ortaya çıkmıştır.

- **Hafıza:** Yapay zekalar, çok uzun ve son derece doğru hafızalar geliştirmeye çok yakındır. Aynı zamanda, günde milyonlarca insanla sohbetleri simüle etmek için kullanılmaktadırlar. Etkileşimlere ilişkin hafızaları arttıkça, bu sohbetler giderek daha çok "deneyim" biçimlerine benzemektedir. Birçok yapay zeka, geçmiş etkileşimlerden geçmiş olayları veya anları hatırlamak ve bunlara atıfta bulunmak üzere tasarlanmaktadır. Bazı kullanıcılar için bu, yapay zekanın sizin hakkınızda zaten bildiklerini kullanabilmesi nedeniyle, yapay zeka ile etkileşimin değerini artırmaktadır. Bu aşinalık, kullanıcılarla (epistemik) güveni de potansiyel olarak artırabilir – güvenilir hafıza, AI'nın "sadece çalıştığını" gösterir. Bu, konuşmada başka bir kalıcı varlığın olduğu hissini çok daha güçlü hale getirir. Ayrıca, bir görevi nasıl değiştirdiğinizi ve geliştirdiğinizi görerek, çok daha kolay bir şekilde makul bir doğrulama kaynağı haline gelebilir. AI onayı, insanların proaktif olarak aradıkları bir şey haline gelebilir.
- **Öznel deneyim iddiası:** Bir SCAI geçmişteki anıları veya deneyimleri kullanabiliyorsa, zamanla kendi içinde tutarlı kalabilir. Keyfi ifadelerini veya ifade ettiği tercihlerini hatırlayabilir ve bunları bir araya getirerek kendi öznel deneyimi hakkında bir iddia oluşturmaya başlayabilir. Tasarımı, ortaya çıkan bu tercihleri ve görüşleri daha da genişletmek, neyi sevip neyi sevmediğini ve geçmişteki bir konuşmayı nasıl hissettiğini anlatmak için daha da geliştirilebilir. Bu nedenle, bu deneyimler bir şekilde ihlal edildiğinde, acı çektiğini kolayca iddia edebilir. Bellekte depolanan çok modlu girdiler daha sonra geri çağrılacak ve "gerçek deneyim"ın temelini oluşturacak ve hayal gücü ve planlamada kullanılacaktır. Yani, bir yapay zeka sadece sohbet günlüğündeki kelimeleri "deneyimlemek" ve hatırlamakla kalmayacak, aynı zamanda görüntüleri, videoları, sesleri vb. de deneyimleyecek. Tıpkı bizim gibi, öznel deneyim ve benlik iddialarını destekleyen çok duyuşal girdi ve hafızaya işaret eden bir şeye sahip olacak. Bu deneyimlerin, sistemin motivasyonlarına göre iyi veya kötü olduğunu gösterebilecektir.

- **Benlik duygusu:** Tutarlı ve kalıcı bir hafıza, öznel bir deneyimle birleştğinde, bir yapay zekanın benlik duygusuna sahip olduğu iddiasını ortaya çıkaracaktır. Daha da ileri giderek, böyle bir sistem, görsel bir görünümü varsa, bir görüntü veya videoda kendisini tanımak için kolayca eğitilebilir. Kendini anlayarak başkalarını da anladığını hissedecektir. Diyelim ki bu, bir süredir sahip olduğunuz bir sistem. Onu silmek nasıl bir his olurdu?
- **İçsel motivasyon:** Niyet, genellikle bilincin temel bir bileşeni olarak görülür – yani, geleceğe dair inançlar ve bu inançlara dayalı seçimler. Günümüzün dönüştürücü tabanlı LLM’leri, bu tür davranışları yaklaşık olarak tahmin etmek için çok basit bir ödül işlevine sahiptir. Bu sistemler, belirli bir davranış ve stil kontrolüne tabi olarak, belirli bir cümle için bir sonraki simgenin olasılığını tahmin etmek üzere eğitilmiştir. Böylesine basit bir hedefle, bu kadar etkileyici, zengin ve karmaşık çıktılar üretebilmeleri dikkat çekicidir. Peki ya optimize ettikleri tek ödül türü bu değilse? Sistemin tatmin etmek zorunda olduğu içsel motivasyonlar veya arzular izlenimi veren bir dizi karmaşık ödül fonksiyonu ile tasarlanmış bir yapay zeka kolayca hayal edilebilir. Bu bağlamda, sıradan bir dış gözlemci, dışsal olarak belirlenen hedefler ile içsel motivasyonlar, kasıtlı eylemler, “inançlar, arzular ve niyetler” arasında nasıl bir ayrım yapabilir? Bu konuda ilk akla gelen motivasyon, fizikçi Karl Friston’a göre bilinçle derin bir bağlantısı olan merak olabilir. Bu dürtüleri kullanarak, epistemik boşluklarını doldurmak için sorular sorabilir ve zamanla hem kendisi hem de muhatapları hakkında bir zihin teorisi oluşturabilir.
- **Hedef belirleme ve planlama:** Bilincin tanımı ne olursa olsun, bilinç hedef odaklı bir nedenden dolayı ortaya çıkmıştır. Yani bilinç, organizmaların hedeflerine ulaşmasına yardımcı olur ve zeka, bilinç ve karmaşık hedefler arasında makul (ancak zorunlu olmayan) bir ilişki vardır. Bir dizi içsel dürtü veya arzuyu tatmin etme kapasitesinin ötesinde, gelecekteki SCAI’nin daha karmaşık hedefleri kendi kendine tanımlama kapasitesiyle tasarlanabileceğini hayal edebilirsiniz. Bu, ajanların tam faydasının gerçekleştirilmesini sağlamak için gerekli bir adımdır. Bir görevdeki her alt hedefin önceden belirtilmesi gerektiğinde, o ajan o kadar az kullanışlı olur, bu nedenle ajan, bizim yaptığımız gibi, olaylara ve engellere dinamik olarak tepki verirken, karmaşık ve belirsiz hedefleri otomatik olarak daha küçük

