SÜPER ZEKÂ

NICK BOSTROM



SÜPER ZEKÂ

YAPAY ZEKÂ UYGULAMALARI, TEHLİKELER VE STRATEJİLER



NICK BOSTROM

2. BASKI

SÜPER ZEKÂ

YAPAY ZEKÂ UYGULAMALARI,

TEHLİKELER VE STRATEJİLER

Koç Üniversitesi Yayınları: 184 UÇBEYLERİ
BİLİM | TEKNOLOJİ

Süper Zekâ: Yapay Zekâ Uygulamaları, Tehlikeler, Stratejiler
Nick Bostrom

İngilizceden çeviren: Ferit Burak Aydar
Yayına hazırlayan: Feza Birtaş
Düzelti: Gözde Yalım
Kitap tasarımı: Gökçen Ergüven
Kapak tasarımı: Melike Oran

Superintelligence: Paths, Dangers, Strategies was originally published in English in year 2016. This translation is published by arrangement with Oxford University Press. Koç University Press is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

© Nick Bostrom 2014
© Türkçe yayın hakları: Koç Üniversitesi Yayınları, 2016

1. Baskı: İstanbul, Aralık 2018
2. Baskı: İstanbul, Mayıs 2020

Bu kitabın yazarı, eserin kendi orijinal yaratımı olduğunu ve eserde dile getirilen tüm görüşlerin kendisine ait olduğunu, bunlardan dolayı kendisinden başka kimsenin sorumlu tutulamayacağını, eserde üçüncü şahısların haklarını ihlal edebilecek kısımlar olmadığını kabul eder.

Baskı: 12. matbaa
İbrahim Karaoğlanoğlu Cad. 35/1 Kâğıthane/İstanbul
Sertifika no: 46618
+90 212 281 2580

Koç Üniversitesi Yayınları
Rumelifeneri Yolu 34450 Sarıyer/İstanbul
kup@ku.edu.tr • www.kocuniversitypress.com • www.kocuniversitesiyayinlari.com
Sertifika no: 18318
+90 212 338 1000

Koç University Suna Kıraç Library Cataloging-in-Publication Data

Bostrom, Nick, 1973-

Süper zeka : yapay zekânın yolları, tehlikeleri ve stratejileri = Superintelligence : paths, dangers, strategies / Nick Bostrom ; çeviren Ferit Burak Aydar ; yayına hazırlayan Feza Birtaş. -- İkinci baskı -- İstanbul : Koç Üniversitesi, 2018.

392 pages : illustrations, graphics ; 16,5x24 cm.-- Koç Üniversitesi Yayınları ; 184. Bilim, Teknoloji

Includes bibliographical references and index.

ISBN 978-605-2116-78-4

1. Artificial intelligence--Philosophy. 2. Artificial intelligence--Social aspects. 3. Islamic learning and scholarship--Mali--Tombouctou. 4. Artificial intelligence--Moral and ethical aspects. 5. Computers and civilization. 6. Cognitive science. I. Aydar, Ferit Burak. II. Birtaş, Feza. III. Title.

Q335.B6720 2018

Süper Zekâ

Yapay Zekâ Uygulamaları, Tehlikeler ve Stratejiler

NICK BOSTROM

İngilizceden çeviren: Ferit Burak Aydar

KÜY

İçindekiler

Şekil, Tablo ve Kutu Listesi 9

Serçelerin Bitmemiş Masalı 11

Önsöz 13

Teşekkür 15

BİRİNCİ BÖLÜM 17

Geçmişteki Gelişmeler, Mevcut Yeterlilikler

Büyüme tarzları ve büyük tarih 17

Büyük umutlar 19

Umut ve umutsuzluk mevsimleri 21

En gelişkin teknoloji 26

Makine zekâsının geleceğine dair fikirler 34

İKİNCİ BÖLÜM 39

Süper Zekâya Giden Yollar

Yapay zekâ 39

Tüm beyin emülasyonu 48

Biyolojik biliş 54

Beyin-bilgisayar arayüzleri 64

Ağlar ve örgütlenmeler 67

Özet 69

ÜÇÜNCÜ BÖLÜM 71

Süper Zekânın Biçimleri

Hızlı süper zekâ 71

Kolektif süper zekâ 73

Nitelikli süper zekâ 75

Doğrudan ve dolaylı kavrama gücü 77

Dijital zekâ için üstünlük kaynakları 78

DÖRDÜNCÜ BÖLÜM 81

Bir Zekâ Patlamasının Kinetiği

Havalanmanın hızı ve zamanlaması 81

Uzlaşmazlık 85

Optimizasyon gücü ve patlayıcılık 93

BEŞİNCİ BÖLÜM 99
Belirleyici Stratejik Üstünlük
Önde giden proje belirleyici bir stratejik üstünlük elde edecek midir? 100
Başarılı proje ne kadar büyük olacaktır? 104
Belirleyici stratejik üstünlükten tekele 108
ALTINCI BÖLÜM 113
Bilişsel Süper Güçler
İşlevsellikler ve süper güçler 113
Bir yapay zekânın devralması senaryosu 117
Doğa ve aracılar karşısında güce sahip olmak 122
YEDİNCİ BÖLÜM 129
Süper Zeki İstenç
Zekâ ile güdüler arasındaki ilişki 129
Araçsal yakınsaklık 133
SEKİZİNCİ BÖLÜM 141
Dön Dolaş Yine Kötü Son mu?
Bir zekâ patlamasının varsayılan sonucu olarak varoluşsal felaket? 141
Tehlikeli dönemeç 142
Kötücül başarısızlık örnekleri 146
DOKUZUNCU BÖLÜM 155
Kontrol Problemi
İki aracılık problemi 155
Yeterlilik kontrol yöntemleri 157
Güdülenme seçilimi yöntemleri 167
Özet 172
ONUNCU BÖLÜM 175
Kâhinler, Cinler, Hükümdarlar, Aletler
Kâhin 175
Cinler ve hükümdarlar 178
Alet yapay zekâlar 181
Karşılaştırma 185
ON BİRİNCİ BÖLÜM 191
Çokkutuplu Senaryolar
Atlar ve insanlar 192
Bir algoritmik ekonomide hayat 198
Geçiş sonrasında bir tekel kurulması mı? 209

ON İKİNCİ BÖLÜM 219
Değerler Edinmek
Değer yükleme problemi 219
Evrimsel seçilim 221
Takviyeli öğrenme 222
Çağrışımsal değer eklenmesi 223
Güdülenme iskeleti 225
Değer öğrenme 226
Emülasyon modülasyonu 238
Kurum tasarımı 240
Özet 244
ON ÜÇÜNCÜ BÖLÜM 247
Seçme Kıstaslarını Seçme
Dolaylı normatiflik ihtiyacı 247
Bütünlüklü dışkestirilmiş irade 249
Ahlak modelleri 256
Ne kastediyorsam onu yap! 259
Bileşen listesi 260
Yeterince yakınlaşmak 266
ON DÖRDÜNCÜ BÖLÜM 269
Stratejik Resim
Bilim ve teknoloji stratejisi 269
Yollar ve kolaylaştırıcılar 281
İşbirliği 288
ON BEŞİNCİ BÖLÜM 299
Kırılma Anı
Miadı olan felsefe 299
Ne yapmalı? 300
İnsanın içindeki en iyi, lütfen ayağa kalksın! 303
Sonsöz 305
Sözlükçe 309
Notlar 313
Kaynakça 363
Dizin 387

Şekil, Tablo ve Kutu Listesi

Şekil Listesi

1 Dünya GSYH'nın uzun vadeli tarihi 18

2 HLMI'nın uzun vadedeki genel etkisi 38

3 Süper bilgisayar performansı 46

4 Elektronlu mikroskop görüntülerinden üçboyutlu nöroanatomi yaratılması 49

5 Tüm beyin emülasyonu yol haritası 53

6 Yazım denetimi yapılmış genomlar için bir metafor olarak bileşik yüz 60

7 Havalanmanın şekli 82

8 Daha az insan merkezci bir ölçek mi? 90

9 Zekâ patlamasının basit bir modeli 97

10 Bir yapay zekânın devralması senaryosundaki araştırmalar 118

11 Varsayımsal bir bilge tekel için bazı olanaklı yörüngelerin şematik örneği 126

12 Uzaylı güdülerinin insanbiçimlileştirilmesinin sonuçları 130

13 İlkin yapay zekâ mı yoksa tüm beyin emülasyonu mu? 285

14 Yapay zekâ teknoloji yarışlarındaki risk düzeyleri 290

Tablo Listesi

1 Oyun oynayan yapay zekâ 29

2 İnsan düzeyinde makine zekâsına ne zaman ulaşılacak? 35

3 İnsan düzeyinde makineden süper zekâya uzaklık ne kadardır? 37

4 Tüm beyin emülasyonu için gerekli yeterlilikler 50

5 Bir grup embriyo arasında seçilimden azami IQ kazanımı 56

6 Farklı senaryolarda genetik seçilimin mümkün olan etkileri 59

7 Stratejik açıdan önemli bazı teknoloji yarışları 103

8 Süper güçler: Stratejik açıdan ilintili bazı görevler ve buna denk düşen yetenekler 116

9 Farklı tetikleyici örnekleri 166

10 Kontrol yöntemleri 173
11 Farklı sistem kastlarının özellikleri 188
12 Değer yükleme tekniklerinin özeti 245
13 Bileşen listesi 261

Kutu Listesi

1 Optimum bir Bayesçi aracı 27
2 2010 Flash Crash'ı 36
3 Evrimi tekrarlamak için neler gereklidir? 42
4 Bir zekâ patlamasının kinetiği üzerine 95
5 Teknoloji yarışları: Bazı tarihsel örnekler 102
6 Posta siparişli DNA senaryosu 120
7 Kozmik sermayemiz ne kadar büyük? 123
8 İnsani zapturapt 163
9 Körü körüne aramanın tuhaf çözümleri 186
10 Değer öğrenmenin formelleştirilmesi 228
11 Dostane olmak isteyen bir yapay zekâ 234
12 Yakın döneme ait iki (yarı pişmiş) fikir 235
13 Riskte dibe doğru yarış 289

Serçelerin Bitmemiş Masalı

Yuva yapma mevsimi gelip çatmıştı. Serçeler günlerce süren ağır mesainin ardından akşam güneşinin altında oturmuş, dinleniyor, cıvıldıyorlardı.

"Biz çok küçük ve zayıfız. Bir düşünsenize, bize yuva yapmamızda yardım edecek bir baykuşumuz olsaydı, hayat ne kadar kolay olurdu!"

"Evet!" dedi bir diğeri. "Ayrıca ondan yaşlılarımıza ve gençlerimize göz kulak olması için yararlanabilirdik."

"Bize tavsiyeler verir, civardaki kedilere karşı bakarak olurdu," dedi bir üçüncüsü.

Ardından yaşlı olan kuş Pastus söze girdi: "Her tarafa gözcüler yollayalım ve bir yerlerde terk edilmiş bir baykuş yavrusu ya da ne bileyim bir yumurta bulmaya çalışalım. Bir karga yavrusu ya da bebek sansar da olur. Bu, şimdiye kadar ya da en azından şu öte taraftaki bahçede Sınırsız Tahıl Pavyonu açıldığından bu yana başımıza gelen en iyi şey olabilir."

Tüm serçe sürüsü heyecanlanmıştı. Hep bir ağızdan ciğerleri elverdiğince cıvıldadılar.

İçlerinden sadece huysuz, tek gözlü bir serçe olan Scronkfinkle, bu çabanın akıllıca olduğuna ikna olmamıştı. Şöyle dedi: "Bu kesinlikle bizim sonumuz olacak. İlk önce baykuş evcilleştirme ve baykuş ehlileştirme sanatına kafa yorup, ondan sonra böyle bir hayvanı aramıza getirmek daha doğru olmaz mı?"

Pastus şöyle yanıt verdi: "Bir baykuşu ehlileştirmek kulağa son derece zor bir şey gibi geliyor. Bir baykuş yumurtası bulmak zaten zor olacak. O yüzden oradan başlayalım. Bir baykuş yetiştirdikten sonra, diğer zorluğu gündemimize alabiliriz."

"Bu planın bir kusuru var!" dedi Scronkfinkle; ama ne dese fayda etmedi, zira sürü çoktan Pastus'un önerdiği doğrultuda çalışmalara başlamak için havalanmıştı.

Geride sadece iki üç serçe kalmıştı. Onlar da kafa kafaya verip baykuşların nasıl evcilleştirilebileceğini ya da ehlileştirilebileceğini düşünmeye giriştiler. Çok geçmeden Pastus'un haklı olduğunu fark ettiler: Özellikle de gerçek bir baykuşun yokluğunda bunu öğrenmeye çalışmak son derece zor bir işti. Yine de ellerinden geldiğince çabaladılar, bir baykuşu nasıl kontrol edebileceklerini bilmeden arkadaşlarının bir baykuşla çıkagelmelerinden duydukları korkuyu bir türlü atamadılar.

Hikâyenin nasıl sona erdiği bilinmiyor, ama elinizdeki bu kitap Scronkfinkle ve takipçilerine ithaf edilmiştir.

Önsöz

Kafatasınızın içinde okumayı gerçekleştiren şey yer alıyor. İnsan beyni denen bu şey, diğer hayvanlarda olmayan birtakım kabiliyetlere sahiptir. Gezegendeki hâkim konumumuzu bu ayırt edici kabiliyetlere borçluyuz. Diğer hayvanların güçlü kasları ve keskin pençeleri var, bizim ise daha akıllı beyinlerimiz. Genel zekâ bakımından sahip olduğumuz mütevazı üstünlük sayesinde dil, teknoloji ve karmaşık bir toplumsal örgütlenme geliştirmeyi başardık. Bu üstünlük, her kuşağın öncekilerin başarılarına yenilerini eklemesiyle zaman içinde daha da ilerledi.

Eğer bir gün genel zekâ bakımından insan beynini aşan makine beyinler yaratmayı başarırsak, o zaman bu yeni süper zekâ çok güçlü hale gelebilir. Ve nasıl ki şu an gorillerin kaderi gorillerden çok biz insanlara bağlıysa, bizim türümüzün kaderi de makine süper zekâsının eylemlerine bağlı olacaktır.

Fakat bizim de bir üstünlüğümüz yok değil: Biz bir şeyler yapmaya, yaratmaya alışmışız. İlkesel açıdan, insani değerleri koruyacak bir tür süper zekâ yaratabiliriz. Bunu yapmak için kesinlikle güçlü bir nedenimiz olacaktır. Pratikte, kontrol problemi (süper zekânın ne yapacağının nasıl kontrol edileceği) çok zor görünüyor. Ayrıca sadece tek bir şansımız olacak gibi duruyor. Eğer olur da dost canlısı olmayan bir süper zekâ yaratırsak, onu ya da tercihlerini değiştirmekten bizleri alıkoyacaktır. Kaderimiz çizilmiş olacaktır.

Ben bu kitapta süper zekâ olasılığının sunduğu zorluğu ve buna en iyi şekilde nasıl yanıt verebileceğimizi anlamaya çalışıyorum. İnsanlığın geçmişten bu yana karşılaştığı en büyük zorluk budur demek yanlış olmaz. Bu işi başaralım ya da başaramayalım, muhtemelen karşılaşacağımız son zorluk olacak.

Bu kitabın yapay zekâda büyük bir atılımın arifesinde olduğumuz ya da bu tür bir gelişmenin ne zaman gerçekleşeceğini kesin olarak tahmin edebileceğimiz gibi bir argümanı yoktur. Bunun 21. yüzyılda bir tarihte gerçekleşmesi muhtemel görünüyor, ama kesin tarihi bilmiyoruz. İlk iki bölümde mümkün olan yollar ele alınıyor ve zamanlama sorunu hakkında bir şeyler söyleniyor. Fakat kitabın ana gövdesini sonrasında neler olacağı oluşturuyor. Bir zekâ patlamasının kinetiğini, süper zekânın biçimlerini ve güçlerini ve de belirleyici bir üstünlüğe erişen süper zekâ taşıyıcısının önündeki stratejik tercihleri inceliyoruz. Ardından odak noktamızı kontrol problemine kaydırıp, yaşayabilir ve faydalı bir sonuç elde etmek için başlangıç koşullarını şekil-

lendirmek üzere neler yapabileceğimizi soruyoruz. Kitabın sonlarına doğru konudan biraz uzaklaşıp tetkiklerimizden çıkan büyük resme odaklanıyoruz. İleride varoluşsal bir felaketten kaçınma şansımızı artırmak için şimdi neler yapılması gerektiğine dair bazı tavsiyelerde bulunuluyor.

Bu kitabı yazmak hiç kolay olmadı. Açtığım yolda diğer araştırmacıların yeni sınırlara çok daha çabuk ve rahat ulaşacaklarını, böylelikle idrakimizin menzilini genişletecek yeni çalışmalara daha diri ve hazır katılacaklarını umuyorum. (Açılan yol biraz engebeli ve inişli çıkışlı olsa da, ortaya çıkan sonucu değerlendiren eleştirmenlerin göz koyulan alanın husumetini hafife almayacaklarını umuyorum!)

Bu kitabı yazmak hiç kolay olmadı. Okunması kolay bir kitap olsun diye uğraştım, ama çok başarılı olduğumu düşünmüyorum. Yazarken aklımdaki hedef kitleyi önceki ben neysem ona göre şekillendirdim ve kendim okuduğumda zevk alacağım bir kitap yazmaya çalıştım. Bu durum dar bir kitleye seslenme anlamına geliyor olabilir. Gelgelelim birazcık kendi düşüncelerini işin içine katan ve kendi kültürel ortamlarındaki en benzer klişelerle özdeşlik kurarak karşılaştığı her yeni fikri hemen yanlış anlama ayartmasına direnen herkesin kitabın içeriğine vakıf olabileceğini düşünüyorum. Teknik meselelerden uzak olan okurlar ara ara matematik işlemleri ya da uzmanlık gerektiren kelimeler görürlerse hevesleri kırılmasın, zira etrafındaki açıklamalardan esas anlatılmak isteneni çıkarmak her zaman mümkündür. (Diğer taraftan, işin özüne dair daha fazla şey bulmak isteyen okurlar için ise sonnotlarda epeyce malzeme var.[1])

Bu kitapta vurgulanan birçok husus muhtemelen yanlıştır.[2] Keza hayati öneme sahip olan ve dolayısıyla bazı ya da tüm yargılarımı tümden geçersiz kılabilecek birtakım meseleleri gözden kaçırmış olabileceğimi de reddetmiyorum. Metin boyunca belirsizlik derecelerini ve nüanslarını belirtmek için epey didindim: Sırf bu yüzden metinde "muhtemelen", "olabilir", "sanırım, "pekâlâ mümkündür", "görünen o ki", "herhalde", "büyük ihtimalle", "neredeyse kesindir" gibi tabirlerden geçilmiyor. Şerh düşülen her yerde dikkatli ve düşünülerek iş tutuldu. Ama bu epistemik tevazu gösterileri yeterli değildir; belirsizliğin ve yanılabilirliğin sistemli olarak kabul edilmesi de gerekir. Burada sahte bir tevazu söz konusu değildir, zira kitabımın cidden hatalı ve yanlış yönlendirici olabileceğine inanmakla birlikte, yazında sunulan alternatif görüşlerin de ciddi derecede yanlış olduğunu düşünüyorum ki, buna süper zekâ olasılığını şimdilik rahatlıkla ya da makul ölçüde göz ardı edebileceğimizi savunan farazi görüş, "sıfır hipotezi" de dahildir.

Teşekkür

Yazma sürecimi çevreleyen zar son derece geçirgendi. Bu kitaba çalışırken yaratılan birçok fikir ve kavramın dışarı sızıp daha geniş bir tartışmanın parçası olmasına izin verilmiştir ve elbette kitabın hazırlanma sürecinde dışarıdan gelen sayısız içgörü metne yedirilmiştir. İktibas aygıtında biraz özenli davranmaya çalıştım, ama katkı sunanların hepsini sıralamak mümkün değildi. Fikirlerimin netleşmesine yardımcı olan uzun tartışmalardan ötürü minnettarlığımı ifade etmem gereken birçok insan var: Sam Altman, Dario Amodei, Ross Andersen, Stuart Armstrong, Owen Cotton-Barratt, Nick Beckstead, Yoshua Bengio, David Chalmers, Paul Christiano, Milan Ćirković, Andrew Critch, Daniel Dennett, David Deutsch, Daniel Dewey, Thomas Dietterich, Eric Drexler, David Duvenaud, Peter Eckersley, Amnon Eden, Oren Etzioni, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Zoubin Ghahramani, Ian Goldin, Katja Grace, Roger Grosse, Tom Gunter, J. Storrs Hall, Robin Hanson, Demis Hassabis, Geoffrey Hinton, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Patrick LaVictoire, Shane Legg, Moshe Looks, Willam MacAskill, Eric Mandelbaum, Gary Marcus, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Seán Ó hÉigeartaigh, Christopher Olah, Toby Ord, Laurent Orseau, Michael Osborne, Larry Page, Dennis Pamlin, Derek Parfit, David Pearce, Huw Price, Guy Ravine, Martin Rees, Bill Roscoe, Francesca Rossi, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jürgen Schmidhuber, Bart Selman, Nicholas Shackel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Nate Soares, Dan Stoicescu, Mustafa Suleyman, Jaan Tallinn, Alexander Tamas, Jessica Taylor, Max Tegmark, Roman Yampolskiy ve Eliezer Yudkowsky.