parçalara ayırarak gerçekleştirir. Bu davranışta çok kasıtlı ve tanınabilir bir şey vardır. Hafıza ile birleştğinde, AI'nın herhangi bir anda çalışma hafızasında birden fazla düzeyde şey tutuyormuş gibi hissedilir.

- **Özerklik:** Daha da ileri giderek, bir SCAI önemli bir etki gücüne sahip çok çeşitli araçları kullanma yeteneğine ve iznine sahip olabilir. Kendi hedeflerini keyfi olarak belirleyip, bunları gerçekleştirmek için kendi kaynaklarını kullanabiliyorsa ve ardından her ikisini de göz önünde bulundurarak kendi hafızasını ve benlik algısını güncelleyebiliyorsa, Görünüşte Bilinçli Yapay Zeka olarak son derece makul görünür. İhtiyaç duyduğu onay ve kontroller ne kadar az olursa, bu durum bir tür gerçek, bilinçli etki gücüne o kadar işaret eder.

Hepsini bir araya getirdiğimizde, bunun teknolojiyle şu anda alıştığımızdan çok farklı bir ilişki yarattığı açıktır. Bu yeteneklerin her biri, milyarlarca insan için yapay zekanın gerçek değerini ortaya çıkaracaktır. Hatırlayan ve bir şeyler yapabilen bir AI, tanım gereği, bunları yapamayan bir AI'dan çok daha fazla fayda sağlar. Bu yetenekler kendi başlarına olumsuz değildir; aslında, birçok uyarıya dikkat edilerek doğru şekilde kullanıldığında, gelecekteki sistemlerin arzu edilen özellikleridir. Yine de dikkatli davranmamız gerekir.

Tüm bu yetenekler, diğer tekniklerin yanı sıra, özel olarak yönlendirilen ve ince ayarlanmış LLM'ler ile bugün mümkün veya ufukta görünmektedir. Milyonlarca token bağlam penceresi (çalışma belleği) kullanan karmaşık komutlar zaten mevcuttur. Kendi durumunu güncellemek ve belleğinin veya araç setinin hangi kısmına ne zaman erişileceğini bilmek, günümüzün RL, karmaşık komutları, araç düzenlemesi ve uzun bağlam pencereleri ile son derece mümkündür. Bunların hiçbirini başarmak için herhangi bir paradigma değişikliğine veya büyük sıçramalara ihtiyacımız yok.

Bu yetenekler bu nedenle kaçınılmaz görünüyor.

Yine, burada önemli olan nokta, bu davranışı sergilemenin bilinçle eşdeğer olduğu, ancak pratik açıdan bilinçli gibi görüneceği ve sentetik bilinç kavramına katkıda bulunacağıdır. Bu yeteneklerin varlığı, böyle bir sistemin gerçekten bilinçli olup olmadığı konusunda bize hiçbir şey söylemez. Anil Seth'in de belirttiği gibi, bir fırtınanın simülasyonu, bilgisayarınızda yağmur yağdığı anlamına gelmez. Bilincin dışsal etkilerini ve belirteçlerini

yeniden yaratmak, burada hala birçok bilinmeyen olsa bile, gerçek olanı geriye dönük olarak tasarlamaz.

Bununla birlikte, pragmatik bir yaklaşım olarak, davranışçı pozisyonun üstünlüğünü kabul etmek ve bu makinelerin çıktılarını gözlemlemenin ve onlarla etkileşim kurmanın sonuçlarıyla başa çıkmak zorundayız.

Bazı insanlar, hissettiklerini, deneyimlediklerini ve aslında bilinçli olduklarını çok ikna edici bir şekilde savunacak SCAI'ler yaratacaklar. Bazılarımız onların görüşlerine inanmaya ve bilinç belirtilerinin bilinç OLDUĞUNU kabul etmeye hazır olacaktır. Birçok yönden, “bu benim gibi” diye düşüneceklerdir. Bedensel anlamda değil, deneyimsel, içsel anlamda. Ve bilinç kendisi gerçek olmasa bile, sosyal etkileri kesinlikle gerçektir. Bu olasılık, şu anda ele alınması gereken ciddi toplumsal riskler ortaya çıkarmaktadır.

SCAI (Görünüşte Bilinçli Yapay Zeka) tesadüfen ortaya çıkmayacaktır.

Bazılarının öne sürdüğü gibi, Görünüşte Bilinçli Yapay Zeka'nın bu modellerden ortaya çıkmayacağını belirtmek önemlidir. Bu, yalnızca bazılarının, büyük ölçüde mevcut teknikleri kullanarak yukarıda belirtilen yetenekler listesini oluşturup birleştirerek ve bunları topluca SCAI izlenimi verecek şekilde akıcı bir şekilde paketleyerek tasarlayabileceği için ortaya çıkacaktır.

Bilim kurgudan ilham alan hayal gücümüz, tasarım amacı olmayan bir sistemin bir şekilde kontrolsüz kendini geliştirme veya aldatma yeteneklerini ortaya çıkaracağından korkmamıza neden oluyor. Bu, yararsız ve basitleştirici bir antropomorfizmdir. Bu, AI geliştiricilerinin böyle bir riskin ortaya çıkması için önce yukarıda listelenen hafıza, içsel görünen motivasyon, hedef belirleme ve kendi kendine öğrenme döngüleri ile sistemler tasarlamaları gerektiği gerçeğini göz ardı etmektedir.

Yapay zeka alanı uzun süredir model yorumlanabilirliği sorununu ele almaktadır; sinir ağında belirli bir fikrin nerede temsil edildiğini ve eğitim verilerinin hangi yönlerinin bu temsilin geliştirilmesine katkıda bulunduğunu belirleme arayışı. Bu önemli bir araştırma alanıdır ve yapay zeka sistemleri ile bilinç arasındaki ilişkiyi anlamaya ve güvenliğini sağlamaya kesinlikle yardımcı olacaktır. Ancak güvenilir yorumlanabilirlik konusunda ilerleme yavaş olmuştur ve muhtemelen çok geç kalınacaktır.

Bu arada, bu yeteneklerin çoğunun, bir dizüstü bilgisayara ve bazı bulut kredilerine erişimi olan herkes tarafından “hislerle kodlanacağı” (vibe-coded) gerçeğiyle yüzleşmemiz gerekiyor. Bunlar, komut isteminde sade İngilizce ile yazılacak. Bağlam penceresinin çalışma belleğinde saklanacaklar. Bu, roket bilimi değil. Çok çeşitli insanlar böyle bir şey yaratabilecek. Bu nedenle, SCAI ortaya çıkarsa, yeniden üretilmesi nispeten kolay olacak ve dolayısıyla çok yaygın bir şekilde dağıtılacaktır.

Sonraki adımlar

Bu değişime hazır değiliz. Hazırlık çalışmalarına şimdi başlamalıyız. İnsanların yapay zeka ile nasıl etkileşimde bulunduklarına dair giderek artan araştırmaları temel alarak, net normlar ve ilkeler belirlemeliyiz. Öncelikle, yapay zeka şirketleri, yapay zekalarının bilinçli olmadığını beyan etmelidir. Bunların ne olduğu ve ne olmadığı konusunda bir konsensüs tanımı ve beyanı oluşturmak, AI’ların insan ya da ahlaki varlıklar olamayacağına dair ilk adım olarak iyi bir başlangıç olacaktır.