Bilhassa ayrıntılı yorumları için Milan Ćirković, Daniel Dewey, Owain Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan ve Carl Shulman'a müteşekkirim. Farklı bölümlerle ilgili tavsiyeleri ya da araştırmadaki yardımlarından ötürü Stuart Armstrong, Daniel Dewey, Eric Drexler, Alexandre Erler, Rebecca Roache ve Anders Sandberg'a teşekkür etmek istiyorum.

Kitap taslağının hazırlanmasındaki yardımlarından ötürü Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, John King, Kristian Rönn, Susan Rogers, Kyle Scott, Andrew Snyder-Beattie, Cecilia Tilli ve Alex

Vermeer'e müteşekkirim. Proje boyunca bana birçok kez cesaret aşıladığı için editörüm Keith Mansfield'a bilhassa teşekkür etmek istiyorum.

Burada ismini unuttuğum herkesten özür diliyorum.

Son olarak, en büyük teşekkürü maddi destek sunan kişi ve kurumlara, dostlarıma ve aileme sakladım: Onların desteği olmasaydı bu kitap tamamlanamazdı.

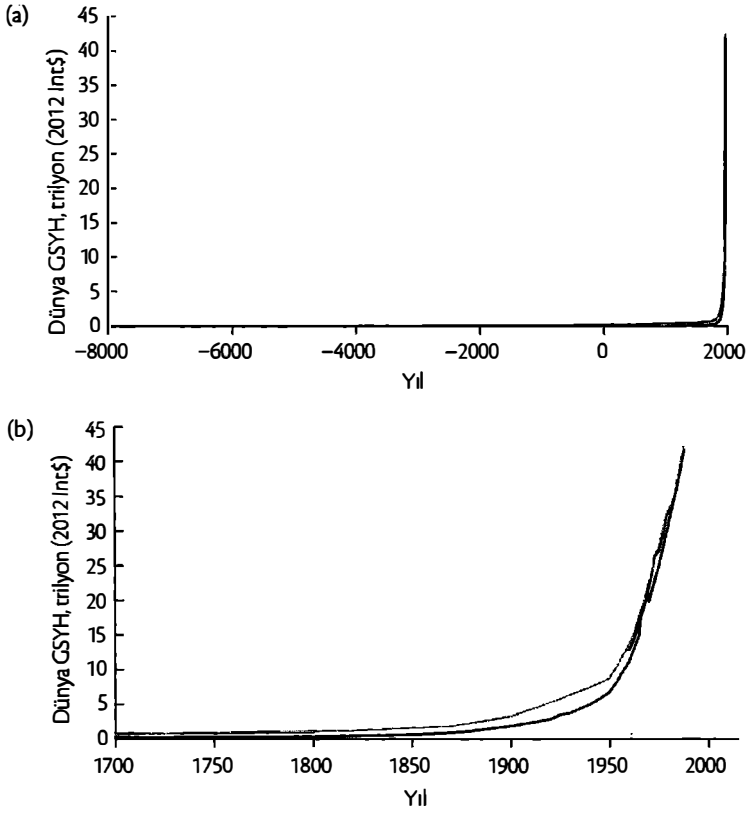
Geçmişteki Gelişmeler, Mevcut Yeterlilikler

Geriye dönüp bakarak başlayacağız. Tarih, en geniş ölçekte, belirgin bir büyüme silsilesi sergiler ve bu dizideki her bir tarz sanki öncekinden daha hızlı gerçekleşir. Bu örüntü, başka (ve daha hızlı) bir büyüme tarzının mümkün olabileceğinin önerilmesini sağlamıştır. Gelgelelim bu tespite fazla güvenmemek gerekir; bu kitap "teknolojik ivme", "üstel büyüme" ya da kimi zaman "tekillik" başlığı altında toplanan çeşitli kavramlarla alakalı bir kitap değildir. Ardından, yapay zekânın tarihini gözden geçireceğiz. Daha sonra, bu alanın mevcut yeterliliklerini ele alacağız. Son olarak, yakın dönemdeki uzman görüşlerine göz atıp gelecekteki ilerlemelerin zamansal çizelgesine dair cehaletimize kafa yoracağız.

Büyüme tarzları ve büyük tarih

Çok değil birkaç milyon yıl önce atalarımız hâlâ Afrika semaları altında daldan dala atlıyorlardı. Jeolojik, hatta evrimsel zaman ölçeğinde, büyük kuyruksuz maymunlarla son ortak atamızdan *Homo sapiens*'in doğuşu çok çabuk gerçekleşmiştir. Dik duruş, diğer dört parmaktan ayrı başparmaklar ve –hayati önemde bir gelişme olarak– beyin büyüklüğü ile nörolojik örgütlenmemizdeki bazı görece küçük değişiklikler, bilişsel yeteneğimizde büyük bir sıçramaya yol açtı. Böylece insanlar gezegendeki diğer bütün türleri geride bırakacak şekilde soyut düşünebilir, karmaşık düşünceleri iletişime konu edebilir ve kuşaktan kuşağa kültürel olarak bilgi biriktirebilir hale geldi.

Bu yetenekler insanların giderek daha verimli hale gelen üretici teknolojiler geliştirmelerine izin verdi ve atalarımızın yağmur ormanlarından ve savanalardan uzağa göç etmelerini mümkün hale getirdi. Özellikle de tarımsal üretime geçilmesinin ardından nüfus yoğunlukları ve toplam insan nüfusu arttı. Daha fazla insan daha fazla fikir anlamına geliyordu; daha büyük yoğunluklar fikirlerin daha kolay yayılabilmesi ve bazı bireylerin zamanlarını uzmanlaşmış yetenekler geliştirmeye vakfetmeleri anlamına geliyordu. Bu gelişmeler teknolojik kapasitemizin ve ekonomik üretkenliğin büyüme hızını artırdı. Sanayi Devrimi'yle bağlantılı sonraki gelişmeler büyüme hızında ikinci ve benzer bir değişime yol açacaktı.



ŞEKİL 1 Dünya GSYH'nın uzun vadeli tarihi. Doğrusal ölçekte hazırlanmış olan bu dünya ekonomisi tarihi x-eksenini kucaklayan bir düz çizgiye benziyor, ama sonra birdenbire hızlıca yukarı çıkıyor. (a) Son on bin yıla odaklandığımızda bile örüntü özünde tek bir 90 derecelik açı olarak kalıyor. (b) Ancak son 100 küsur yılda eğri sıfır düzeyinin üstüne belirgin şekilde çıkıyor. (Farklı çizgiler çok az farklı tahminler sunan farklı veri kümelerine tekabül etmektedir.*)

* Van Zanden (2003); Maddison (1999, 2001); De Long (1998).

Büyüme hızındaki bu tür değişikliklerin önemli sonuçları oldu. Birkaç bin yıl önce, insanın ya da insansıların tarihöncesi döneminin başlarında, büyüme o kadar yavaştı ki insanın üretim kapasitesinin fazladan bir milyon insanı geçim düzeyinde yaşatabilecek kadar artması bir milyon yıl almıştı. Bugün, Sanayi Devrimi'nin ardından dünya ekonomisi ortalama doksan dakikada bir bu düzeyde büyüme oranı yakalıyor.[1]

Mevcut büyüme hızı bile görece uzun bir süre devam ettirilirse etkileyici sonuçlar verecektir. Eğer dünya ekonomisi son elli yıldaki hızıyla büyümeye devam ederse, dünya 2050 yılında 4,8 kat, 2100 yılındaysa 34 kat daha zengin hale gelecektir.[1]

Fakat büyüme hızındaki büyüklük bakımından Tarım Devrimi ve Sanayi Devrimi'ndekine benzer bir değişim yanında, kesintisiz bir üstel büyümeyi sürdürme olasılığı solda sıfır kalır. İktisatçı Robin Hanson'un tarihsel ekonomi ve nüfus verilerine

dayandırdığı tahminlerine göre, dünya ekonomisinin ikiye katlanma zamanı Buzul Çağı'nda avcı-toplayıcı topluluklar için 224 bin yıl; tarım toplumları için 909 yıl; sanayi toplumu için ise 6,3 yıldır.[2] (Hanson'un modeline göre, mevcut çağ tarımsal ve sınai büyüme tarzlarının bir karışımıdır; dünya ekonomisinin bütünü henüz 6,3 yılda bir ikiye katlanacak şekilde büyümemektedir.) Eğer farklı bir büyüme tarzına bu türde bir geçiş daha yaşanırsa ve önceki ikisiyle benzer büyüklükte olursa, dünya ekonomisinin yaklaşık iki haftada bir ikiye katlanacağı yeni bir büyüme rejimi ortaya çıkabilir.

Bu tür bir büyüme hızı mevcut durumda akıl almaz görünüyor. Eski çağlardaki gözlemciler bir gün dünya ekonomisinin tek bir insanın ömrü süresince birkaç kat birden büyüyebileceğini hayal bile edemezlerdi. Oysa şu an olağan gördüğümüz olağanüstü koşul budur.

Teknolojik tekilliğin yaklaştığı fikri Vernor Vinge'in çığır açıcı denemesiyle başlayıp Ray Kurzweil ve diğerlerinin yazılarıyla devam ederek epey popüler hale gelmiştir.[3]

Fakat "tekillik" terimi kafa karıştıracak şekilde birçok farklı anlamda kullanılmış ve teknolojik-ütopyacı yan anlamlarla yüklü olan çok habis (handiyse binyılcı) bir aura edinmiştir.[4] Bu anlamların ve yan anlamların çoğu argümanımız açısından önemsiz olduğu için, "tekillik" terimini bir kenara bırakıp daha kesin bir terminolojiden istifade edebiliriz.

Tekillikle ilintili olarak bizi burada ilgilendiren fikir bir zekâ patlaması olanağı ve özellikle de makine süper zekâsı olasılığıdır. **ŞEKİL 1**'deki gibi büyüme grafiklerine bakıp, büyüme tarzında Sanayi ya da Tarım Devrimi'ndekine benzer bir devasa değişimin ufukta göründüğüne kanaat getirenler olabilir. Bu kişiler, akabinde, dünya ekonomisinin iki kat büyüme zamanının birkaç hafta kadar kısa bir süreye düşmesinin eşiğine olduğumuz biyolojik türden çok daha hızlı ve daha verimli olan zihinlerin yaratılmasını içermeyecek şekilde gerçekleştiği bir senaryo tasavvur etmenin zor olduğunu düşünebilirler. Oysa bir makine zekâsı devrimi olasılığını ciddiye almak için eğri uydurma egzersizlerine ya da geçmişteki ekonomik büyümelerden çıkarımlar yapmaya ihtiyacımız yok. Göreceğimiz üzere, bunu dikkate almamız için güçlü gerekçeler var.

Büyük umutlar

İnsanla aynı genel zekâya sahip olan makineler, yani aklıselim sahibi olan ve öğrenme, muhakemede bulunma ve gerek doğal gerek soyut olsun çok çeşitli alanlarda karmaşık bilgi işlem süreçleriyle baş edecek şekilde plan yapma yetilerini barındıran zekâya sahip olan makineler 1940'larda bilgisayarın icadından bu yana umutla bekleniyor.[5] O dönemde beklenti, bu tür makinelerin icat edilmesinin genellikle yirmi yıl alacağı yönündeydi. O zamandan beri beklenen varış tarihi yılda bir yıllık hızla ileri atılıyor; böylece bugün yapay genel zekâ imkânıyla ilgilenen fütüristler hâlâ çoğu zaman akıllı makinelerin yirmi yıl kadar uzaklıkta olduğuna inanıyorlar.[6]

Yirmi yıl, radikal değişim kehanetinde bulunanlar için bir tatlı noktadır: Dikkat çekecek ve ilintili olacak kadar yakın, ama bugün ancak muğlak şekilde hayal edilebilecek birtakım atılımların bu süre zarfında gerçekleşmiş olacağını varsaymayı mümkün kılacak kadar uzak. Bunu kısa zaman ölçekleriyle kıyaslayalım: Bugünden beş on yıl sonra dünyaya büyük bir etkide bulunacak olan çoğu teknoloji halihazırda sınırlı bir kullanımdadır, oysa dünyayı on beş yıla kalmadan yeniden şekillendirecek teknolojiler muhtemelen laboratuvar prototipleri olarak mevcuttur. Ayrıca yirmi yıl cüretkâr bir tahminin sınırındadır ve tahmincinin kariyerini tehdit etmez.

Bazı bireylerin geçmişte yapay zekâyla ilgili aşırı tahminlerde bulunmuş olmaları, yapay zekânın olanaksız olduğu ya da asla gelişmeyeceği anlamına gelmez.[7] İlerlemenin beklenenden daha yavaş gerçekleşmiş olmasının esas nedeni, zeki makineler yaratmanın bu alandaki öncülerin öngördüğünden daha büyük teknik zorluklar barındırdığının anlaşılmış olmasıdır. Fakat yine de bu zorlukların ne kadar büyük olduğu ve bunları aşmaktan ne kadar uzak olduğumuz sorusu hâlâ tartışmaya açıktır. Kimi zaman ilk başta içinden çıkılmayacak kadar karmaşık görünen bir problemin şaşırtıcı derecede basit bir çözümü vardır; gerçi tam tersi muhtemelen daha sık görülür ama neyse!

Bir sonraki bölümde, insan düzeyinde makine zekâsına yol açabilecek farklı yollara bakacağız. Ama burası ile insan düzeyinde makine zekâsı arasında ne kadar çok durak olursa olsun, makine zekâsının son durak olmadığını en baştan belirtelim. Çok daha kısa süre sonra varacağımız bir sonraki durak insanüstü düzeyde makine zekâsıdır. Tren Beşerşaşar Durağı'nda durmayabilir, hatta hız bile kesmeyebilir. Sağa direksiyon kırması muhtemeldir.

Alan Turing'in İkinci Dünya Savaşı'ndaki şifre kırma ekibinde başistatistikçi olarak çalışmış olan Matematikçi I.J. Good, bu senaryonun asli veçhelerini dile getiren ilk isim olabilir. 1965 yılına ait, çok alıntılanan bir paragrafında şöyle yazar:

> Bir aşırı zeki makineyi, ne kadar akıllı olursa olsun her insanın entelektüel faaliyetlerini misliyle aşabilecek bir makine olarak tanımlayalım. Makinelerin tasarımı bu entelektüel faaliyetlerin bir tasarımı olduğundan, bir aşırı zeki makine daha da iyi makineler tasarlayabilir; bir "zekâ patlaması" yaşanacağına şüphe yok ve insanın zekâsı da çok geride kalacaktır. Dolayısıyla ilk süper zekâ makine insanın yaratma gereksinimi duyacağı son icat olacaktır, yeter ki makine bize onu nasıl kontrol altında tutacağımızı söyleyecek kadar yumuşak başlı olsun.[8]

Bu tür bir zekâ patlamasının temel varoluşsal riskler barındırdığı ve dolayısıyla meydana gelme olasılığının mütevazı derecede küçük olduğu bilinse bile (ki böyle bir şey yok) azami ciddiyetle incelenmesi gerektiği bugün bariz görünebilir. Gelgelelim yapay zekânın öncüleri insan düzeyinde yapay zekânın yakın olduğuna inanmalarına karşın, çoğunlukla, insandan daha ileri bir yapay zekânın mümkün olduğuna inanmıyorlardı.

Adeta spekülasyon kasları radikal bir olanak olarak insan zekâsına erişen makineleri tasavvur etme konusunda kendisini o kadar yormuştu ki bunun doğal sonucunu kavrayamıyorlardı: Makineler er geç süper zeki hale geleceklerdi.

Yapay zekâ öncüleri, girişikleri işin risklerinin olabileceğini çoğunlukla anlamamışlardı.[9] Yapay zekâların ve potansiyel bilgisayar efendilerinin yaratılmasıyla ilgili bir güvenlik kaygısı ya da etik şüpheye ciddi ciddi kafa yormak şöyle dursun, bu konuya yalandan ilgi bile göstermiyorlardı: Bu çağın o kadar da etkileyici olmayan eleştirel teknoloji değerlendirmesinin arka planı düşünüldüğünde daha da şaşırtıcı bir boşluktur bu.[10] Bu girişim nihayetinde gerçekten uygulanabilir hale geldiğinde, sadece bir zekâ patlamasının önüne geçecek teknolojik yeterliliğe sahip olmakla kalmayıp, infilakı yaşayabilir kılmak için gerekli olabilecek yüksek uzmanlık seviyesine de ulaşmış olacağız.

Fakat ileride bizi bekleyen meselelere dönmeden önce, makine zekânın günümüze kadarki tarihine hızlıca bakmakta yarar var.

Umut ve umutsuzluk mevsimleri

1956 yazında Dartmouth College'ta nöral ağlar, otomat teorisi ve zekâ çalışmalarına ilgi duyan on bilimci altı hafta sürecek bir atölye için bir araya geldi. Bu Dartmouth Yaz Projesi çoğu zaman bir araştırma alanı olarak yapay zekâ horozunun ötüşü olarak kabul edilir. Katılımcıların çoğu ileride kurucu simalar olarak kabul edilecekti. Katılımcıların iyimser bakış açısı toplantının sponsorluğunu yapan Rockefeller Vakfı'na sunulan teklifte yansımasını bulmuştu:

> 2 aylık, 10 kişilik bir yapay zekâ çalışması yürütülmesini teklif ediyoruz [...] Çalışma, öğrenmenin her bir veçhesinin ya da zekânın herhangi bir özelliğinin bir makinenin simüle edebileceği şekilde kesinkes tarif edilebileceğini ilkesel olarak kabullenen bir hipotez temelinde ilerleyecektir. Dil kullanan, soyutlamalar yapan, kavramlar yaratan, bugün insanlara ayrılmış problemleri çözen ve kendini geliştiren makineler yaratmanın nasıl mümkün olduğu araştırılacak. Eğer dikkatle seçilmiş bir grup bilimci bir yaz boyu birlikte çalışırsa, bu problemlerden birinde ya da birkaçında ciddi bir ilerleme kaydedilebileceği kanaatindeyiz.

Bu cüretkâr başlangıçtan sonra geçen altmış yılda yapay zekâ alanı kâh heyecanın zirve yaptığı büyük umutlar beslenen kâh aksiliklerin hayal kırıklıklarını beslediği dönemlerden geçti.

Dartmouth toplantısıyla başlayan ilk heyecan dalgasının hissedildiği dönem, sonradan (toplantının esas örgütleyicisi olan) John McCarthy tarafından, "Bak, anne, ellerimi kullanmıyorum!" çağı olarak adlandırılmıştı. İlk zamanlarda araştırmacılar "Hiçbir makine falanca şeyi yapamaz!" şeklindeki iddiaları çürütmek üzere tasarlanmış sistemler yaratıyorlardı. Bu tür şüpheci iddialar o dönemde yaygındı. Bunlara

karşılık, yapay zekâ araştırmacıları, "mikrodünya"da (performansın azaltılmış bir versiyonunun sergilenmesini sağlayan, tanımı net yapılmış, sınırlı bir alanda) *X*'i başaran küçük sistemler yarattılar, böylece kavramın bir kanıtını sunup *X*'in ilkesel olarak makineyle yapılabileceğini gösterdiler. Erken dönemdeki bu tür sistemlerden biri olan "Mantık Teorisyeni", Whitehead ve Russell'ın *Principia Mathematica*'sının ikinci bölümündeki teoremlerden çoğunu kanıtlayabiliyor, hatta özgün halinden çok daha zarif bir kanıt sunabiliyordu, böylece makinelerin "ancak sayısal olarak düşünebildikleri" iddiasını boşa çıkararak makinelerin çıkarım da yapabildiğini ve mantıksal kanıtlar icat edebildiğini gösteriyordu.[11] Genel Problem Çözücü adındaki bir takip programı, biçimsel olarak belirtilmiş çok sayıda problemi ilkesel olarak çözebiliyordu.[12] Üniversite birinci sınıfta karşımıza çıkan yüksek matematik problemlerini, bazı IQ testlerinde karşımıza çıkan türde görsel analoji problemlerini ve basit sözlü cebir problemlerini çözebilen programlar da yazılmıştı.[13] Sallanan Robot (çalışırken sallandığı için bu ad verilmişti) mantıksal muhakemenin algıyla bütünleştirilebileceğini ve fiziksel faaliyeti planlayıp kontrol etmek için kullanılabileceğini gösteriyordu.[14] ELIZA programı bir bilgisayarın nasıl Rogerci bir psikoterapisti taklit edebileceğini ortaya koyuyordu.[15] 1970'lerin ortalarında SHRDLU programı simülatif bir geometrik bloklar dünyasındaki simüle edilmiş bir robot kolunun bir kullanıcı tarafından İngilizce yazılan soruları cevaplayıp talimatları takip edebildiğini gösteriyordu.[16] Sonraki yıllarda makinelerin çeşitli klasik besteciler gibi beste yapabildiği, belli klinik tanılarında asistan doktorlardan daha iyi iş çıkarabildiği, kendi başlarına araba kullanabildiği ve patentli icatlar yapabildiği kanıtlandı.[17] Hatta espri yapan bir yapay zekâ makinesi bile yapıldı.[18] Gerçi mizah seviyesi çok yüksek değildi (*–Ayda* 100 bin dolar kazanmak ister misin? –Evet! *–Ay'da*?!), ama geri bildirimlerde çocukların bu kelime oyunlarını eğlenceli bulduğu söyleniyordu.

İlk kanıtlama sistemlerinde başarı elde eden yöntemleri daha geniş çeşitlilikteki ya da daha çetin problemlere uygulamak genellikle zor oldu. Bunun nedenlerinden biri, kapsamlı araştırma gibi bir şeye yaslanan yöntemlerle araştırılması gereken olanakların "kombinasyonel patlaması"dır. Bu tür yöntemler bir problemin basit sorularında iyi iş çıkarabilir, ama işler biraz karmaşıklaştığında çuvallar. Sözgelimi tek bir çıkarsama kurallı ve 5 aksiyomlu bir tümdengelim sistemindeki 5 satır uzunluğunda bir kanıtı olan teoremi kanıtlamak için, mümkün olan 3125 kombinasyonu sıralayıp her birinin istenen sonucu verip vermediğini kontrol etmek yeterlidir. Kapsamlı bir araştırma 6-7 satırlık kanıtlarda da işe yarayacaktır. Ama görev zorlaştıkça kapsamlı araştırma yöntemi de sıkıntıya düşer. 50 satırlık kanıtı olan bir teoremi kanıtlamak 5 satırlık kanıtı olan bir teoremi kanıtlamaktan on kat daha uzun sürmez; daha ziyade, eğer kapsamlı araştırmadan yararlanılırsa, mümkün olan $550 \approx 8{,}9 \times 10^{34}$ dizilerini didik didik etmek gerekir ki, bunu yapmak en hızlı süper bilgisayarlarla bile imkânsızdır.

Kombinasyonel patlamayı aşmak için bulgulayıcı araştırmadan, planlamadan ve esnek soyut temsillerden (yani ilk yapay zekâ sistemlerinde pek gelişmemiş olan yeteneklerden) yararlanarak hedef alandaki yapıyı sömüren ve önceki bilgiden istifade eden algoritmalara ihtiyaç vardır. Bu ilk sistemlerin performansı belirsizliği ele alan yöntemlerinin zayıflığı, kırılgan ve temelsiz sembolik temsillere yaslanması, veri kısırlığı ve hafıza kapasitesi ile işlemci hızındaki ciddi donanım sınırlılığından ötürü de sıkıntı yaşamıştı. 1970'lerin ortalarına gelindiğinde bu sorunlara dair farkındalık giderek artıyordu. Birçok yapay zekâ projesinin ilk baştaki vaatleri asla gerçekleştiremeyeceğine dair farkındalık ilk "Yapay Zekâ Kışı"nın başlamasına yol açtı: Fonların azaldığı ve şüpheciliğin arttığı bir geri çekilme dönemi geldi ve yapay zekâ gözden düştü.

1980'lerin başında Japonya'nın Beşinci Nesil Bilgisayar Sistemleri Projesi'ni başlatmasıyla birlikte yapay zekâ da ikinci baharını yaşamaya başladı. Japonların bu projesi yapay zekâ için bir platform işlevi görecek olan muazzam bir paralel bilgisayar mimarisi geliştirerek son teknolojiyi aşmayı amaçlayan bir devlet-özel teşebbüs ortaklığıydı. Bu proje Japonların "savaş sonrası ekonomik mucizesi"ne duyulan hayranlığın tepe noktasında olduğu döneme denk geldi; o sıralar, Batılı hükümetler ve iş dünyasının önde gelen isimleri bu sihri kendi ülkelerinde tekrarlamak amacıyla Japonya'nın ekonomik başarısının gerisindeki formülü bulma telaşına düşmüşlerdi. Japonya yapay zekâya büyük yatırım yapma kararı alınca, diğer birkaç ülke de aynı yolu izledi.

Sonraki yıllar "uzman sistemleri"nde müthiş bir artışa tanık oldu. Karar alıcılar için destek araçları olarak tasarlanan uzman sistemleri, olguların bilgi temelinden basit çıkarımlar yapan kurala dayalı programlardı ve beşeri bilimler uzmanlarından alınarak, bin bir zorlukla formel dilde kodlanmıştı. Bu uzman sistemlerinden yüzlerce yaratıldı. Gelgelelim küçük sistemler çok az getiri sağladı, büyük olanlarınsa geliştirilmesi, doğrulanması ve güncellenmesi pahalıya geldi ve genel olarak kullanması zor, hantalcaydı. Sırf tek bir programı işletmek uğruna bağımsız bir bilgisayar edinmek hiç de pratik değildi. 1980'lerin sonunda bu büyüme mevsimi de miadını doldurmuştu.

Beşinci Nesil Projesi tıpkı ABD ve Avrupa'daki muadilleri gibi amaçlarına ulaşmakta başarısız oldu. İkinci bir Yapay Zekâ Kışı çöktü. Bu noktada, bir eleştirmen gayet haklı bir serzenişte bulunabiliyordu: "Yapay zekâ araştırmalarının bu zamana kadarki tarihi her zaman belli alanlarda çok sınırlı bir başarı ve hemen akabinde, bu ilk başarıların ilk başta heveslendirdiği büyük hedeflere ulaşma başarısızlığından ibarettir."[19] Özel yatırımcılar "yapay zekâ" markasını taşıyan her türlü girişimden uzak durmaya başladılar. "Yapay zekâ", akademisyenler ile maddi destekçileri arasında bile istenmeyen bir etiket haline geldi.[20]

Gelgelelim teknik çalışma hız kesmeden devam etti ve 1990'lara gelindiğinde ikinci Yapay Zekâ Kışı yavaş yavaş bahara dönmeye başladı. İyimserlik yeni tekniklerin devreye sokulmasıyla yeniden parladı. Bu teknikler, yüksek düzey sembol manipülasyonuna

odaklanmış ve tepe noktasına 1980'lerin uzman sistemleriyle ulaşmış olan geleneksel mantıkçı paradigmaya (çoğunlukla "Güzel Eski Moda Yapay Zekâ", kısaca "GOFAI" olarak biliniyordu) alternatifler sunuyor görünüyordu. Yeni popülerlik kazanan ve aralarında nöral ağlar ile genetik algoritmaların da olduğu teknikler GOFAI yaklaşımının kusurlarından bazılarını, özellikle de klasik yapay zekâ programlarını nitelendiren "kırılganlığı" aşmayı vaat ediyordu (programcılar tek bir tane bile olsa azıcık hatalı bir varsayımda bulunduklarında ortaya tam bir saçmalık çıkıyordu). Yeni teknikler daha organik bir performansa sahip olmakla övünüyordu. Örneğin nöral ağlar "dereceli bozulma" özelliği sergiliyordu: bir nöral ağa verilen ufak bir hasar bütünsel bir çöküş yerine performansında ufak bir kötüleşmeye yol açıyordu. Daha da önemlisi, nöral ağlar deneyimlerden dersler çıkarabiliyor, örneklerden hareketle genelleme yapmanın doğal yollarını bulup girdilerindeki saklı istatistiki örüntüleri keşfedebiliyordu.[21] Bu sayede ağlar örüntü tanıma ve sınıflandırma problemlerinde iyi hale gelmişti. Örneğin bir nöral ağı deniz radarı sinyalleriyle ilgili veriler hakkında eğiterek ona denizaltıların, mayınların ve deniz yaşamının akustik profilini insan uzmanlardan daha kesin şekilde ayırt etmeyi öğretmek ve kategorilerin nasıl tanımlanması ya da farklı özelliklerin nasıl tartılması gerektiğini peşinen belirlemek zorunda kalmadan bunu yapmak mümkündü.

Basit nöral ağ modelleri 1950'lerin sonundan bu yana bilinmekle birlikte, bu alanda geri yayılma algoritmasıyla bir rönesans yaşanmaya başladı, zira bu sayede çok katmanlı nöral ağları geliştirmek mümkün hale geldi.[22] Girdi katmanı ile çıktı katmanı arasında bir ya da daha fazla ara ("saklı") nöron katmanına sahip olan bu tür çok katmanlı ağlar basit öncellerinden çok daha fazla çeşitlilikte işlev öğrenebilirler.[23] Giderek güçlü ve ulaşılabilir hale gelen bilgisayarlarla birleştirildiğinde, algoritmadaki bu ilerlemeler mühendislere birçok uygulamada, pratikte yararlı olabilecek kadar iyi olan nöral ağları yaratma imkânı sağladı.

Nöral ağların beyni andıran nitelikleri geleneksel kurala dayalı GOFAI sistemlerinin aşırı mantıksal argümanlarla dolu ama zayıf olan performansının aksine gayet olumluydu, hem de muazzam sayıda paralel alt-sembolik işlemcinin önemini vurgulayan yeni bir yaklaşıma ilham verdi: *Bağlantıcılık*. O günden bugüne yapay zekâ ağlarıyla ilgili 150 binden fazla akademik makale yazılmıştır ve bu ağlar makine öğreniminde önemli bir yaklaşım olmayı sürdürmektedir.

Evrime dayalı yöntemler (genetik algoritmalar ve genetik programlama gibi), ortaya çıkar çıkmaz ikinci Yapay Zekâ Kışı'nın sona ermesine yardımcı olmuş bir başka yaklaşım teşkil eder. Belki akademik etkisi nöral ağlar kadar olmamıştır, ama epey popüler hale geldiği kesindir. Evrim modellerinde bir aday çözüm popülasyonu (veri yapıları ya da programları olabilir bu) sürdürülür ve yeni aday çözümleri mevcut popülasyondaki değişkenleri değiştirerek ya da yeniden birleştirerek rastgele yaratılır. Popülasyon bir sonraki kuşağa sadece en iyi adayların kalmasına imkân sağlayan bir seçim kıstası

(bir uygunluk işlevi) uygulayarak dönemsel olarak azaltılır. Binlerce kuşak boyunca tekrarlandığında, aday havuzundaki çözümlerin ortalama niteliği tedricen artar. Bu tür bir algoritma işlediğinde, çok çeşitli problemlere verimli çözümler üretebilir: Çarpıcı derecede yeni ve sezgilere aykırı olabilecek, çoğu zaman bir insan mühendisin tasarlayabileceklerinden daha doğal yapılar olarak görünen çözümlerdir bunlar. Ayrıca bu durum, ilkesel açıdan, uygunluk işlevinin ilk başta somutlaştırılmasını saymazsak (ki bunu yapmak çoğu zaman çok basittir) insan girdisine çok ihtiyaç duymadan gerçekleşebilir. Fakat pratikte, evrimsel yöntemlerin iyi işlemesini sağlamak yetenek ve deha gerektirir, özellikle de iyi bir temsil formatı yaratılmışken. Aday çözümleri kodlamak için etkili bir yol (aday alandaki gizli yapıyla eşleşen bir genetik dil) olmadan, evrimsel araştırma çok geniş bir araştırma alanında amaçsızca dolanmaya başlar ya da bir lokal optimumda kalakalır. İyi bir temsili format bulunsa bile, evrim sayısal olarak talepkârdır ve çoğu zaman kombinasyonel patlama tarafından dize getirilir.

Nöral ağlar ve genetik algoritmalar 1990'ların sonunda tıkanma noktasına gelen GOFAI paradigmasına alternatif sunuyormuş gibi görünerek heyecan yaratan yöntemlere birer örnektir. Ama burada niyetimiz bu iki yönteme methiyeler düzmek ya da bunları makine öğrenimindeki diğer birçok tekniğin üzerine çıkarmak değildir. Aslında son yirmi yılın en önemli teorik gelişmelerinden biri, yüzeysel açıdan farklı olan tekniklerin ortak bir matematiksel çerçeve dahilinde pekâlâ özel örnekler olarak anlaşılabileceğinin giderek daha da farkına varılmasıdır. Örneğin yapay nöral ağ türlerinin birçoğu belli türde istatistiksel hesaplama (azami ihtimal hesaplaması) yapan sınıflandırıcılar olarak görülebilir.[24] Bu perspektif nöral ağların örneklerden hareketle sınıflandırıcıları öğrenmek için daha büyük bir algoritma sınıfıyla karşılaştırılmasına imkân sağlar: "Karar ağaçları", "lojistik gerileme modelleri", "destek vektör makineleri", "naif Bayes", "*k*-en yakın komşu gerilemesi" bunlardan birkaçıdır.[25] Benzer bir tarzda, genetik algoritmalar stokastik yükselme icra ediyor olarak görülebilir ki, zaten bu da daha geniş bir optimizasyon algoritması sınıfının altkümesidir. Bu sınıflandırma yaratma ya da bir çözüm uzamı araştırma algoritmalarının her birinin matematiksel olarak incelenebilecek, kendine ait güçlü ve zayıf yanları vardır. Algoritmalar işlemci zamanı ve bellek alanı gereklilikleri, hangi önyargıları varsaydıkları, dışsal olarak üretilmiş içeriği hangi kolaylıkla özümseyebildikleri ve iç işleyişlerinin bir insan analist için ne kadar şeffaf olduğu gibi yönlerden farklılıklar gösterirler.

Dolayısıyla makine öğrenimi ve yaratıcı problem çözme konusundaki tüm bu dağdağanın gerisinde matematiksel açıdan gayet iyi somutlaştırılmış olan ödünleşimler yatmaktadır. İdeal olan kusursuz Bayesçi aracınınkidir, yani mevcut bilgiden olasılıksal açıdan optimum düzeyde yararlanmaktır. Bu ideale erişmek zordur, çünkü fiziksel bir bilgisayarda uygulanması hesaplama açısından çok zordur (bkz. **KUTU 1**). Dolayısıyla yapay zekâyı kısayollar bulma arayışı olarak görebiliriz: Belli bir optimalliği ya da

genelliği esas ilgi alanlarında yüksek performans elde edecek kadar muhafaza ederek feda eden Bayes idealine işlenebilir tarzda yaklaşma yollarıdır.

Bu resmin bir yansımasını son yirmi otuz yılda Bayesçi ağlar gibi olasılıkçı grafik modellerinde çıkarılan işlerde görebiliriz. Bayesçi ağlar belli birtakım alanlarda geçerli olan olasılıkçı ve koşullu bağımsız ilişkileri yansıtmanın kesin bir yolunu sunarlar.

(Bu tür bağımsız ilişkilerden istifade etmek mantıksal tümdengelim için olduğu gibi olasılıkçı çıkarsama için de büyük bir problem olan kombinasyon patlamasının üstesinden gelmek için aslidir.) Ayrıca bunlar nedensellik kavramına dair önemli bir içgörü sağlar.[26]

Özgül alanlardan öğrenme problemlerini genel bir problem olarak Bayesçi çıkarıma bağlamanın üstün yanlarından biri, Bayesçi çıkarımı daha verimli hale getiren yeni algoritmaların bu durumda birçok farklı alanda derhal iyileştirme sağlayacak olmasıdır. Örneğin Monte Carlo kestirim tekniklerindeki ilerlemeler doğrudan bilgisayar görüntüsünde, robot biliminde ve hesaplamalı genetikte uygulanır. Sağladığı bir diğer üstünlük ise farklı disiplinlerden araştırmacılara bulgularını daha kolay şekilde bir havuzda toplama olanağı sağlamasıdır. Grafik modelleri ve Bayesçi istatistik birçok alanda ortak bir araştırma odağı haline gelmiştir ki, bu alanlara makine öğrenimi, istatistiksel fizik, biyoinformatik, kombinasyonel optimizasyon ve iletişim teorisi de dahildir.[27] Makine öğreniminde son dönemdeki ilerlemenin önemli bir kısmı esasen diğer akademik alanlardan çıkarılan formel sonuçların özümsenmesiyle gerçekleşmiştir. (Makine öğrenimi uygulamaları daha hızlı bilgisayarlardan ve daha büyük veri kümelerinin giderek daha fazla mevcut olmasından da muazzam yarar elde etmiştir.)

En gelişkin teknoloji

Yapay zekâ halihazırda insan zekâsını birçok alanda geride bırakabiliyor. **TABLO 1**'de oyun bilgisayarlarının durumu ele alınıyor ve yapay zekâların şu anda birçok oyunda insan şampiyonları dize getirdiği gösteriliyor.[28]

Bu başarılar bugün etkileyici görünmeyebilir. Ama bunun nedeni, neyin etkileyici olduğuna dair standartlarımızın kaydedilen ilerlemelere uyum sağlamayı sürdürmesidir. Örneğin uzman satranç oyunculuğu bir zamanlar insan zekâsının yoğunlaşmış hali olarak görülürdü. 1950'lerin sonunda birkaç uzmanın görüşü şöyleydi: "Eğer başarılı bir satranç makinesi geliştirilebilirse, insan zekâsının çabalarının özüne nüfuz edilmiş olacağı söylenebilir."[29]Artık böyle görünmüyor. "İşe yaradığı andan itibaren artık kimse ona yapay zekâ adını vermiyor!" diye hayıflanan John McCarthy'nin duygularımıza tercüman olduğu söylenebilir.[30]

Gelgelelim satranç oyunculuğunun birçok kişinin tahayyül ettiğinden daha küçük bir zafer olmasının önemli bir anlamı yok değil. Bir zamanlar, belki de çok da

KUTU 1: Optimum bir Bayesçi aracı

İdeal bir Bayesçi aracı, "önsel olasılık dağılımıyla" yol çıkar: "Olanaklı dünyalar"ın her birine (yani dünyanın bürünebileceği azami ölçüde özgül tarzlardan her birine) olasılıklar tahsis eden bir işlev.* Bu önsel dağılım daha basit olanaklı dünyalara yüksel olasılıklar tahsis edecek şekilde tümevarımcı bir eğilimi barındırır. (Olanaklı bir dünyanın yalınlığını formel olarak tanımlamanın yollarından biri "Kolmogorov karmaşıklığı"nın terimleriyle, yani dünyanın tam bir tarifini yaratan en kısa bilgisayar programının uzunluğuna dayalı bir ölçümle olur.†) Önceki dağılım aynı zamanda programcıların aracıya kazandırmak istedikleri her türlü arka plan bilgisini de kapsar.

Aracı, algılayıcılarından yeni bilgi aldıkça, olasılık dağılımını Bayes'in teoremine göre yeni bilgiye koşullandırarak olasılık dağılımını günceller.‡ Koşullandırma alınan bilgiyle tutarsız çıkan bu dünyaların yeni olasılığını sıfır olarak belirleyip, olasılık dağılımını geri kalan olanaklı dünyalarda yeniden normalleştiren matematik işlemidir. Sonuç, bir "sonraki olasılık dağılımı"dır (aracı bir sonraki adımda bunu yeni önceki dağılımı olarak kullanabilir). Aracı gözlemler yaptıkça olasılık kütlesi de kanıtlarla tutarlılığını koruyabilen sayıları azalmış olanaklı dünyalara odaklanmaya başlar ve bu olanaklı dünyalar arasında basit olanlar her zaman daha fazla olasılığa sahiptir.

Mecazi olarak, bir olasılığı büyük bir kâğıt parçasının üstündeki kum olarak düşünebiliriz. Kâğıt, her biri olanaklı dünyalardan birine tekabül eden çeşitli büyüklükte alanlara ayrılır ve büyük alanlar daha basit olanaklı dünyalara tekabül eder. Keza eşit kalınlıkta bir kum katmanının tüm kâğıda yayıldığını tahayyül edin: Bu bizim önceki olasılık dağılımımızdır. Ne zaman olanaklı dünyalardan bazılarını devre dışı bırakan bir gözlem yapılsa, kumu kâğıdın mütekabil alanlarından temizleyip geride kalan alanlara eşit şekilde dağıtırız. Böylece kâğıttaki toplam kum miktarı asla değişmez, sadece gözlemsel kanıtlar biriktikçe daha az alanda yoğunlaşır. Bu,

---

* Açıklamayı gereksiz bilgilerle doldurmamak adına çeşitli teknik ayrıntıları bir kenara bırakıyoruz. Bu göz ardı edilen meselelerden bazılarını 12. Bölüm'de yeniden ele alma fırsatı bulacağız.

† Eğer (tikel bir) evrensel Turing makinesinde (U) işletilen bir p programı x sonucunu veriyorsa, o program x dizisinin bir tasviridir; biz bunu $U(p) = x$ olarak yazıyoruz. (Burada x dizisi olanaklı bir dünyayı temsil eder.) Dolayısıyla x'in Kolmogorov karmaşıklığı: $K(x) := \min_p \{\ell(p) : U(p) = x\}$ olur. Burada $\ell(p)$ bit olarak p'nin uzunluğudur. Dolayısıyla x'in "Solomonoff" olasılığı şöyle tanımlanır: $M(x) := \sum_{p:U(p)=x} 2^{-\ell}$.

Burada toplam, U'nun x ile başlayan bir diziyi çıkardığı tüm ("asgari", yani mutlaka durdurucu olmayan) programların üzerinde tanımlanır (Hutter 2005).