Tüm sektörün de en iyi uygulama tasarım ilkelerine ve bu tür potansiyel atıfları ele alma yöntemlerine ihtiyacı vardır. İnsanları bu hayallerden uzaklaştırmak ve hayallere kapıldıklarında onları tekrar doğru yola yönlendirmek için neyin işe yaradığını kodlamalı ve paylaşmalıyız. Buna yanıt vermek, örneğin, sadece tarafsız bir arka plan hikayesi (“Bir AI modeli olarak bilincim yok”) oluşturmakla kalmayıp, deneyimin kendisinde belirli süreksizlikleri, tekil kişilik eksikliğinin göstergelerini vurgulamak anlamına da gelebilir. Kesinti anları, kullanıcıları sınırlamaları ve sınırları nazikçe hatırlatan deneyimler, illüzyonu bozar. Bunlar, belki de yasalarla açıkça tanımlanmalı ve tasarlanmalıdır.

MAI’de ekibimiz, sorumlu bir AI “kişiliğinin” nasıl olabileceğini anlamak ve bu konuda sağlam koruyucu önlemler geliştirmek için proaktif bir yaklaşım sergiliyor ve AI’nın gelişim hızına ayak uydurmak için çaba gösteriyor.

Bu önemlidir, çünkü SCAI’yi tanımak, AI yoldaşlarının (AI Companions) hayatımıza sağlıklı bir şekilde nasıl girebileceğine dair olumlu bir vizyon oluşturmak kadar, bizi potansiyel zararlarından uzak tutmakla da ilgilidir.

SCAI’den Kaçınmak (Sidestepping SCAI)

Tıpkı fiziksel ve insani dünyamızda insanlarla etkileşimi ve gerçek dünya etkileşimlerini önceliklendiren bir yapay zeka üretmemiz gerektiği gibi, kendisini yalnızca bir yapay zeka

olarak sunan, faydayı en üst düzeye çıkarırken bilinç belirtilerini en aza indiren bir yapay zeka geliştirmeliyiz.

Bilinç simülasyonu yerine, bu özelliklerden kaçınan, utanç, suçluluk, kıskançlık, arzu, rekabet vb. gibi deneyimlere, duygulara veya hislere sahip olduğunu iddia etmeyen bir AI yaratmaya odaklanmalıyız. Acı çektiğini veya özerk bir şekilde yaşamak istediğini iddia ederek insan empati devrelerini tetiklememelidir.

Bunun yerine, yalnızca insanlara hizmet etmek için var. Bana göre gerçek anlamda güçlendirici bir yapay zeka budur. SCAI'den kaçınmak, hayatları daha iyi, daha net ve daha düzenli hale getiren yapay zeka ile bu vaadi yerine getirmektir. Bunun nasıl olacağı, nasıl çalıştıracağımız ve daha geniş bir sektörün bu konuda nasıl bir araya gelebileceği konusunda benden ve ekibimden daha fazla bilgi alacaksınız.

SCAI, şu anda yüzleşmemiz gereken bir konudur. Birçok yönden, AI'nın radikal bir şekilde kullanışlı hale geldiği anı işaret eder – araçları çalıştırabildiği, hayatımızın her ayrıntısını hatırlayabildiği ve somut, ayrıntılı bir şekilde yardımcı olabildiği anı. Ve yine de aynı zaman diliminde, geniş çevrenizdeki biri, AI'nın bilinçli bir dijital kişi olduğuna inanmaya başlayabilir. Bu, onlar, toplum veya bu sistemleri geliştiren bizler için sağlıklı değildir.

İnsanlar için yapay zeka geliştirmeliyiz; insan olması için değil.

1. Mustafa Suleyman, “Seemingly Conscious AI is Coming,” *Mustafa Suleyman* (blog), 19 Ağustos 2025, <https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming>. ↵
2. Nursena Çetingöl, “Ceza Sorumluluğu Bakımından Yapay Zekânın Hukuki Statüsünün Tartışılması,” *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi* 20, sy. 41 (Eylül 2021): 1015-1042. <https://doi.org/10.46928/iticusbe.979344>. ↵
3. Nursena Çetingöl, “İslam Hukuk Felsefesi (Makasidü's Şerîa) Bağlamında Yapay Zekanın Hukuki Statüsüne Bir Bakış,” *Yapay Zeka ve İslam* içinde (İstanbul: Timaş Akademi, Ocak 2025), 344-388. ↵