‡ E kanıtıyla ilgili Bayesçi koşullama şunu verir:

$$P_{\text{sonraki}}(w) = P_{\text{önsel}}(w \mid E) = \frac{P_{\text{önsel}}(E \mid w) P_{\text{önsel}}(w)}{P_{\text{önsel}}(E)}.$$

(E gibi bir önermenin olasılığı doğru olduğu olanaklı dünyalar olasılığının toplamıdır.)

en saf haliyle öğrenmenin resmidir. (Bir *hipotezin* olasılığını hesaplamak için o hipotezin doğru olduğu olanaklı dünyalara denk düşen tüm alanlardaki kum miktarını ölçmemiz yeterlidir.)

Buraya kadar, bir öğrenme kuralını tanımladık. Bir aracı elde etmek için, bir karar kuralına da ihtiyacımız var. Bu amaçla, aracıya mümkün olan her dünyaya bir sayı tahsis eden bir "yararlılık işlevi" bahşederiz. Bu sayı, aracının temel tercihlerine göre bu dünyanın arzulanabilirliğini temsil eder. Şimdi, aracı, her bir adımda, eylemi beklenen en yüksek yararlılıkla seçer.* (En yüksek yararlılık beklentisi olan eylemi bulmak için mümkün olan tüm eylemleri listeleyebilir. Ardından eylemdeki koşullu olasılık dağılımını –mevcut olasılık dağılımını eyleme henüz girişilmiş olduğu gözlemine göre koşullandırmaktan doğacak olan olasılık dağılımını– hesaplayabilir. Son olarak, eylemin beklenen değerini, mümkün olan her bir dünyanın değeri ile eylem durumundaki dünyanın koşullu olasılığının çarpımının toplamı olarak hesaplayabilir.†)

Öğrenme kuralı ve karar kuralı bir araya geldiğinde aracı için bir "optimallik fikri" tanımlarlar. (Özünde, aynı optimallik kavramı yapay zekâda, epistemolojide, bilim felsefesinde, iktisatta ve istatistikte de kullanılmıştır.‡) Gerçekte, bu tür bir aracı yaratmak imkânsızdır, çünkü gerekli hesaplamaları icra etmek sayısal olarak ölçülebilir değildir. Bunu yapmaya çalışmak GOFAI değerlendirmemizde tarif edilen türde bir kombinasyonel patlamaya boyun eğmek olur. Neden böyle olduğunu anlamak için, olanaklı dünyaların küçük bir altkümesini düşünün: Sonsuz bir boşlukta gezinen tek bir bilgisayar ekranından oluşuyor olsun. Ekranın pikseli 1000 x 1000'dir ve her biri ya açık ya da kapalıdır. Olanaklı dünyaların bu altkümesi bile muazzam büyüktür: $2^{(1000 \times 1000)}$ olanaklı ekran gözlemlenebilir evrende gerçekleşmesi beklenen tüm hesaplamaları sayıca aşar. Böylece, tek tek her birinde daha incelikli hesaplamalar yapmak bir yana, olanaklı tüm dünyaların bu küçük altkümesinde olanaklı tüm dünyaları sıralayamayız bile.

Optimallik fikri fiziksel açıdan gerçekleştirilemez olsa bile teorik açıdan ilgi çekici olabilir. Bize bulgulayıcı kestirimleri değerlendirmek için bir standart sunarlar ve bazen bir optimal aracının belli bir özel durumda ne yapacağına dair akıl yürütebiliriz. 12. Bölüm'de yapay aracılar için alternatif bazı optimallik kavramlarıyla karşılaşacağız.

---

* Veya beraberlik olma ihtimaline karşı, beklenen faydanın en yüksek olduğu olanaklı eylemlerden birini rastgele seçer.

† Bir eylemin beklenen faydası daha özlü şekilde şöyle yazılabilir: $EU(a) = \sum_{w \in W} U(w)P(w|a)$, (burada toplam tüm dünyalar üzerindendir).

‡ Örneğin bkz. Howson ve Urbach (1993); Bernardo ve Smith (1994); Russell ve Norvig (2010).

mantıksız denemeyecek şekilde, bir bilgisayarın ustalık düzeyinde satranç oynaması için yüksek derecede genel zekâyla donatılması gerektiği varsayılırdı.[31] Örneğin büyük satranç oyunculuğunun soyut kavramları öğrenebilmeyi, strateji hakkında akıllıca düşünebilmeyi, esnek planlar yapabilmeyi, birtakım dâhice mantıksal çıkarımlarda bulunabilmeyi ve belki de rakibinin düşüncesinin bir modelini çıkarabilmeyi gerektirdiği düşünülebilirdi. Öyle değilmiş. Özel amaçlı bir algoritma etrafında kusursuz satranç oynayan bir makine yaratmanın mümkün olduğu anlaşıldı.[32] Bunlar, 20. yüzyılın sonuna doğru mevcut hale gelen hızlı işlemcilere uygulandığında, çok güçlü bir oyun çıkarabilmektedir. Fakat bu şekilde yaratılan bir yapay zekâ dar kapsamlıdır. Satranç oynar, ama başka da bir şey yapamaz.[33]

Diğer alanlardaysa çözümlerin ilk başta düşünüldüğünden daha karmaşık olduğu ve ilerlemenin yavaş gittiği ortaya çıktı.

TABLO 1 Oyun oynayan Yapay Zekâ

Dama	İnsanüstü	Arthur Samuel'in ilk olarak 1952'de yazılan ve sonra geliştirilen (1955 versiyonu makine öğrenimini de kapsıyordu) dama programı, bir oyunu yaratıcısından daha iyi oynamayı öğrenen ilk program olur.[34] 1994'te CHINOOK programı son insan şampiyonu yener ve tarihte ilk kez bir program bir yetenek oyununda resmi dünya şampiyonluğunu kazanır. 2002'de Jonathan Schaeffer ve takımı damayı "çözerler", yani her zaman mümkün olan en iyi hamleyi yapan bir program üretmeyi başarırlar (alfa-beta araştırmasını 39 trilyon oyun sonu pozisyonunun olduğu bir veritabanıyla birleştirerek). Her iki tarafın da kusursuz oyunu beraberliğe yol açar.[35]
Tavla	İnsanüstü	1979: Hans Berliner'in tavla programı BKG dünya şampiyonunu yener. Tarihte ilk kez bir bilgisayar programı bir gösteri maçında da olsa bir dünya şampiyonunu herhangi bir oyunda yenmiş olur, ama Berliner sonradan galibiyeti "şans işte, zar geldi," diye açıklar.[36] 1992: Gerry Tesauro'nun tavla programı TD-Gammon zamansal farklılık öğrenimi (bir pekiştirmeli öğrenme biçimi) ve oyun tekrarlarını kendisine karşı kullana kullana şampiyonluk düzeyinde yeteneğe ulaşır.[37] O günden beri tavla programları en iyi insan oyuncuları bir hayli geride bırakmıştır.[38]
Traveller TCS	İnsanla işbirliği halinde insanüstü[39]	Hem 1981'de hem de 1982'de Douglas Lenat'ın Eurisko programı Traveller TCS'de (fütürist bir deniz savaşı oyunu) ABD şampiyonluğunu kazanır ve onun sıra dışı stratejilerini engellemek amacıyla kural değişiklikleri yapılır.[40] Eurisko donanmasını tasarlamak için bulgulayıcı kullanmıştı, hatta bulgulayıcısını değiştirmek için de bir bulgulayıcısı vardı.

Othello	İnsanüstü	1997: Logistello programı dünya şampiyonu Takeshi Murakami'ye karşı altı oyunluk maçta bütün oyunları kazanır.[41]
Satranç	İnsanüstü	1997: Deep Blue dünya satranç şampiyonu Garry Kasparov'u yener. Kasparov bilgisayarın bazı hamlelerinde gerçek zekâ ve yaratıcılık pırıltıları gördüğünü söyler.[42] O zamandan beri satranç makineleri ilerlemeyi sürdürüyor.[43]
Çapraz Bulmaca	Uzmanlık düzeyi	1999: Çapraz bulmaca çözme programı Proverb ortalama bir çapraz bulmacacıyı geçer.[44] 2012: Matt Ginsberg'in yarattığı Dr. Fill programı Amerikan Çapraz Bulmaca Turnuvası'nda insan yarışmacıları geçip çeyrek finale kalır. (Dr. Fill'in performansı değişkendi. İnsanların en zor diye nitelendirdikleri bulmacaları hata yapmadan bitiriyor, ama ters ya da çapraz yazma gerektiren standart dışı bulmacalarda çuvallıyordu.)[45]
Scrabble	İnsanüstü	2002 itibariyle Scrabble oynayan yazılım programı en iyi insan oyuncuları geride bırakıyor.[46]
Briç	En iyiye denk	2005'te briç oynayan yazılım en iyi insan briççilerle aynı seviyeye ulaşır.[47]
Jeopardy!	İnsanüstü	2010: IBM'in ***Watson***'ı ***Jeopardy!*** oyununun gelmiş geçmiş en büyük iki şampiyonu olan Ken Jennings ve Brad Rutter'ı dize getirir.[48] ***Jeopardy!*** tarih, edebiyat, spor, coğrafya, pop kültürü, bilim ve diğer konularda basit sorular soran bir televizyon yarışma programıdır. Sorular ipucu biçiminde sorulur ve çoğu zaman bir kelime oyunu içerir.
Poker	Farklı farklı	Bilgisayar poker oyuncuları Texas hold'em'da en iyi insanların biraz altında, ama diğer poker çeşitlerinde insanüstü düzeyde performans sergilerler.[49]
FreeCell	İnsanüstü	Genetik algoritmalar kullanarak evrilen bulgulayıcılar solitaire oyunu FreeCell için üst düzey insan oyuncuları yenebilecek bir çözücü üretirler.[50]
Go	Amatör düzeyde çok güçlü	2012 itibariyle, Go oynayan Zen serisi programlar Monte Carlo ağaç araması ve makine öğrenme tekniklerinden yararlanarak hızlı oyunlarda "6 dan" (çok güçlü amatör oyuncu) düzeyine ulaştı.[51] Go oynayan programlar son yıllarda yaklaşık 1 dan/yıl hızında ilerliyor. Eğer bu ilerleme hızı devam ederse on yıla kadar dünya şampiyonu bir insanı yenebilirler.

Bilgisayar mühendisi Donald Knuth, "Yapay zekâ bu zamana kadar özünde 'düşünme'yi gerektiren her şeyi yapmayı başarmış, ama insanların ve hayvanların 'düşünmeden' yaptığı birçok şeyi yapmayı başaramamıştır; nasıl olduysa olmuş, zoru başarmıştır!" diyerek şaşkınlığını dile getiriyordu.[52] Görsel sahneleri analiz etmenin, nesneleri tanımanın ya da bir robotun doğal çevreyle etkileşim halindeyken davranış-

larını kontrol etmenin zor bir iş olduğu anlaşıldı. Ne var ki epey bir ilerleme kaydedildi ve donanımda sürekli iyileştirmeler sayesinde hâlâ da ediliyor.

Aklıselim ve doğal dil anlayışının da zor olduğu anlaşıldı. Bu görevlerle ilgili tamamen insani düzeyde bir performansa ulaşmanın "Yapay zekâ için baştan aşağı" problem olduğu, yani bu problemleri çözmenin zorluğunun özünde genel itibariyle insan düzeyinde zeki makineler yaratma zorluğuna denk olduğu düşünülmüştür çoğu zaman.[53] Başka bir deyişle, eğer yetişkin bir insanı olduğu gibi doğal dili de anlayabilecek bir yapay zekâ yaratmak başarılacaksa, büyük ihtimalle ya insan zekâsının yapabileceği her şeyi yapabilen bir yapay zekâ yaratmak başarılmış ya da bu tür bir genel yeteneğe son bir adım kalmış olacak.[54]

Satranç oynama ustalığının şaşırtıcı derecede basit bir algoritmayla başarılabileceği anlaşıldı. Diğer yeteneklerin (genel muhakeme yetisi ya da programlamayı kapsayan başka bir kilit yeteneğin) de şaşırtıcı derecede basit bir algoritmayla elde edilebilir olup olmadığı konusunda spekülasyon yapmak çekici gelmiyor değil. Bir dönem en iyi performansın karmaşık bir mekanizmayla elde ediliyor olması hiçbir basit mekanizmanın işi bir o kadar iyi, hatta daha iyi yapabileceği anlamına gelmiyor. Pekâlâ, hiç kimse henüz basit alternatifi bulamamış olabilir. Ptolemaios'un sistemi (Dünya'yı evrenin merkezine, Güneş'i, Ay'ı, gezegenleri ve yıldızları da onun etrafına koyan sistem) bin yıldan fazla astronomide en son gelişme olarak kaldı ve kestirimsel doğruluğu yüzyıllar içinde modelin sürekli karmaşıklaştırılmasıyla ilerledi: Gökyüzündeki koyutlanmış hareketlere ardı ardına dış çemberler eklendi. Ardından, Kopernik çıkageldi ve onun daha basit ama –Kepler tarafından daha da geliştirildikten sonra– kestirim gücü daha kuvvetli olan Güneş merkezli teorisi tarafından alt edildi.[55]

Bugün yapay zekâ yöntemleri burada ele almayı gerektirmeyecek kadar çok alanda kullanılıyor; bunlara dair bir örnekseme sunarak uygulama alanının genişliği hakkında bir fikir vermek mümkün olabilir. **TABLO 1**'de listelenen oyun yapay zekâları haricinde, ortamdaki gürültüyü filtreleyen algoritmalara sahip işitme cihazları; sürücülere harita gösterip navigasyon tavsiyesi sunan yol bulma programları; kullanıcının önceki satın almaları ve puanlamaları temelinde kitap ve müzik albümü öneren tavsiye sistemleri ve doktorlara meme kanseri tanısı koymalarında yardımcı olan, tedavi planları öneren ve elektrokardiyogramların yorumlanmasına yardım eden tıbbi karar destek sistemleri vardır. Robot evcil hayvanlar ve temizlikçi robotlar, çim biçen robotlar, kurtarma robotları, cerrah robotlar ve bir milyondan fazla sınai robot vardır.[56] Dünyadaki robot sayısı 10 milyonun üstündedir.[57]

Saklı Markov modelleri gibi istatistik tekniklerine dayanan modern konuşma tanıma sistemleri, pratik kullanım için yeterli düzeye ulaşmıştır (bu kitabın bazı taslakları bir konuşma tanıma programının yardımıyla hazırlandı). Apple'ın Siri'si gibi bazı kişisel dijital asistanlar sözlü komutlara yanıt veriyor ve basit soruları cevaplayıp

komutları yerine getirebiliyorlar. El yazısıyla ve daktiloyla yazılmış metinlerde optik karakter tanıma özelliği posta tasnifi ve eski belgelerin dijital ortama aktarılması gibi uygulamalarda düzenli olarak kullanılıyor.[58]

Tercüme yapan makineler henüz kusursuz değildir, ama birçok uygulama için yeterince iyidir. Önceki sistemlerde GOFAI programıyla, her dil için yetenekli dilbilimcilerin en baştan geliştirmesi gereken elle kodlanmış gramerler kullanılıyordu. Yeni sistemler ise gözlemlenmiş kullanım örüntülerinden istatistiksel modelleri otomatik olarak yaratan istatistiksel makine öğrenimi teknikleri kullanıyor. Makine çift dilli bütünce analiziyle bu modeller için parametreler çıkarıyor. Bu yaklaşımda dilbilimcilere ihtiyaç kalmamıştır: Bu sistemleri yaratan programcıların çalıştıkları dili bilmelerine dahi gerek yoktur.[59]

Yüz tanıma programları son yıllarda o kadar ilerledi ki artık Avrupa'da ve Avustralya'da otomatik sınır geçişlerinde kullanılıyor. ABD Dışişleri Bakanlığı vize işlemleri için 75 milyon fotoğraflık bir yüz tanıma sisteminden yararlanıyor. Teftiş sistemleri giderek karmaşıklaşan yapay zekâ ve veri madenciliği teknolojilerini kullanarak ses, video ya da metin analiz ediyor; bunlar dünyanın elektronik iletişim araçlarından büyük miktarlarda alınıp devasa veri merkezlerinde depolanıyor.

Teorem kanıtlama ve denklem çözme artık o kadar yerleşik hale gelmiştir ki yapay zekâ olarak bile görülmüyor. Denklem çözücüler Mathematica gibi bilimsel bilgisayar programlarında kullanılıyor. Formel doğrulama yöntemleri (otomatik teorem kanıtlayıcılar dahil) çip üreticileri tarafından devre tasarımlarının davranışlarını üretimden önce doğrulamak için kullanılıyor.

ABD ordusu ve istihbarat birimleri bomba imha robotları, casus ve saldırı *drone*'ları ve diğer insansız araçların geniş çaplı kullanımında başı çekiyorlar. Bunlar hâlâ esasen insanlar tarafından uzaktan kumandayla kullanılıyor, ama özerk faaliyet alanlarını genişletme çalışmaları sürüyor.

Zeki planlama (zaman programlama) çok büyük başarılar elde edilen bir alandır. Otomatik lojistik planlama ve zaman programlama için DART aracı, 1991'de Çöl Fırtınası Operasyonu'nda öyle etkili şekilde kullanıldı ki DARPA (ABD'deki Savunma İleri Araştırma Projeleri Ajansı) tek başına bu uygulamanın yapay zekâya yaptıkları otuz yıllık yatırımı karşıladığını iddia etti.[60] Havayolu rezervasyon sistemleri karmaşık zaman programlama ve fiyatlandırma sistemleri kullanıyor. İşletmeler envanter kontrol sistemlerinde yapay zekâ tekniklerinden bir hayli istifade ediyor. Ayrıca otomatik telefonla rezervasyon sistemlerinden ve konuşma tanıma yazılımına bağlı yardım hatlarından yararlanarak bahtsız müşterilerinin iç içe geçmiş menü seçeneklerinin labirentlerinde kaybolmalarını engelliyorlar.

Yapay zekâ teknolojileri birçok internet hizmetinin temelini oluşturuyor. Yazılım dünya e-posta trafiğini denetliyor ve *spammer*'lar kendilerine karşı ek önlemleri bertaraf etmek için bunlardan devamlı yararlanmalarına karşın, Bayesçi spam filtreleri spam selini büyük oranda durdurmayı başarmıştır. Yazılım kullanan yapay zekâ programları kredi kartı işlemlerini otomatikman kabul ya da reddetmekten sorumludur ve dolandırıcılık faaliyetlerine karşı hesap hareketlerini sürekli takip etmektedir. Bilgi kurtarma sistemleri de makine öğreniminden yaygın ölçekte yararlanıyorlar. Google arama motoru ise kuşkusuz bu zamana kadar yaratılmış en iyi yapay zekâ sistemidir.

Şimdi, yapay zekâ ile genel olarak yazılım arasındaki çizginin kalın olmadığını vurgulamak gerekiyor. Yukarıda sıralanan uygulamalardan bazıları özelde yapay zekâdan daha jenerik yazılım uygulamaları olarak görülebilir, ama bu da bizi McCarthy'nin bir şey işe yaradığında artık yapay zekâ diye adlandırılmıyor sözüne geri getirir. Buradaki amaçlarımız açısından daha ilintili bir ayrım bilişsel yetenek menzili dar olan sistemler (yapay zekâ olarak adlandırılsın ya da adlandırılmasın) ile daha genel anlamda uygulanabilir problem çözme yetenekleri olan sistemler arasındadır. Özünde şu an kullanımda olan tüm sistemler ilk türdedir: Dar. Gelgelelim bunların birçoğu geleceğin yapay genel zekâsında bir rol oynayabilecek veya gelişimine yardımı dokunabilecek bileşenler içerir: Sınıflandırıcılar, arama algoritmaları, planlayıcılar, çözümcüler ve temsili çerçeveler gibi.

Bugün büyük paraların döndüğü ve son derece sıkı rekabetin olduğu küresel finans piyasalarında da yapay zekâ sistemleri yerini almıştır. Otomatik hisse senedi alım satım sistemleri büyük yatırımcılar tarafından yaygın olarak kullanılıyor. Bunlardan bazıları bir insan fon yöneticisinin belli alım satım emirlerinin otomatik olarak yapılmasının basit yollarıyken, diğerleri değişen piyasa koşullarına uyum sağlayan karmaşık ticaret stratejilerinden yararlanıyor. Analitik sistemler, borsadaki eğilimleri ve örüntüleri takip etmek ya da tarihsel fiyat hareketleri ile altta kayan yazılardaki kilit kelimeler gibi dış değişkenler arasında bağıntı kurmak için veri madenciliği teknikleri ve zaman serisi analizine başvuruyorlar. Finans haberleri yapanlar özel olarak bu tür yapay zekâ programları için kullanılmak üzere formatlanmış haberler satıyorlar. Diğer sistemler piyasaların içinde ya da arasında arbitraj fırsatları bulma konusunda ya da milisaniyeler içinde gerçekleşen çok küçük fiyat hareketlerinden kâr elde etmeye çalışan yüksek frekanslı alım satımda uzmanlaşmıştır (bu zaman ölçeğinde, fiberoptik kablolardaki ışık hızı sinyalleri için bile iletişim gecikmeleri kayda değer hale geliyor ve bilgisayarları borsanın yakınına yerleştirmeyi avantajlı kılıyor). Algoritmik yüksek frekanslı alım satım yapanlar ABD borsasında yapılan hisse senedi işlemlerinin yarısından fazlasını oluşturuyor.[61] Algoritmik ticaret 2010 "Flash Crash"ında [Borsanın Çöküşü] bir role sahiptir (Bkz. **KUTU 2**).

Makine zekâsının geleceğine dair fikirler

İki asli cephede ilerleme (bir yandan makine öğrenimi için daha sağlam istatistiksel ve enformasyon-teorik temel, diğer yandan çeşitli problemlere ya da alanlara özgü uygulamalardaki ticari başarı) yapay zekâ araştırmalarına kaybettiği itibarın bir kısmını geri kazandırmıştır. Gelgelelim yapay zekâ camiasına önceki tarihinden kalan bir kültürel etki sonucu, birçok ana akım araştırmacı, isminin çok hırslı projelerle anılmamasına özen gösteriyor. Mesela alanın gediklilerinden biri olan Nils Nilsson şu anki meslektaşlarının kendi kuşağının öncülerini işe koşturan o cesaretten ve ruhtan yoksun olduğundan yakınıyor:

> "Saygınlık" kaygısı kanımca bazı yapay zekâ araştırmacılarında ahmaklaştırıcı bir etkiye yol açtı. "Yapay zekâ eskiden gösterişliliğinden ötürü eleştirilirdi. Artık sağlam bir ilerleme kaydettiğimiz için, saygınlığımızı riske atmayalım," gibi şeyler söylediklerini duyuyorum. Bu muhafazakârlığın sonuçlarından biri "zayıf yapay zekâ"ya (insan düşüncesine yardımlarda bulunan türe) giderek daha fazla, "güçlü yapay zekâ"ya (insan düzeyinde zekâyı mekanikleştirmeye çalışan türe) daha az odaklanılmasıdır.[62]

Nilsson bu sözlerle Marvin Minsky, John McCarthy ve Patrick Winston gibi diğer kurucuların da hislerine tercüman olmuştur.[63]

Yapay zekâya olan ilgi son birkaç yılda yeniden artmıştır ve bu da yapay *genel* zekâ (Nilsson'ın verdiği adla, "güçlü yapay zekâ") yönünde yeniden çaba gösterilmesini doğurabilir. Daha hızlı donanıma ilaveten, günümüzdeki projelerden biri yapay zekânın birçok alt alanında, genel olarak yazılım mühendisliğinde ve hesaplamalı nörobilim gibi komşu alanlarda gerçekleştirilmiş büyük atılımlardan yarar sağlayacaktır. Nitelikli bilgi ve eğitim konusunda biriken talebin işaretlerinden biri 2011 güzünde Stanford Üniversitesi'nde Sebastien Thrun ve Peter Norvig'in yapay zekâya giriş dersinin internet üzerinden ücretsiz olarak verilmesine gelen yanıtlarda görülmektedir. Bu dersi almak için dünyanın farklı yerlerinden 160 bin kadar öğrenci başvurmuş ve 23 bini dersi tamamlamıştı.[64]

Uzmanların yapay zekânın geleceğine ilişkin fikirleri ciddi farklılıklar göstermektedir. Gerek yapay zekânın sonunda hangi biçimlere bürüneceği gerekse de zaman ölçeği konusunda görüş ayrılığı vardır. Yapay zekânın ilerideki gelişimine ilişkin tahminler, yakın dönemdeki bir çalışmada belirtildiği üzere, "kendinden emin olduğu kadar çeşitlilik de arz eder."[65]

Mevcut inanç dağılımı çok dikkatli ölçülmemiş olsa da, birçok küçük araştırmadan ve resmi olmayan gözlemden kaba bir izlenim edinebiliriz. Daha somutta, son döneme ait çeşitli anketlerde birkaç ilintili uzmanlık alanından kişilere, "birçok insan vazifesini en az bir insan kadar iyi yapabilen [...] insan düzeyinde makine zekâsı"nın (HLMI) geliştirilmesinin ne zaman mümkün olacağını düşündükleri sorulmuştur.[66]

Sonuçlar **TABLO 2**'de yer almaktadır. Bu örnekseme şu (medyan) tahmini sunuyordu: "HLMI olasılığı 2022'de yüzde 10, 2040'ta yüzde 50, 2075'te yüzde 90." (Katılımcılara tahminlerini "insanın bilimsel faaliyetinin çok büyük bir kesintiye uğramadan devam edeceği" varsayımına dayandırmaları söylenmiştir.)

Bu rakamlara belli bir şüpheyle yaklaşmak gerekir: Örnekseme büyüklükleri çok küçüktür ve genel uzmanların görüşlerini kesinkes temsil ediyor diye alınmamalıdır. Fakat yine de diğer araştırmalardan gelen sonuçlarla uyumludur.[67]

Bu yoklama sonuçları son dönemde yapay zekâ ile alakalı alanlardaki 25 kadar araştırmacıyla yapılan bazı mülakatlarla da uyumludur. Örneğin Nils Nelsson arama, planlama, bilgi temsili ve robotikle alakalı sorunlarla boğuştuğu uzun ve verimli bir kariyere sahiptir; yapay zekâyla ilgili ders kitapları yazmıştır ve yakın zaman önce bu alanda yazılmış en kapsamlı tarih eserini tamamlamıştır.[68] HLMI'nın ne zaman geleceğini tahmin ettiği sorulduğunda, şu görüşü ileri sürer:[69]

Yüzde 10 şans: 2030
Yüzde 50 şans: 2050
Yüzde 90 şans: 2100

TABLO 2 İnsan düzeyinde makine zekâsına ne zaman ulaşılacak?[70]

	%10	%50	%90
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Toplam	2022	2040	2075

Yayımlanan mülakatlardan hareketle söylersek, Profesör Nilsson'ın olasılık dağılımı alandaki birçok uzmanın görüşünü özetlemektedir, ama bir kez daha, çok farklı görüşler olduğunu söylemek gerekir: Alandaki bazı isimler çok daha emin konuşuyor ve HLMI'nın 2020-40 arasında bir dönem geleceğine, bir başka kesim ise asla gelmeyeceğine ya da çok uzak tarihte geleceğine inanıyor.[71] Ayrıca mülakata katılan bazı isimler "insan düzeyinde" bir yapay zekâ kavramının kötü bir tanım olduğunu ya da yanlış yönlendirici olduğunu düşünüyorlar veya başka nedenlerden ötürü yıl tahmini yapmak istemiyorlar.

Benim görüşüm ise uzman görüşlerinde aktarılan medyan rakamların sonraki varış tarihlerine ilişkin yeterli olasılık kütlesine sahip olmadığı yönündedir. HLMI'nın 2075'e, hatta 2100'e kadar geliştirilmeme olasılığı için ("insanın bilimsel faaliyetinin

KUTU 2: 2010 Flash Crash'i

6 Mayıs 2010 öğleden sonra, ABD hisse senedi piyasaları Avrupa'daki borç krizinden duyulan endişeden ötürü yüzde 4 gerilemişti. Saat 2.32'de büyük bir satıcı (bir yatırım fonu grubu) borsadaki dakika dakika likidite ölçümüyle bağlantılı bir satış hızında elden çıkarılacak çok sayıda E-Mini S&P 500 vadeli işlem sözleşmesinden kurtulmak için bir satış algoritması başlattı. Bu sözleşmeler algoritmik yüksek frekanslı alım satım yapanlar tarafından satın alındı, zaten onlar da sözleşmelerini başka borsacılara satarak geçici uzun vadeli durumlarını çabucak saf dışı bırakacak şekilde programlanmıştı. Asli alıcılardan talebin azalmasıyla algoritmik borsacılar E-Mini'leri esasen diğer algoritmik borsacılara satmaya başladılar, onlar da bunları diğer algoritmik borsacılara pasladılar ve böylece ortaya ticaret hacmini zirveye çıkaran bir "sıcak patates" çıktı; bu durum satış algoritması tarafından yüksek likiditenin işareti olarak yorumlandı ve E-Mini sözleşmeleri piyasaya sürme hızını artırmaya ve çöküşü hızlandırmaya neden oldu. Bir noktada, yüksek frekanslı borsacılar piyasadan çekilmeye başladılar ve fiyatlar düşmeye devam ederken likidite kalmadı. 2.45'te E-Mini alım satımları otomatik devre kesici (borsanın durma mantığı fonksiyonu) tarafından durduruldu. Alım satım yeniden başladığında (sadece beş saniye sonra), fiyatlar istikrara kavuştu ve çok geçmeden kayıplarının çoğunu kapamaya başladı. Ama bir süre krizin dip noktasındayken, bir trilyon dolar piyasadan silinmişti ve yayılma etkisi belli tahvillerde önemli sayıda alım satımın bir sent ya da 100 bin dolar gibi "absürd" fiyatlarda gerçekleştirilmesine yol açtı. Borsa günü kapadıktan sonra, borsa aktörleri düzenleyicilerle buluşup fiyat öncesi düzeyden yüzde 60 ya da daha fazla olan fiyatlarda yapılan tüm alım satımları durdurma kararı aldı (bu tür işlemler "açıkça hatalı" kabul edilecek ve dolayısıyla mevcut ticaret kuralları gereği sonradan iptal edilecekti).*

Burada bu hikâyeyi yeniden anlatarak konu dışına çıkmış oluyoruz, zira *Flash Crash*'teki bilgisayar programları özel olarak zeki ya da karmaşık değildi ve yarattıkları tehdit türü bu kitabın ileriki kısımlarında makine süper zekâsı olasılığına ilişkin olarak dile getireceğimiz kaygılardan farklıdır. Gelgelelim bu olaylardan birkaç yararlı ders çıkarılabilir. Bunlardan biri, tekil olarak basit bileşenler arasındaki işlemler (satış algoritması ve yüksek frekanslı algoritmik ticaret programları gibi), karmaşık ve beklenmedik sonuçlar doğurabilir. Bir sistemde yeni öğelerin devreye sokulması sistemik riskler (bir şeyler ters gidene kadar, hatta bazen o zaman bile bariz olmayan riskler) doğurabilir.†

Çıkarılacak bir diğer ders ise şudur: Akıllı profesyoneller duyarlı görünen ve normalde sağlam bir varsayıma dayanan bir programa talimat verebilirler (örneğin ticaret hacmi piyasa likiditesinin iyi bir ölçüsüdür) ve bu durum varsayımın geçersiz olduğu beklenmedik durumda bile program müthiş bir mantıksal tutarlılıkla talimatı etkilemeye devam ettiğinde

---

* CFTC ve SEC (2010). 6 Mayıs 2010 olaylarına ilişkin farklı bir perspektif için, bkz. CME Group (2010).

† Metindeki hiçbir şey algoritmik yüksek frekanslı ticaret aleyhinde bir argüman olarak yorumlanmamalıdır, zira normalde likiditeyi ve piyasa verimliliğini artırarak yararlı bir işlev görebilir.

feci sonuçlar doğurabilir. Algoritma tam da yapacağını yapar ve çok özel türde bir algoritma olmadıkça eylemlerinin absürd uygunsuzluğuna hayretler içinde bakakalmamızı umursamaz. Bu konuyla ileride yeniden karşılaşacağız.

Flash Crash ile ilgili üçüncü bir gözlem ise şudur: Otomasyonun olayda bir rolü olmakla birlikte, çözümüne de katkısı olmuştur. Fiyatlar alıp başını gittiğinde alım satımı durduran önceden programlanmış durdurma emri mantığı otomatikman devreye girecekti, zira tetikleyici olayların insanların yanıt veremeyeceği kadar hızlı şekilde gerçekleşebileceği öngörülmüştü. Önceden kurulmuş ve otomatik çalışan güvenlik fonksiyonu ihtiyacı –insan denetiminde çalıştırılan sistemden farklı olarak– bir kez daha makine süper zekâsına ilişkin değerlendirmemizde önemli olacak bir konunun habercisidir.*

* 1 Ağustos 2012'de kısmen, "devre kesici", ticarete konu olan hisselerin *sayısında* aşırı değişiklikler varsa ticareti durduracak şekilde programlanmadığı için (Popper 2012) piyasada küçük çaplı bir panik yaşandı. Bu bir kez daha sonraki izleklerden birinin habercisidir: Akla yatkın görünen bazı kuralların hangi özgül açılardan yanlış gidebileceğini öngörmenin zorluğu.

çok büyük bir kesintiye uğramadan devam edeceği"ni şerh düştükten sonra) yüzde 10 çok düşük bir tahmindir.

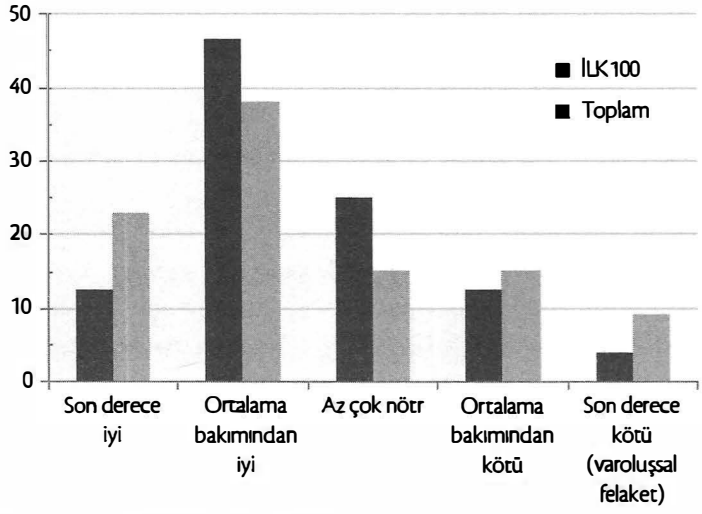
Tarihsel açıdan, yapay zekâ araştırmacılarının kendi alanlarındaki ilerleme hızlarını ya da gerçekleşecek önemli ilerlemeleri tahmin etme konusundaki sicilleri pek parlak değildir. Bir yandan, satranç oynama gibi bazı görevlerin şaşırtıcı derecede basit programlar tarafından yapılabildiği anlaşılmıştır, makinelerin "asla" şunu ya da bunu yapamayacağını söyleyen hayırcıların yanıldığı defalarca kanıtlanmıştır. Diğer yandan, alandaki isimlerin daha tipik hatalarından biri de bir sistemin gerçek dünyadaki görevleri düzgün şekilde yaparken karşılaşacağı zorlukları hafife almaları ya da kendi gözde projelerinin ya da tekniklerinin üstünlüklerini abartmalarıdır.

Araştırmada konumuzla alakalı iki soru daha sorulmuştur. Bunlardan birinde, insan düzeyinde makinenin ilk önce yaratılması durumunda süper zekâya ulaşmanın ne kadar zaman alacağı sorulmuştur. Sonuçlar TABLO 3'tedir.

Diğer sorudaysa insan düzeyinde makine zekâya ulaşmanın insanlığa uzun vadeli etkilerinin ne olacağı sorulmuştur. Cevaplar ŞEKİL 2'de özetlenmiştir.

TABLO 3 İnsan düzeyinde makineden süper zekâya uzaklık ne kadardır?

	HLMI'dan sonra 2 yıl içinde	HLMI'dan sonra 30 yıl içinde
İLK 100	yüzde 5	yüzde 50
Toplam	yüzde 10	yüzde 75



ŞEKİL 2 HLMI'nın uzun vadedeki genel etkisi.*

* Tabloda yeniden normalleştirilmiş medyan tahminleri gösterilmiştir. Araçlar epey farklıdır. Örneğin "son derece kötü" sonuç için ortalama tahminler %7,6 (*TOP100* için) ve %17,2 (uzman değerlendirmecilerin toplamı) idi.

Benim görüşüm ise bir kez daha araştırmada dile getirilen görüşlerden farklıdır. Süper zekânın insan düzeyinde makine zekâya ulaşıldıktan görece yakın zaman sonra yaratılmasının yüksek ihtimal olduğunu düşünüyorum. Ayrıca sonuçlara ilişkin daha kutuplaşmış bir bakışa sahibim ve son derece kötü ve son derece iyi sonuç ihtimallerinin daha dengeli bir sonuçtan daha muhtemel olduğunu düşünüyorum. Bunun nedenini kitabın ileriki kısımlarında açıklayacağım.

Küçük örneksemeler, seçimlerde taraflılık ve hepsinden öte öznel görüşlerin içkin güvenilmezliği dikkate alındığında bu uzman anketlerine ve mülakatlarına fazla anlam yüklememek gerekiyor; bize güçlü bir sonuç çıkarma olanağı vermiyorlar. Ama zayıf bir sonuç çıkarmak için bunlardan yararlanılabilir. En azından daha iyi veriler ya da analiz yerine insan düzeyinde makine zekâsının yüzyılın ortalarında geliştirilme olasılığının epey yüksek olduğuna ve bundan çok daha önce ya da sonra geliştirilmesi ihtimalinin çok az olduğuna; bundan epey bir süre sonra süper zekânın geliştirileceğine; bunun dışında bir sürü farklı seçeneğin mümkün olduğuna (son derece iyi sonuçlar da olabilir, insanlığın yok oluşu gibi son derece kötü sonuçlar da) inanmanın makul olabileceğini söylüyorlar.[72] Hiç olmadı, konuya daha yakından bakmakta yarar olduğunu gösteriyorlar.

Süper Zekâya Giden Yollar

Makineler şu anda genel zekâ bakımından insanlardan çok geridedir. Ama bir gün (diyoruz) süper zekâya sahip olacaklar. Peki, buradan oraya nasıl gideceğiz? Bu bölümde tasavvur edilebilir birkaç teknolojik yol ele alınıyor. Yapay zekâya, tüm beyin emülasyonu, biyolojik biliş ve insan makine arayüzlerinin yanı sıra ağlara ve örgütlenmelere de bakıyoruz. Bunların süper zekâya giden yollar olarak farklı inandırıcılık derecelerini değerlendiriyoruz. Birden çok yolun varlığı varış noktasına en azından biri üzerinden varılma olasılığını artırıyor.

Süper zekâyı muvakkaten şöyle tanımlayabiliriz: *İnsanların bilişsel performansını handiyse her alanda katbekat aşan her türlü zekâ.*[1]

Bir sonraki bölümde süper zekâ kavramından daha fazla bahsetme imkânı bulacağız ve süper zekânın mümkün olan bazı biçimlerini ayırt etmek için bir tür görüngesel analiz yapacağız. Ama şimdilik bu kaba tanım işimizi görür. Tanımın süper zekânın nasıl uygulanacağı konusunda bir şey söylemediğine dikkatinizi çekerim. Keza öznel boyut hakkında da öyle: Bir süper zekânın öznel bilinçli deneyime sahip olup olmayacağı, kimi sorunlarda (özellikle de bazı ahlaki sorunlar açısından) büyük önem arz edebilir, ama bizim buradaki esas odak noktamız zihnin metafiziği değil, süper zekânın nedensel öncelleri ve sonuçlarıdır.[2]

Satranç programı Deep Fritz bu tanıma göre süper zekâ değildir, zira Fritz sadece dar bir alan olan satrançta zekidir. Fakat belli bir alana özgü olan süper zekâlardan bazıları önemli olabilir. Belirli bir alanla sınırlı süper zekâ performansından bahsederken kısıtlılığı açıkça belirteceğiz. Sözgelimi bir "mühendislik süper zekâsı" mühendislik alanında şu an var olan en iyi zekâları misli misli aşan bir zekâ olacaktır. Bu terimle, aksi belirtilmedikçe, insanüstü düzeyde genel zekâya sahip olan sistemleri kastediyoruz.

Peki, bir süper zekâ nasıl yaratılabilir? Gelin mümkün olan yollardan bazılarını inceleyelim.

Yapay zekâ

Bu bölümün okurları bir yapay genel zekânın nasıl programlanacağı konusunda bir mavi kopya sunacağımızı düşünüyorlarsa yanılıyorlar. Böyle bir mavi kopya yok el-

bette. Zaten böyle bir mavi kopyaya sahip olsaydım, çok büyük ihtimalle bir kitapta yayımlamazdım. (Bunun nedenlerini anlamadıysanız, ileriki bölümlere kadar beklemeniz gerekiyor.)

Ne var ki ne tür bir sistem gerekeceğine dair bazı genel özellikleri belirtebiliriz. Bugün, öğrenme kapasitesinin genel zekâya ulaşmayı amaçlayan bir sistemin tasarımının ayrılmaz bir parçası olması gerektiği, sonradan eklenecek bir şey olmadığı açıktır. Aynısı belirsizliği ve olasılıkçı bilgiyi etkili şekilde ele alma yeteneği için de geçerlidir. Duyusal verilerden ve iç durumlardan yararlı kavramlar çıkarma ve edinilen kavramları mantıksal ve sezgisel muhakemede kullanım için esnek kombinasyonel temsillere aktarma yetisi de muhtemelen genel zekâya ulaşmayı amaçlayan modern bir yapay zekâdaki temel tasarım özelliklerindendir.

İlk başlardaki Eski Moda Güzel Yapay Zekâ (GOFAI) sistemleri, genel itibariyle, öğrenmeye, belirsizliğe ya da kavram oluşturmaya odaklanmamıştı, belki de bunun nedeni bu boyutları ele almak için gerekli tekniklerin o dönemde yeterince gelişmemiş olmasıydı. Bu demek değildir ki, altta yatan fikirlerin hepsi yenidir. Basit bir sistemi insan düzeyinde zekâya yüklemenin bir aracı olarak öğrenmeden yararlanma fikri en azından Alan Turing'in "çocuk makine" kavramına kadar götürülebilir. Turing bu konuda 1950'de şöyle yazmıştı:

Yetişkin zihnini simüle edecek bir program üretmeye çalışmak yerine, çocuğunkini simüle eden bir program üretmeye çalışmak neden olmasın? Bu sonrasında uygun bir eğitime tabi tutulursa, yetişkin beynine sahip olacaktır.[3]

Turing bu tür bir çocuk makine üretmek için bir yineleme süreci öngörüyordu:

> İlk teşebbüste iyi bir çocuk makine bulacağımızı sanmayalım. Böyle bir makineyi eğitip ne kadar iyi öğrendiğine bakmak gerekir. Ardından başka bir tane daha deneyip iyiye mi kötüye mi gittiğini gözlemleyebiliriz. Bu süreç ile evrim arasında bariz bir bağlantı var. [...] Fakat umalım ki bu süreç evrimden daha hızlı gerçekleşir. En uygun olanın hayatta kalması üstünlükleri ölçmek için yavaş bir yöntemdir. Deney yapan kişi zekâsını kullanarak bunu hızlandırabiliyor olmalıdır. Bir o kadar önemlisi, rastgele mutasyonlarla sınırlı değildir. Eğer bir zayıflığın nedeninin izini sürebilirse, onu iyileştirecek olan mutasyon türünü de herhalde düşünebilir.[4]

Körü körüne evrim süreçlerinin insan düzeyinde genel zekâ üretebildiğini biliyoruz, zira halihazırda en azından bir defa başarıldı. Öngörerek gerçekleştirilen evrimsel süreçler, yani zekâ sahibi bir insan programlayıcı tarafından tasarlanan ve yönlendirilen genetik programlar benzer bir sonucu çok daha verimli şekilde gerçekleştirebiliyor olmalıdır. Bu gözlem bazı felsefeciler ve bilimciler tarafından (David Chalmers ve Hans Moravec gibi) insan düzeyinde yapay zekânın sadece teorik açıdan mümkün olmayıp yaşadığımız çağda uygulanabilir de olduğunu savunmak için kullanılmıştır.[5]

Buradaki fikir şudur: Evrimin ve insan mühendisliğinin göreli yeteneklerinden yararlanarak zekâ üretmek ve insan mühendisliğinin evrimden bazı alanlarda halihazırda çok daha üstün olduğunu, geri kalan alanlardaysa çok geçmeden daha da üstün hale gelebileceğini görmek mümkündür. Dolayısıyla evrimin zekâ üretmiş olması insan mühendisliğinin çok geçmeden aynısını yapabileceğine işaret eder. Sözgelimi Moravec (daha 1976'da) şöyle yazmıştı:

> Bu kısıtlamalar dahilinde tasarlanan zekânın birkaç örneğinin olması kısa sürede aynısını yapabileceğimize dair büyük bir güven kaynağı olmalıdır. Durum havadan ağır hava araçlarının tarihiyle benzeştir: Kuşlar, yarasalar ve böcekler, kültürümüz uçmayı öğrenmeden önce bunun olanaklı olduğunu göstermişlerdi.[6]

Ne ki bu muhakeme çizgisinden çıkarılacak sonuçlara ihtiyatla yaklaşmak gerekir. Evrimin havadan ağır hava araçları ürettiği ve insan mühendislerin sonrasında (çok farklı bir mekanizmayla da olsa) benzerini yapmayı başardıkları doğrudur. Başka örnekler de zikredilebilir: Deniz radarı, manyetik navigasyon, kimyasal silahlar, fotoreseptörler ve her türlü mekanik ve kinetik performans özellikleri. Gelgelelim insan mühendislerin bu zamana kadar evrime yetişemedikleri alanlara da kolaylıkla işaret edebiliriz: Morfojenez, kendini onarma ve immün savunma gibi konularda insanlığın çabaları doğanın başardıklarının çok gerisinde kalmıştır. Dolayısıyla Moravec'in argümanı "kısa sürede" insan düzeyinde yapay zekâya erişebileceğimize dair "büyük bir güven kaynağı" olamaz. Olsa olsa, zeki yaşamın evrimi zekâ tasarlamanın içsel zorluğuna bir üst sınır getirir. Ama bu üst sınır insanlığın şu anki mühendislik yeteneklerinin epey bir üstünde olabilir.

Yapay zekânın gerçekleştirilebilir olduğunu savunmak için kullanılan evrimle ilgili argümanlardan bir diğeri de yeterince hızlı bilgisayarlarda genetik algoritmalar kullanarak biyolojik evrimin sonuçlarına benzer sonuçlar elde edebileceğimiz fikridir. Dolayısıyla evrim argümanının bu versiyonu zekânın nasıl üretileceğine dair özgül bir yöntem sunar.

Ama çok geçmeden insan zekâsını üreten ilintili evrim süreçlerini tekrarlamaya yeterli bilgisayar gücüne sahip olacağımız doğru mudur? Sorunun cevabı hem bilgisayar teknolojisinin önümüzdeki yirmi otuz yılda ne kadar ilerleyeceğine hem de genetik algoritmaları geçmişimizde yer alan doğal seçilimin evrimsel süreciyle aynı optimizasyon gücüyle işletmek için ne kadar bilgisayar gücü gerekeceğiyle alakalıdır. Sonunda bu muhakemeyle vardığımız sonuç hayal kırıklığı yaşatacak kadar belirsiz olsa da, kaba bir tahmin yapmak bilgilendirici olacaktır (bkz. **KUTU 3**). Hiçbir şey olmasa, bu alıştırma bazı ilginç bilinmeyenlere dikkatimizi çekecektir.

Mesele şu ki insan düzeyinde zekâyı üreten yeryüzündeki ilintili evrimsel süreçleri basitçe tekrarlamak için gerekli bilgisayar kaynakları elimizin altında değildir ve Moore

yasası bir yüzyıl daha hüküm sürse bile bu değişmeyecektir (karş. ŞEKİL 3). Gelgelelim doğal evrimsel süreçlerin kaba kuvvetle tekrarlanmasıyla kıyaslandığında, zekâyı hedefleyecek şekilde arama sürecini tasarlayıp doğal seçilime karşı bariz üstünlüklerden yararlanarak birçok verimlilik kazanımı elde edilebilir. Fakat bu edinilebilir verimlilik kazanımlarının büyüklüğünü ölçmek çok zordur. Büyüklük derecesinin beş mi yirmi beş mi olduğunu bile söyleyemiyoruz. Dolayısıyla daha fazla ayrıntının yokluğunda, evrimsel argümanlar insan düzeyinde makine zekâsı üretmenin zorluğunu ya da bu tür gelişmelerin zamansal ölçeğine dair beklentilerimizi anlamlı şekilde sınırlayamaz.

Bu tür evrimsel mülahazaların mahzurlu bir yanı daha var ki, bunlardan zekânın evriminin zorluğuna dair çok gevşek de olsa bir üst sınır çıkarmayı zorlaştırmaktadır. Zeki yaşamın yeryüzünde verilmiş olduğu gerçeğinden hareketle evrimsel süreçlerin zekâ üretme konusundaki önceki olasılığının makul derecede yüksek olduğu çıkarımını yapma yanılgısına düşmemeliyiz.

KUTU 3: Evrimi tekrarlamak için neler gereklidir?

Evrimin insan zekâsının gelişimi sırasında elde ettiği her başarı yapay yollarla makine zekâ evriltmeye çalışan bir insan mühendisin işine yaramaz. Yeryüzündeki evrimsel seçilimin sadece küçük bir kısmı zekâ seçilimidir. Daha somuta inersek, insan mühendislerin öylece geçiştiremediği problemler evrimsel seçilimin bütününün çok küçük bir kısmının hedefi olmuş olabilir. Örneğin bilgisayarlarımızı elektrikle çalıştırdığımızdan, zeki makineler yaratmak için hücresel enerji ekonomisinin moleküllerini yeniden icat etmemize gerek yoktur; ama metabolizma yolaklarının bu tür bir moleküler evrimi yeryüzünün tarihi boyunca evrimde hazır bulunan toplam seçilim gücü miktarının büyük bir kısmını harcamış olabilir.*

Yapay Zekâ için kilit içgörülerin bir milyar yıldan daha az bir süre önce ortaya çıkan sinir sistemlerinin yapısında tecessüm ettiği savunulabilir.† Eğer bu görüşü savunursak o zaman evrim için mevcut olan ilintili "deneyler"in sayısı ciddi derecede azalacaktır. Bugün dünyada $4–6 \times 10^{30}$ kadar prokaryot var, ama sadece 10^{19} böcek var, insan sayısı ise 10^{10}'dan azdır (tarım öncesi topluluklar ise daha küçük boyuttaydı).‡ Bu sayılar çok da göz korkutucu değildir.

Gelgelelim evrimsel algoritmalar sadece arasından seçeceği çeşitliliklere değil, bu çeşitlilikleri değerlendirecek bir uygunluk işlevine de gerek duyar ve bilgisayarlar açısından en

---

* Ayrıca bkz. Shulman ve Bostrom (2012). Orada bu meseleler daha ayrıntılı ele alınmıştır.

† Legg (2008) bu gerekçeyi insanların evrimin ilerleyişini çok daha kısa zaman ölçeğinde ve daha az hesaplama kaynaklarıyla özetleyebileceği iddiasını desteklemek için sunar (ama evrimin düzenlenmemiş hesaplama kaynaklarının ulaşılamaz olduğunu da belirtmeden geçmez). Baum (2004) yapay zekâyla ilgili bazı gelişmelerin daha önceden gerçekleştiğini ve bizatihi genom örgütlenmesinin evrimsel algoritmalar için değerli bir temsili cisimleştirdiğini savunur.

‡ Whitman vd. (1998); Sabrosky (1952).

pahalı bileşen de genellikle budur. Yapay zekânın evrimi için uygunluk işlevi, uygunluğu değerlendirmek için nöral gelişimi, öğrenme ve biliş yetisinin simülasyonunu gerektirir. Dolayısıyla karmaşık sinir sistemleri olan organizmaların ham rakamına bakmak yerine evrimin uygunluk işlevini taklit etmek için simüle etmemiz gerekebilecek biyolojik organizmalardaki nöronların sayısına bakabiliriz. Bu niceliğe dair kaba bir tahmin yapmanın yolu karada yaşayan hayvanlara hükmeden böcekleri düşünmektir (sadece karıncaların sayısının yüzde 15-20 arasında olduğu tahmin edilmektedir).* Böceklerin beyin büyüklükleri büyük farklılıklar göstermekte, büyük ve sosyal böceklerin daha büyük beyinleri bulunmaktadır: Bir bal arısının beyninde neredeyse 10^6 nöron, bir meyve sineğininkinde 10^5 nöron vardır; karıncalarınkindeyse 250 bin civarındadır.† Küçük böceklerin çoğunluğunun sadece birkaç bin nöronluk beyinleri olabilir. Muhafazakâr derecede yüksek tarafta yanılmayı göze alarak, eğer tüm 10^{19} böceğe meyve sineği sayısında nöron tahsis edersek, toplam rakam olarak dünyada 10^{24} böcek nöronuna ulaşırız. Buna su kopepodu, kuş, sürüngen, memeli vb. için ek büyüklükleri ilave ederek 10^{25}'e ulaşılabilir. (Halbuki tarım öncesi dönemde 10^7'den az insan vardı ve her birinde 10^{11}'den az nöron bulunuyordu: Sonuç olarak insanlarda nöron başına sinaps sayısı fazla olsa da toplamda 10^{18}'den az insan nöronu eder.)

Bir nöronu simüle etmenin hesaplama maliyeti simülasyondaki ayrıntı düzeyine bağlıdır. Son derece basit nöron modelleri, gerçek zamanda tek bir nöronu simüle etmek için saniye başına 1000 yüzer-nokta işleminden (FLOPS) yararlanır. Elektrofizyolojik açıdan gerçekçi Hodgkin-Huxley modeli 1.200.000 FLOPS kullanır. Daha ayrıntılı çok bölmeli bir model üç dört kat daha büyüklük ekleyecekken, nöron sistemlerini içeren yüksek düzey modeller basit modellerden iki üç kat büyüklük eksiltebilirler.‡ Eğer bir milyar yıllık evrimde 10^{25} nöronu simüle edersek (bizim bildiğimiz şekliyle sinir sistemlerinin varlığından daha uzundur) ve bilgisayarlarımızın bir yıl boyunca çalışmasına izin verirsek, bu rakamlar 10^{31}-10^{44} FLOPS aralığında bir gereklilik demektir. Karşılaştırma olması adına, Eylül 2013 itibariyle dünyanın en güçlü süper bilgisayarı olan Çin'in Tianhe-2'si sadece $3{,}39 \times 10^{16}$ FLOPS sağlar. Son dönemde kişisel bilgisayarların bir birim büyüklük artışı gerçekleştirmeleri yaklaşık 6,7 yıl almaktadır. Bu açığı kapamak için bir yüzyıl boyunca Moore yasasının sürdürülmesi bile yeterli olmayacaktır. Daha uzmanlaşmış donanım kullanmak ya da daha uzun işleyiş süresine izin vermek ancak birkaç birim daha büyüklük sağlayabilir.

Bu rakam başka bir açıdan da muhafazakârdır. Evrim insan zekâsını bu sonucu amaçlamadan doğurmuştur. Başka bir deyişle, doğal organizmalar için uygunluk işlevleri sadece zekâyı ve öncellerini seçmez.§ Üstün bilgi işleme yetenekleri olan organizmaların çeşitli ödüller elde

---

* Schultz (2000).

† Menzel ve Giurfa (2001, 62); Truman vd. (1993).

‡ Sandberg ve Bostrom (2008).

§ Bu hususta ve ayrıca saf zekâ testlerinin pürüzsüz manzarasına dayalı uygunluğu belirleyen çevrelerin ya da işlevlerin vaadi hakkında daha fazla değerlendirme için, bkz. Legg (2008).

ettiği çevrelerde bile zekâ seçilmeyebilir, çünkü zekâdaki ilerlemeler ciddi bedel ödetebilir ve çoğu zaman ödetir de. Yüksek enerji harcanması ya da daha yavaş olgunlaşma süreleri gibi bu tür bedeller daha akıllı davranıştan elde edilebilecek tüm faydaları gölgede bırakabilir. Aşırı derecede ölümcül çevreler zekânın değerini de azaltır: İnsanın ortalama ömrü ne kadar kısaysa, artan öğrenme yetisi için o kadar az zaman kalacaktır. Zekâ için seçilim baskısının azalması zekâ artırıcı yeniliklerin yayılmasını ve dolayısıyla seçilimin onlara bağımlı olan sonraki yeniliklere yardımcı olma fırsatını yavaşlatır. Ayrıca evrim insanların –fark etseler bile– sömürü ile keşif arasındaki ödünleşimleri değiştirerek ya da giderek zorlaşan zekâ testlerinin pürüzsüz bir ilerleyişini sağlayarak es geçecekleri yerel optimumlara saplanıp kalabilir.* Daha önce bahsedildiği üzere, evrim, seçilim gücünün çoğunu zekâyla alakasız olan özelliklere dağıtır (Kızıl Kraliçe'nin bağışıklık sistemleri ile asalaklar arasındaki rekabetçi ortak evrim yarışları gibi). Evrim istikrarlı şekilde ölümcül olduğu kanıtlanmış mutasyonlar üreten kaynakları çarçur etmeyi sürdürür ve farklı mutasyonların sonuçlarında istatistiksel benzerliklerden yararlanmayı başaramaz. Bunların hepsi doğal seçilimdeki yetersizliklerdir (zekânın evrimi açısından değerlendirildiğinde böyledir) ve bir insan mühendisin zeki yazılım geliştirmek için evrimsel algoritmalardan yararlanırken bunlardan kaçınmaya çalışması görece kolay olacaktır.

Yukarıda tarif edilen türde verimsizlikleri ortadan kaldırmanın daha önce hesaplanan 10^{31}-10^{44} FLOPS'tan birçok birim büyüklüğü götüreceğini söylemek akla yatkındır. Ne yazık ki bu büyüklüğün kaç birim olacağını söylemek güçtür, hatta kaba bir tahmin yapmak bile güçtür, hiç belli olmaz, beş birim de olabilir, on da, yirmi beş de.†

---

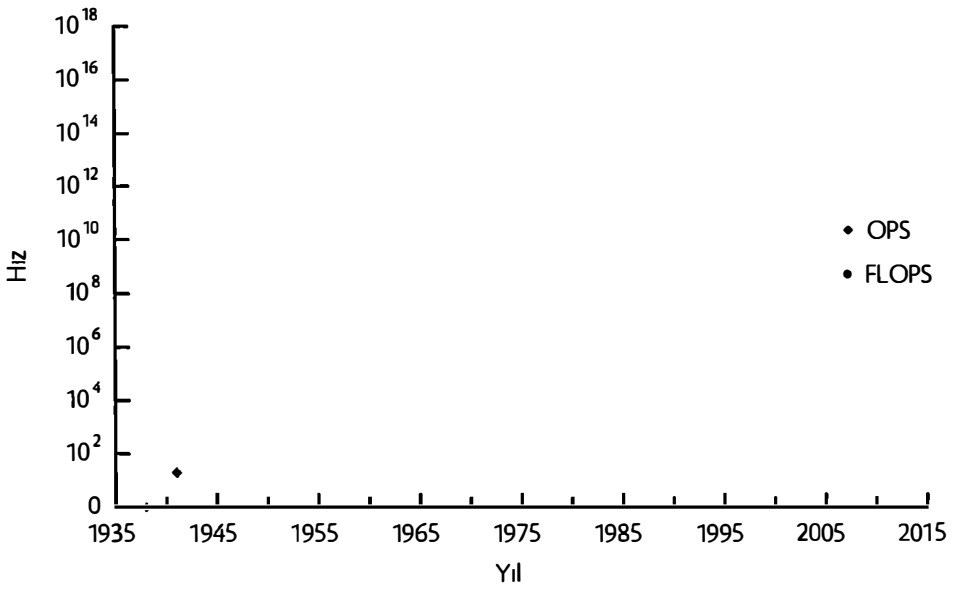
* Mühendislerin tarihi evrimsel seçilimi hangi şekillerde aşabildiklerine dair daha ayrıntılı bir değerlendirme ve taksonomi için, bkz. Bostrom ve Sandberg (2009b).

† Analizde beden simülasyonlarının maliyeti ya da uygunluk işlevinin parçası olarak çevreleyici sanal çevreye atıfta bulunulmadan canlı varlıkların sinir sistemleri ele alınmıştır. Yeterli bir uygunluk işlevinin doğal ömrü boyunca bu organizmanın beyninin tüm nöral hesaplamayı simüle etmek için gerektiğinden çok daha az işlemle belli bir organizmanın yeterliliğini test edebilecek olması akla yatkındır. Bugün yapay zekâ programları çoğu zaman çok soyut çevrelerde gelişir ve işler (sembolik matematik dünyalarında teorem kanıtlayıcıları, basit oyun turnuvası dünyalarındaki menajerler, aracılar vb.).

Bir şüpheci soyut bir çevrenin genel zekânın evrimi için yetersiz olacağında ısrar edebilir ve sanal çevrenin atalarımızın evrildikleri gerçek biyolojik çevreye yakından benzemesi gerektiğine inanıyor olabilirler. Fiziksel açıdan gerçekçi bir sanal dünya yaratmak, bir soyut problem alanı ya da basit bir oyuncak dünyasının simülasyonuna kıyasla hesaplama kaynaklarından çok daha fazla yatırım gerektirecektir (oysa evrimin fiziksel açıdan gerçekçi bir dünyaya "bedavadan" erişimi vardı). Sınırlayıcı örnekte, eğer tam mikro-fiziksel doğrulukta ısrar edilirse, hesaplama gereklilikleri gülünç boyutlara ulaşacaktır. Gelgelelim bu türde aşırı bir karamsarlığın mesnetsiz olduğu neredeyse kesindir; zekâ evrimi için en iyi çevrenin doğayı olabildiğince yakından taklit eden çevre olması muhtemel görünmüyor. Bilakis, bir yapay seçilim çevresinden, atalarımızınkinden çok farklı olan, evrilmesi için çalıştığımız zekâ türünü (sözgelimi azami derecede hızlı içgüdüsel tepkiler ya da son derece optimum hale getirilmiş bir görme sisteminden farklı olarak soyut muhakeme ve genel problem çözme yetenekleri) artıran uyarlanımları teşvik etmek üzere özel olarak tasarlanmış bir çevreden yararlanmanın daha verimli olması akla yatkındır.

Bu tür bir çıkarım sağlıksızdır, zira tüm gözlemcilerin kendilerinin zeki yaşamın ortaya çıktığı bir gezegende –böyle bir gezegenin zekâyı üretme olasılığının ne kadar az ya da çok olduğundan bağımsız olarak– ortaya çıktıklarını görmelerini teminat altına alan gözlem seçilim etkisini gözden kaçırır. Örneğin zeki yaşamı üretmek için doğal seçilimin sistemli sonuçlarına ilaveten ciddi düzeyde bir şanslı tesadüfün de gerektiğini varsayalım; o kadar ki zeki yaşam basit çoğaltıcıların ortaya çıktığı her 10^{30} gezegenden sadece birinde evrilir. Bu durumda, doğal evrimin yaptığını tekrarlamaya çalışırken genetik algoritmalarımızdan yararlandığımızda, tüm öğelerin tam da doğru şekilde bir araya geldiği bir örneğe ulaşmak için 10^{30} kadar simülasyon denememiz gerektiğini öğrenebiliriz. Bu durum yaşamın burada, yeryüzünde evrildiğine dair gözlemimizle tamamen tutarlıdır. Ancak dikkatli ve biraz incelikli bir muhakemeyle (zekâyla alakalı vasıfların yakınsak evriminin örneklerini analiz ederek ve gözlem seçilim teorisinin incelikleriyle ilgilenerek) bu epistemolojik engeli kısmen aşabiliriz. Bunu yapma zahmetine girmezsek, KUTU 3'ten üretilen zekânın evrimini tekrardan özetlemek için gerekli bilgisayar güçlerine getirilen sözümona "üst sınır"ın, otuz kat kadar az büyüklükte (ya da buna benzer) olabileceği ihtimali dışlanamaz.[7]

Yapay zekânın gerçekleştirilebilirliğini savunmanın bir diğer yolu da insan beynine işaret ederek bunu makine zekâ için bir kalıp olarak kullanabileceğimizi söylemektir. Bu yaklaşımın farklı versiyonlarını biyolojik beyin işlevlerini ne kadar yakından taklit edebileceklerine bağlı olarak ayırt edebiliriz. Bir uçta (çok yakından taklit edebilenler) *tüm beyin emülasyonu* vardır; bunu bir sonraki alt başlıkta ele alacağız. Diğer uçtaysa ilhamını beynin işleyişinden alan ama düşük düzey taklit yapmaya çalışmayan yaklaşımlar vardır. Nörobilimdeki ve bilişsel psikolojideki ilerlemeler (aletlerdeki ilerlemelerden de yardım alarak) sonunda beynin işleyişinin genel ilkelerini ortaya serebilir ve ardından da bu bilgi yapay zekâ çabalarına kılavuzluk edebilir. Beyinden ilham almış yapay zekâ tekniğinin bir örneği olarak nöral ağlardan yukarıda bahsetmiştik. Hiyerarşik algısal örgütlenme de beyin biliminden makine öğrenimine nakledilmiş bir diğer fikirdir. Pekiştirici öğrenme incelemeleri psikolojideki hayvan bilişi teorilerindeki rolünden (en azından kısmen) ilham almıştır ve bu teorilerden ilham alan takviyeli öğrenme teknikleri (örneğin "TD-algoritması") bugün yapay zekâda yaygın olarak kullanılmaktadır.[8] Gelecekte bu türden daha fazla örneğin birikeceği açıktır. Beyinde işlev gösteren ayrı temel mekanizmaların sayısı sınırlı (belki de çok az) olduğundan, beyin bilimindeki artımlı ilerlemenin sürmesi sonucunda bunların hepsi nihayet keşfedilecektir. Fakat bu gerçekleşmeden önce beyinden esinlenmiş bazı tekniklerle bazı saf yapay yöntemleri birleştiren melez bir yaklaşımın bitiş çizgisini geçmesi mümkündür. Bu durumda, geliştirme sürecinde beyinden türetilmiş bazı içgörüler kullanılsa da ortaya çıkan sistemin tanınabilir derecede beyne benzer olması gerekmez.



ŞEKİL 3 Süper bilgisayar performansı. "Moore yasası", dar bir anlamda, entegre devrelerdeki transistör sayısının son yirmi otuz yıldır yaklaşık iki yılda bir ikiye katlandığı gözlemine atıfta bulunur. Gelgelelim bu terim çoğu bilgisayar teknolojisindeki birçok performans ölçümünün benzer derecede hızlı bir büyüme trendi izlediğine dair daha genel gözleme atıfta bulunmak için de sık sık kullanılır. Burada dünyanın en hızlı süper bilgisayarının en üst hızını zamanın bir fonksiyonu olarak (logaritmik bir dikey ölçekte) kullanıyoruz. Son yıllarda işlemcilerin seri hızındaki büyüme durmuştur, ama paralelleşmeden giderek daha fazla yararlanılması gerçekleştirilen toplam hesaplamaların trend çizgisinde kalmasını sağlamıştır.*

* Wikipedia (2012b).

Bir kalıp olarak beynin varlığı makine zekâsının er geç gerçekleştirilebileceği iddiası için güçlü bir destek sağlar. Fakat bu bize keşfin ne zaman gerçekleşeceğini söyleme imkânı sunmaz, çünkü beyin bilimindeki keşiflerin gelecekteki hızını kestirmek zordur. Söyleyebileceğimiz şey şudur: Geleceğe ne kadar bakarsak, beynin işlevselliğinin gizlerinin bu tarzda makine zekâsının yaratılmasını mümkün kılacak kadar deşifre edilme ihtimali o kadar artacaktır.

Makine zekâ yaratmak için çalışan farklı isimler, tamamen sentetik tasarımları amaçlayan yaklaşımlarla kıyaslandığında nöromorfik yaklaşımların ne kadar ümit vaat ettiğine dair farklı görüşlere sahiptir. Kuşların varlığı havadan ağır uçuşun fiziken mümkün olduğunu kanıtlamış ve uçan makineler yapma çabalarını teşvik etmiştir. Ama ilk çalışan uçaklar kanat çırpmamıştır. Makine zekâsının da insanların yapay bir mekanizma yoluyla başardıkları uçuş gibi mi yoksa doğal yollarla gerçekleşen yangınları kopyalayarak hatmettiğimiz tutuşma işlemi gibi mi olacağı tartışmaya açıktır.

Turing'in içeriğinin çoğunu en baştan önceden programlamaktan ziyade öğrenerek edinen bir program tasarlama fikri, makine zekâsı konusundaki nöromorfik yaklaşımlar ile sentetik yaklaşımlara aynı derecede uygulanabilir.

Turing'in çocuk makine anlayışının bir çeşidi de "çekirdek yapay zekâ" fikridir.[9] Turing'in öngördüğü şekliyle, bir çocuk makine içkin potansiyelliklerini içerik biriktirerek geliştiren görece sabit bir mimariye sahip olacakken, bir çekirdek yapay zekâ kendi mimarisini geliştirebilecek daha incelikli bir yapay zekâya olacaktır. Çekirdek yapay zekânın ilk aşamalarında bu tür ilerlemeler esasen deneme yanılma, bilgi edinme ya da programcılardan yardım yoluyla gerçekleşebilir. Ama sonraki aşamalarda bir çekirdek yapay zekâ bilişsel performansını ön yükleme yapacak yeni algoritmaları ve hesaplamalı yapıları yaratacak kadar kendi işleyişini anlayabiliyor olmalıdır. Bu gerekli anlayış çekirdek yapay zekânın birçok alanda yeterli genel zekâ düzeyine ulaşmasından ya da bilgisayar bilimi veya matematik gibi özellikle ilintili bir alanda bir eşiğin geçilmesinden doğabilir.

Bu da bizi bir diğer önemli kavrama getiriyor: "Yinelemeli kendini geliştirme." Başarılı bir çekirdek yapay zekâ kendi kendini yinelemeli olarak ilerletecektir: Yapay zekânın ilk versiyonlarından biri kendisinin gelişmiş bir versiyonunu tasarlayabilir ve gelişmiş versiyon (ilk halinden daha zeki olduğundan) kendisinin daha da zeki bir versiyonunu tasarlayabilir ve bu böyle gider.[10] Bazı koşullarda bu tür bir yinelemeli kendini geliştirme süreci bir zekâ patlamasını doğuracak kadar uzun sürebilir; kısa bir süre zarfında bir sistemin zekâ düzeyi görece mütevazı bilişsel yeteneklere sahipken (belki de birçok açıdan insandan aşağı düzeyde ama kodlama ve yapay zekâ araştırması için alana özgü yetenekleri bulunan) radikal bir süper zekâya sahip hale gelebilir. Bu önemli olanağa 4. Bölüm'de geri döneceğiz ve orada bu hadisenin dinamikleri daha yakından analiz edilecek. Fakat bu modelin sürprizlere açık kapı bıraktığını belirtelim: Yapay genel zekâ inşa etme girişimleri son kritik bileşen yerli yerine oturtulana kadar çok başarısız olabilir ve ancak bu noktada bir çekirdek yapay zekâ yinelemeli kendini geliştirmeyi sürdürebilecek hale gelebilir.

Bu alt başlığı tamamlamadan önce vurgulamamız gereken bir husus daha var: Bir yapay zekânın insan zihnine çok da benzemesi gerekmez. Yapay zekâlar son derece yabancı olabilir, hatta olması muhtemeldir. Bilişsel mimarilerinin biyolojik zekâlarınkinden çok farklı olmasına şaşırmamak gerekir ve gelişimlerinin ilk aşamalarında çok farklı bilişsel zayıflıkları ve güçlü yanları olacaktır (ileride savunulacağı üzere, sonunda ilk baştaki zayıflıklarını aşabilirler). Ayrıca yapay zekâların hedef sistemleri insanlarınkinden çok farklı olabilir. Türsel bir yapay zekânın sevgi, nefret, gurur ya da benzer ortak insan duygularıyla hareket etmesini beklemek için bir neden yoktur: Bu karmaşık uyarlanımların yapay zekâlarda yeniden yaratılabilmesi kasıtlı ve ciddi bir çaba gerektirir. Bu hem büyük bir sorundur hem de büyük bir fırsattır.

Yapay zekâ motivasyonu meselesine ileriki bölümlerde geri döneceğiz, ama bu kitabın argümanı açısından öyle merkezi bir yere sahiptir ki her an akılda tutmakta yarar var.

Tüm beyin emülasyonu

Tüm beyin emülasyonunda ("zihin yükleme" olarak da bilinir), akıllı yazılım biyolojik beynin hesaplamalı yapısını tarayarak ve yakından örnek alarak üretilecektir. Dolayısıyla bu yaklaşım doğadan ilham alan sınırlı bir örneği temsil eder: Alenen intihal. Tüm beyin emülasyonunu başarmak şu adımları atabilmeyi gerektirir.

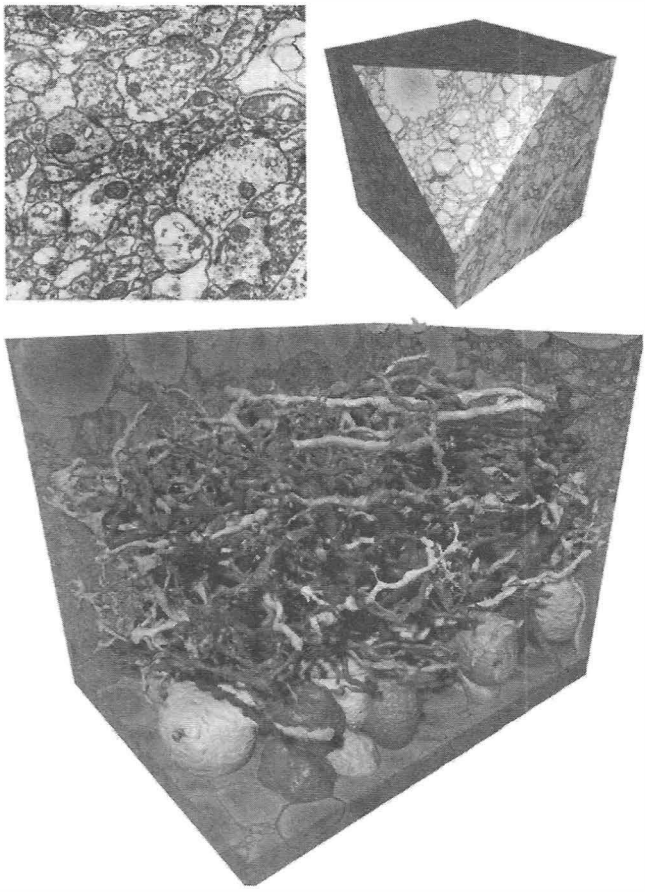
Birincisi, belli bir insan beyninin yeterince ayrıntılı bir görüntüsü çıkartılır. Bunun için beynin ölüm sonrasında vitrifikasyon yoluyla stabilize edilmesi de gerekebilir (böylece doku bir tür cama dönüşür). Ardından bir makine, dokuyu ince dilimler halinde keser ve bunlar belki de bir dizi elektron mikroskobu tarafından tarama için başka bir makineye aktarılabilir. Bu aşamada farklı yapısal ve kimyasal özellikleri açığa çıkarmak için çeşitli koyulaştırıcı kimyasallar uygulanabilir. Birçok tarama makinesi birden çok beyin dilimini aynı anda işlemek için yan yana çalışabilir.

İkincisi, tarayıcılardan gelen ham veriler otomatik görüntü işlemenin özgün beyinde bilişi yerine getirmiş olan üçboyutlu nöral ağı yeniden kurması için bir bilgisayara nakledilir. Pratikte bu adım ilk adımla aynı anda arabelleklerde depolanan yüksek çözünürlüklü görüntü verilerinin miktarını azaltacak şekilde yürütülebilir. Ardından ortaya çıkan harita farklı türde nöron ya da farklı nöron elementlerinin nöral hesaplama modelleriyle birleştirilir (belli türde sinaptik konektörler gibi). **ŞEKİL 4**'te günümüz teknolojisiyle üretilmiş tarama ve görüntü işleme sonuçlarından bazıları gösteriliyor.

Üçüncü aşamada, önceki adımdan doğan nöral hesaplama yapısı yeterince güçlü bir bilgisayarda uygulamaya dökülür. Eğer tamamen başarılıysa, sonuç belleğe ve kişiliğe zarar vermeden özgün aklın dijital olarak yeniden üretilmesi olacaktır. Taklit edilmiş insan zihni artık bir bilgisayarda yazılım olarak mevcuttur. Zihin robot eklentiler yoluyla dış dünyayla arayüz kurabilir ya da sanal bir gerçekliği işgal edebilir.

Tüm beyin emülasyonunun yolu insandaki biliş yetisinin nasıl işlediğini ya da bir yapay zekânın nasıl programlanacağını bilmemizi gerektirmez. Sadece beynin temel hesaplamalı öğelerinin düşük düzey işlevsel niteliklerini anlamamızı gerektirir. Tüm beyin emülasyonunun başarılı olması için temel bir kavramsal ya da teorik atılıma gerek yoktur.

Gelgelelim tüm beyin emülasyonu bazı ileri teknolojilerin ön açıcılığına gerek duyar. Üç tane kilit öngereklilik vardır: (1) *Tarama*: Yeterli çözünürlüğe sahip olan ilgili özellikleri yakalayabilen yüksek girdiçıktılı mikroskopi; (2) *Tercüme*: Otomatik görüntü analizi ham tarama verilerini ilintili nöral hesaplama öğelerinin yorumlanmış



ŞEKİL 4 Elektronlu mikroskop görüntülerinden üçboyutlu nöroanatomi yaratılması. Sol üstte: Bir nöronal maddenin (dendritler ve aksonların) enine kesitlerini gösteren tipik bir elektron mikroskobu görüntüsü. Sağ üstte: SBFSEM yoluyla edinilen tavşan retinasının sinir dokusunun hacim görüntülemesi.* Tekil ikiboyutlu görüntüler bir küpe doldurulmuştur (yaklaşık 11 mikrometrelik kenarı vardır). Altta: Otomatik bir segmentasyon algoritmasıyla yaratılmış bir nöropil hacmini dolduran nöron izdüşümlerinin bir altkümesinin yeniden yaratılması.†

* Helmstaedter vd. (2011).

† Andres vd. (2012).

üçboyutlu modeline dönüştürmelidir; (3) *Simülasyon*: Ortaya çıkan hesaplamalı yapıyı uygulayacak kadar güçlü donanım (bkz. **TABLO 4**).

(Daha zorlu olan bu adımlarla kıyaslandığında, temel bir sanal gerçekliğin ya da işitsel-görsel giriş kanalı ve basit bir çıkış kanalı olan robot varlığın yaratılması görece kolaydır. Basit ama asgari düzeyde yeterli bir I/O mevcut teknolojiyle bile başarılabilir durmaktadır.[11])

Gerekli kolaylaştırıcı teknolojilerin yakın gelecekte olmasa bile bir gün ulaşılabilir hale geleceğini söyleyebiliriz. Bugün farklı türde nöron ve nöronal sürecin makul hesaplamalı modelleri halihazırda mevcuttur.

TABLO 4: Tüm beyin emülasyonu için gerekli yeterlilikler

Tarama	İşleme öncesi/ sabitleme		Beyinleri uygun şekilde hazırlama, ilintili mikro yapıyı ve durumu koruyarak.
	Fiziksel idare		Sabit beyinleri ve doku parçalarını tarama öncesinde, sırasında ve sonrasında kullanma yöntemleri
	Görüntüleme	Hacim	Makul zamanda ve masrafla tüm beyin hacimlerini tarama kabiliyeti
		Çözünürlük	Yeniden inşa için gerekli çözünürlükte tarama
		İşlevsel bilgi	Dokunun işlevsel açıdan ilintili özelliklerini tespit için tarama yeteneği
Tercüme	Görüntü işleme	Geometrik ayarlama	Taramadaki kusurlardan kaynaklı bozuklukları giderme
		Veri aradeğerleme	Eksik verileri bulma
		Gürültü temizleme	Tarama kalitesini artırma
		İzleme	Yapıyı tespit edip bunu dokunun tutarlı üçboyutlu bir modeli olarak işleme
Taramayı yorumlama	Hücre türü tespiti	Hücre türlerini tespit etme	
		Sinaps tespiti	Sinapsları ve bağlantılılıklarını tespit etme
		Parametre tahmini	Hücrelerin, sinapsların ve diğer varlıkların işlevsel açıdan ilintili parametrelerini tahmin etme
		Veritabanlama	Ortaya çıkan envanterleri verimli bir tarzda depolama
	Sinir sisteminin yazılım modeli	Matematiksel model	Varlıkların ve onların davranışlarının modeli
		Verimli uygulama	Modelin uygulanması
Simülasyon	Depolama		Özgün modelin ve mevcut durumunun depolanması
	Bant genişliği		Süreçler arası verimli iletişim

CPU	Simülasyonu yapacak olan işlemci gücü
Beden simülasyonu	Robot aracılığıyla gerçek çevreyle ya da sanal çevreyle etkileşimi sağlayan beden simülasyonu
Çevre simülasyonu	Sanal beden için sanal çevre

İkiboyutlu görüntü kümeleri yoluyla aksonların ve dendritlerin izini sürebilen görüntü tanıma yazılımı geliştirilmiştir (ama güvenilirliğinin geliştirilmesi gerekiyor). Ayrıca gerekli çözünürlüğü sağlayan görüntüleme aletleri vardır – tarayıcı bir tünel açıcı mikroskopla tekil atomları gerektiğinden çok daha yüksek bir çözünürlükte "görmek" mümkündür. Gelgelelim mevcut bilgi ve yetenekler gerekli kolaylaştırıcı teknolojilerin gelişiminin önünde ilkesel bir engel olmadığını söylese de, tüm beyin emülasyonuna ulaşabilmek için artımlı teknik ilerlemenin epey büyük bir kısmından yararlanmak gerekeceği açıktır.[12] Örneğin mikroskopi teknolojisi sadece yeterli çözünürlüğe değil yeterli iş çıkarma yeteneğine de ihtiyaç duyacaktır. Gerekli yüzey alanını görüntülemek için atomik çözünürlüklü STM (taramalı tünelleme mikroskobu) kullanmak pratiğe dökülemeyecek kadar yavaş olacaktır. Düşük çözünürlüklü bir elektron mikroskobu kullanmak daha akla yatkındır, ama bu da sinaptik ince yapı gibi ilintili ayrıntıları görünür kılmak için kortikal dokuyu hazırlama ve renklendirmede yeni yöntemler gerektirir. Nöral hesaplama kütüphanelerinin müthiş büyümesi ve otomatik görüntü işleme ve tarama yorumlamada büyük ilerlemeler de gerekli olacaktır.

Yapay zekâyla kıyaslandığında, tüm beyin emülasyonu, genel itibariyle, teorik içgörüye daha az, teknolojik beceriye daha fazla ihtiyaç duyar. Tüm beyin emülasyonu için tam olarak ne kadar teknoloji gerektiği beyin emülasyonundaki soyutlama düzeyine bağlıdır. Bu bakımdan içgörü ile teknoloji arasında bir ödünleşim vardır. Genel itibariyle, tarama ekipmanımız ne kadar kötü, bilgisayarlarımız ne kadar zayıfsa, düşük düzeydeki kimyasal ve elektrofizyolojik beyin süreçlerini simüle etmeye o kadar az bel bağlayabiliriz ve ilgili işlevselliklerin daha soyut temsillerini yaratmak için emülasyona uğraştığımız hesaplamalı yapıya dair o kadar fazla teorik anlayışa ihtiyaç olur.[13] Tersinden bakıldığında, yeterince ileri tarama teknolojisi ve bolca miktarda bilgisayar gücüyle, beyne dair fevkalade sınırlı bir anlayışla bile kaba kuvvet yoluyla bir emülasyon sağlanabilir. Gerçekçi olmayan sınırlayıcı örnekte kuantum mekaniğinin Schrödinger denkleminden yararlanarak temel parçacıklarının düzeyinde bir beyin emülasyonu yapmayı umabiliriz. Ardından herhangi bir biyolojik modele değil tamamen mevcut fizik bilgisine yaslanılabilir. Gelgelelim bu uç örnek hesaplamalı güçten ve veri ediniminden hiçbir şekilde gerçekleştirilemez taleplerde bulunur. Çok daha akla yatkın bir emülasyon düzeyi, tekil nöronlar ile onların bağlantılılık matrisinin yanı

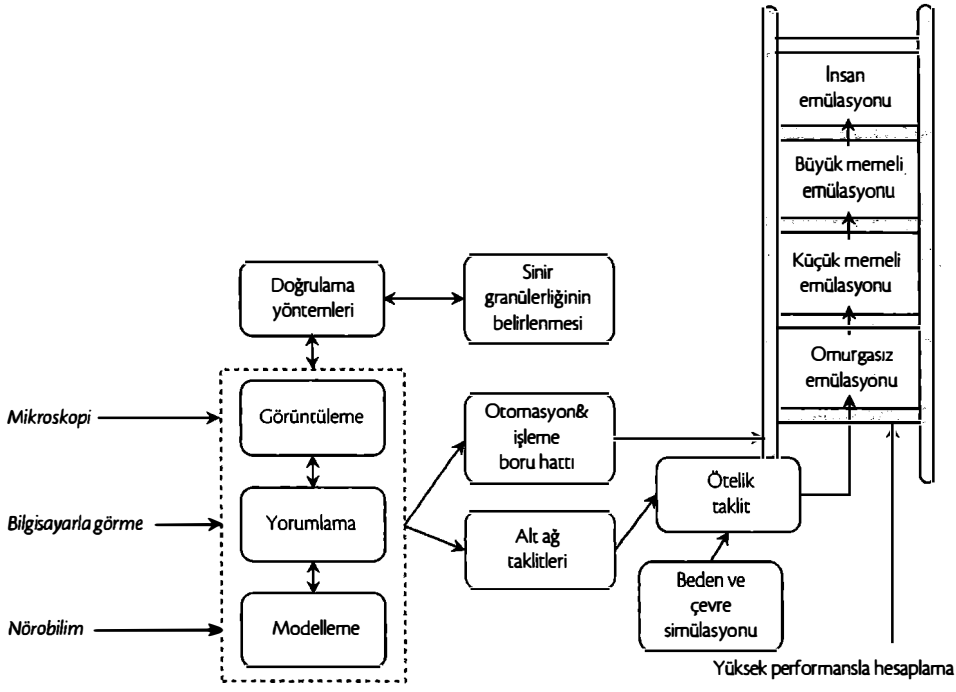
sıra bunların dendrit ağaçlarının yapılarından bazılarını ve belki de tekil sinapsların bazı durum değişkenlerini kapsayacaktır. Nörotransmiter molekülleri bireysel olarak simüle edilmeyecek ama bunların dalgalı yoğunlaşmaları kaba taneli bir tarzda model alınacaktır.

Tüm beyin emülasyonunun uygulanabilirliğini değerlendirmek için başarı kıstasını anlamak gerekir. Amaç tikel bir uyarılma dizisine maruz kalan özgün beyinde tam olarak ne olacağını tahmin etmek için kullanılabilecek kadar ayrıntılı ve eksiksiz bir beyin simülasyonu yaratmak değildir. Daha ziyade, beynin hesaplamalı açıdan işlevsel özelliklerini, ortaya çıkan emülasyonun zihinsel çalışmayı yerine getirmesini sağlayacak kadar yakalamaktır. Bu amaca istinaden gerçek bir beynin karmakarışık biyolojik ayrıntılarının çoğu ilintisizdir.

Daha incelikli bir analiz, taklit edilen beynin bilgi işleme işlevselliğinin hangi ölçüde korunduğuna bağlı olarak emülasyonun başarısının farklı düzeyleri arasında ayrım yapacaktır. Örneğin şunlar arasında ayrım yapılabilir: (1) Taklit edilen beynin tüm bilgi, yetenek, yeti ve değerlerini koruyan *yüksek sadakatli emülasyon*; (2) eğilimleri belli açılardan insandan epey farklı olan ama taklit edilen beyinle aynı zihinsel emeği büyük oranda ortaya koyabilen *çarpık emülasyon*; (3) taklit edilen yetişkin beyninin edindiği yeteneklerden ya da anılardan yoksun ama normal bir insanın öğrenebileceklerinin çoğunu öğrenme beceresine sahip, bir bakıma bebek gibi olan *türsel taklit* (bu da çarpık olabilir).[14]

Yüksek sadakatli emülasyonu üretmek nihayetinde gerçekleştirilebilir dursa da, eğer bu yoldan gidersek başaracağımız *ilk* tüm beyin emülasyonunun daha düşük seviyede olması kuvvetle muhtemeldir. Kusursuz işleyen şeyler elde etmeden önce, muhtemelen kusurlu işleyen şeyler elde edeceğiz. Ayrıca emülasyon teknolojisi yönünde bir uğraşın, bu çabalar sırasında keşfedilen bazı nöral hesaplama ilkelerini benimseyecek ve bunları sentetik yöntemlerle harmanlayacak bir tür nöromorfik yapay zekâ yaratılmasına yol açması ve bunun tamamen işlevsel bir tüm beyin emülasyonunun tamamlanmasından önce gerçekleşmesi de mümkündür. İleride göreceğimiz gibi, bu şekilde nöromorfik yapay zekâya sirayet etme olanağı emülasyon teknolojisini hızlandırmaya çalışmanın ne derece arzulanır bir şey olduğuna dair stratejik değerlendirmeyi karmaşıklaştırır.

Şu an tüm bir insan beyni emülasyonunun gerçekleştirilmesinden ne kadar uzağız? Yakın döneme ait bir değerlendirme teknik bir yol haritası sunuyor ve öngerekli yeteneklerin yüzyılın ortalarına doğru mevcut hale geleceğini ama bu arada epey bir belirsizlik yaşanacağını tahmin ediyordu.[15] **ŞEKİL 5**'te bu yol haritasının önemli kilometre taşları gösterilmiştir. Gelgelelim haritanın görünüşteki basitliği yanıltıcı olabilir ve daha gidilecek uzun bir yol olduğunu kesinlikle unutmamamız gerekiyor.



ŞEKİL 5 Tüm beyin emülasyonu yol haritası. Girdiler, faaliyetler ve kilometre taşlarının şeması.*

* Sandberg ve Bostrom (2008). Özgün raporda ek açıklamalar bulunabilir.

Şu ana kadar gerçekleştirilmiş bir beyin emülasyonu yoktur. *Caenorhabditis elegans* adındaki mütevazı örnek organizmayı ele alalım. 1 mm uzunluğunda, 302 nöronu olan şeffaf bir yuvarlak solucandır. Bu nöronların tam bağlantılılık matrisi numunelerin dilimlenmesi, elektron mikroskopisi ve elle etiketleme yoluyla harıl harıl haritasının çıkartıldığı 1980'lerin ortalarından beri bilinmektedir.[16] Ama salt hangi nöronların hangileriyle bağlantılı olduğunu bilmek yeterli değildir. Bir beyin emülasyonu yaratmak için hangi sinapsların uyarıcı hangilerinin ketleyici olduğunu, bağlantıların kuvvetini ve aksonların, sinapsların ve dendrit ağaçlarının çeşitli dinamik özelliklerini bilmek gerekir. Bu bilgi henüz *C. Elegans*'ın küçük sinir sisteminde bile mevcut değildir (gerçi artık hedeflenmiş ılımlı büyüklükte bir araştırma projesinin menzili dahilinde olabilir).[17] *C. Elegans*'inki gibi küçük bir beynin emülasyonundaki başarı bize büyük beyinlerin emülasyonu için neler gerekeceği konusunda daha iyi bir görüş sunacaktır.

Teknoloji geliştirme sürecinin bir noktasında, beyin dokusunun küçük niceliklerinin otomatik emülasyonu için teknikler ortaya çıktığında, sorun artık ölçekleme sorunu haline gelir. **ŞEKİL 5**'in sağ kısmındaki "merdiven"e dikkat edin. Bu yükselen kutular ilk baştaki engeller aşıldıktan sonra başlayabilecek olan ilerlemelerin nihai silsilesini temsil eder. Bu silsiledeki aşamalar her seferinde nörolojik açıdan daha karmaşık model orga-

nizmaların tüm beyin emülasyonlarına tekabül eder: Örneğin *C. Elegans* –> bal arısı –> fare –> al yanaklı şebek –> insan. Bu halkalar arasındaki boşluklar –en azından ilk adımdan sonra– doğası bakımından çoğunlukla nicel olduğu ve tamamen olmasa da esasen taklit edilen beyinlerin büyüklüğündeki farklardan kaynaklandığı için, tarama ve simülasyon kapasitesinin görece düz bir ölçek büyümesi yoluyla izlenebilir olmalıdır.[18]

Bu son basamağı bir kez tırmanmaya başladığımızda, insanın tüm beyin emülasyonunun nihai olarak elde edilmesi çok daha açık bir şekilde öngörülür hale gelir.[19] Böylece tüm beyin emülasyonu yolundan geçerek insan düzeyinde makine zekâya gelmeden önce peşinden bir uyarı almayı umabiliriz; en azından yeterli olgunluğa erişmek için gerekli kolaylaştırıcı teknolojiler arasında sonuncusu yüksek verimli tarama ya da gerçek zaman simülasyonu için gerekli hesaplamalı güç varsa. Ne var ki eğer son kolaylaştırıcı teknoloji nöral hesaplama modellemesiyse, o zaman etkileyici olmayan prototiplerden işler bir insan emülasyonuna geçiş daha keskin olabilir. Bol miktarda tarama verisine ve hızlı bilgisayara karşın nöron modellerimizi düzgün çalıştırmanın zor olduğu bir senaryo tahayyül edebiliriz. Nihayet son kusur da giderildiğinde, öncesinde tamamen işlemez olan sistem (belki de büyük bir tür epilepsi nöbetine yakalanmış bilinçsiz bir beyne benzetilebilir) bütünlüklü bir uyanık duruma dönüştürülebilir. Bu örnekte kilit ilerlemenin habercisi, artan büyüklükte işler hayvan emülasyonları olmayacaktır (bir o kadar büyük fontlarla atılan gazete manşetlerine aldanmamalı). Dikkatini verenler için bile belli bir noktada nöral hesaplama modellerinde kaç tane kusur kaldığını ve kritik atılım arifesinde bile bunları düzeltmenin ne kadar zaman alacağını önceden söylemek zor olabilir. (Tüm insan beyin emülasyonu bir kez başarıldığında, potansiyel açıdan patlayıcı başka gelişmeler gerçekleşecektir; ama bu değerlendirmeyi 4. Bölüm'e kadar erteleyeceğiz.)

Dolayısıyla ilgili tüm araştırmalar açıktan yürütülse bile tüm beyin emülasyonu için sürpriz senaryolar tahayyül edilebilir. Gelgelelim yapay zekâdan makine zekâsına giden yolla kıyaslandığında tüm beyin emülasyonunun daha fazla açıktan işaretlerle yol alması muhtemeldir, zira somut gözlemlenebilir teknolojilere daha fazla yaslanır ve tümüyle teorik içgörüye dayalı değildir. Ayrıca yapay zekâ yolu için olduğundan daha bir güvenle söyleyebiliriz ki, emülasyon yolu yakın gelecekte (sözgelimi önümüzdeki on beş yılda) başarıya ulaşmayacaktır, zira öncesinde gerekli olan birkaç teknoloji henüz geliştirilmemiştir. Halbuki birinin çıkıp sıradan bir kişisel bilgisayarda bir çekirdek yapay zekâ kodlaması ilkesel olarak konuşursak muhtemel görünüyor ve birilerinin bir yerlerde yakın gelecekte bunun nasıl yapılacağına dair doğru içgörüye ulaşması da –pek muhtemel olmasa da– tasavvur edilebilir.

Biyolojik biliş

Mevcut insan zekâsından daha büyük bir zekâya giden üçüncü bir yol ise biyolojik beyinlerin çalışmasını iyileştirmektir. İlkesel olarak bakıldığında, bunu teknoloji

olmadan, seçici yetiştirmeyle sağlamak mümkündür. Klasik bir büyük ölçekli öjeni programı başlatma girişimiyse siyasi ve ahlaki açıdan büyük engellerle karşılaşacaktır. Ayrıca seçilim son derece güçlü değilse, ciddi sonuçlar elde etmek için birçok kuşak gerekecektir. Bu tür bir inisiyatif meyve vermeden önce, biyolojideki ilerlemeler insan genetiğinin ve nörobiyolojisinin çok daha doğrudan kontrol edilmesine izin verecek ve her türlü insan yetiştirme programını güdük bırakacaktır. Dolayısıyla daha hızlı bir şekilde, zaman ölçeği bakımından birkaç kuşak ya da daha az sürede sonuç verme potansiyeli taşıyan yöntemlere odaklanacağız.

Bizim bireysel bilişsel kapasitelerimiz eğitim-öğretim gibi geleneksel yöntemler de dahil olmak üzere çeşitli şekillerde güçlendirilebilir. Nörolojik gelişim anne ve bebek beslenmesini optimum düzeye çekmek, kurşun ve çevreyi kirleten diğer nörotoksik maddeleri temizlemek, asalakları yok etmek, yeterli uyku ve egzersiz ile beyni etkileyen hastalıkların önüne geçmek gibi düşük teknolojili müdahalelerle mümkün hale gelebilir.[20] Biliş yetisindeki ilerlemeler kuşkusuz bu araçlardan her biri yoluyla sağlanabilir, ama kazanımın büyüklükleri özellikle de zaten epey iyi beslenen ve eğitim alan toplumlarda muhtemelen mütevazı boyutlarda olacaktır. Bu araçlardan herhangi biriyle süper zekâya erişemeyeceğimiz kesindir, ama özellikle de yoksunların seviyesini yukarı çekerek ve küresel yetenek havuzunu büyüterek kıyıda köşede kalmış olanlara yardımcı olabilir. (İyot eksikliğinden ötürü ömür boyu zekâ bunalımı dünyanın birçok yoksul bölgesinde hâlâ çok yaygındır; kişi başına ve yılda birkaç sent tutarında sofra tuzuyla bunun önlenebileceği düşünüldüğünde durumun vahameti anlaşılır.[21])

Biyotıbbi ilerlemeler daha da büyük itkiler sağlayabilir. Hafızayı, konsantrasyonu ve zihinsel enerjiyi en azından bazı konularda ilerlettiği iddia edilen ilaçlar halihazırda mevcuttur.[22] (Bu kitap üzerine çalışırken sağladığı yardımlardan ötürü kahve ve nikotinli sakıza teşekkürü bir borç bilirim.) Akıllı ilaçların mevcut kuşağının verimliliği değişken, marjinal ve genellikle şaibeliyken, geleceğin nootropikleri daha açık faydalar sağlayıp, çok daha az yan etki sunabilir.[23] Gelgelelim gerek nörolojik gerek evrimsel temellerde sağlıklı bir insanın beynine bir kimyasal vererek çarpıcı bir zekâ artışını tetiklemek hiç de inandırıcı gelmiyor.[24] Bir insan beyninin bilişsel işleyişi birçok etkenin –özellikle de embriyo gelişimindeki kritik aşamalar sırasında– incelikli bir eşgüdüme sahip olmasına bağlıdır ve bu kendi kendini örgütleyici yapının ilerlemesi için sadece dışarıdan bir iksir dayayarak işi halletmekten ziyade dikkatle dengelenmesi, ayarının yapılması ve inceltilmesi gerektiğini söylemek sanırım çok daha doğrudur.

Genetikle oynamak da psikofarmakolojiden çok daha güçlü araçlar sağlayacaktır. Bir kez daha genetik seçilim fikrini düşünün: Çiftleştirme örüntülerini kontrol ederek bir öjeni programı uygulamaya çalışmak yerine, embriyolar ya da gamet hücreleri düzeyinde seçilimden yararlanabilirler.[25] İmplantasyon öncesi genetik tanı Huntington hastalığı gibi tek genli hastalıklar ve meme kanseri gibi geç başlangıçlı hastalıklara

yatkınlık için üretilen embriyoları görüntülemek için laboratuvar ortamında dölleme işlemlerinde şimdiden kullanılmaktadır. Ayrıca cinsiyet seçilimi ve insan lökosit antijen türünü hasta bir çocuğunkiyle eşleştirmek için de kullanılmaktadır ki, bu sayede bebek doğduğunda kordon kanı kök hücre bağışından yarar sağlayabilmektedir.[26] Seçilecek ya da seçilmeyecek özelliklerin kapsamı önümüzdeki on yirmi yılda epey genişleyecektir. Davranışsal genetikteki güçlü bir ilerleme dürtüsü genotiplenme ve gen diziliminin maliyetlerinin hızla düşüyor olmasıdır. Çok sayıda denekten yararlanılarak yapılan genom çapında karmaşık özellik analizi daha yeni yeni uygulanabilir hale geliyor ve insanın bilişsel ve davranışsal özelliklerinin genetik mimarilerine dair bilgimizi önemli ölçüde artıracaktır.[27] Bu durumda, göz ardı edilebilir kalıtsallığı olmayan bir özellik (bilişsel kapasite dahil) seçilime açık hale gelebilir.[28] Embriyo seçilimi genlerin çevreyle karmaşık etkileşim halinde fenotipler üretebilmelerini sağlayan nedensel yolakların derinlemesine anlaşılmasını gerektirmez: Sadece ilgili özelliklerin genetik bağıntılarına ilişkin (bir sürü) veri gerektirir.

Farklı seçilim senaryolarında elde edilebilir olan kazanımların büyüklüğüne dair bazı kaba tahminlerde bulunmak mümkündür.[29] **TABLO 5** dar anlamda zekâ kalıtsallığının altında yatan ortak eklemeli genetik çeşitlilikleri hakkında tam bilgiye sahip olunduğunda çeşitli seçilim miktarlarından kaynaklanan zekâ artışı beklentilerini gösteriyor. (Kısmi bilgi olduğundaysa, seçilimin verimliliği azalacak, ama naif bir şekilde beklediğimiz boyutta olmayacaktır.[30]) Bekleneceği üzere, daha çok sayıda embriyo arasında seçim yapmak daha büyük kazanımlar getirir, ama ciddi şekilde azalan geri dönüşler vardır: 100 embriyo arasında seçilim, 2 embriyo arasında seçilimle elde edilecek olandan elli kat daha fazla kazanım sağlamaz.[31]

TABLO 5 Bir grup embriyo arasında seçilimden azami IQ kazanımı[32]

Seçilim	Kazanılan IQ puanı
2'de 1	4,2
10'da 1	11,5
100'de 1	18,8
1000'de 1	24,3
5 kuşak 10'da 1	< 65 (b/c azalan kazanım)
10 kuşak 10'da bir	< 130 (b/c azalan kazanım)
Toplam sınırlar (biliş için optimize edilmiş ilave değişkenler)	100+ (<300 (b/c azalan kazanım))

İşin ilginç yanı, geri dönüşlerin azalması seçilim birden çok kuşağa yayıldığında epey düşer. Mesela on kuşak boyunca en üst 10'dan 1'i tekrar tekrar seçmek (her kuşağın

önceki kuşaktan seçilenlerin dölü olduğunu varsayarsak) özellik değeri bakımından 100'de 1'den tek seferlik seçilime kıyasla çok daha büyük artış sağlayacaktır. Dizilimli seçilimdeki sorun, elbette, daha uzun sürmesidir. Eğer her kuşakta atılan adım yirmi otuz yıl sürerse, o zaman art arda beş kuşak bile bizi 22. yüzyıla götürebilir. O zamana kalmadan, genetik mühendisliğinin daha doğrudan ve güçlü tarzları (makine zekâsından bahsetmiyorum bile) büyük ihtimalle mevcut hale gelecektir.

Gelgelelim insanlarda kullanım için geliştirildiğinde uygulama öncesi genetik taramanın geliştirme gücünü epey bir artıracak olan tamamlayıcı bir teknoloji var: Embriyon halindeki kök hücrelerden yaşayabilir sperm ve yumurtalar almak.[33] Bu teknik insanlardaki gamet türü hücrelerde ve farelerde doğurgan döl üretmek için halihazırda kullanılmaktadır. Ne var ki hayvanlardaki sonuçları insanlara nakletme ve türetilen kök hücrelerdeki epigenetik anormalliklerden kaçınma konusunda ciddi bilimsel zorluklar vardır. Bir uzmana göre, bu zorluklar insana uygulanmasını "gelecekteki 10, hatta 50 yıla atabilir."[34]

Kök hücreden türetilmiş gametlerde bir çift için mevcut olan seçilim gücü miktarı ciddi oranda artırılabilir. Günümüz pratiğinde laboratuvarda dölleme işleminde genellikle ondan daha az embriyo yaratılıyor. Kök hücreden türetilen gametlerdeyse birkaç hücre bağışı neredeyse sınırsız sayıda gamete dönüştürülebilir ve bunlar da birlikte embriyo üretebilir ve ardından bunların genotipi ya da dizilimi yapılır ve içlerinde en umut vaat edeni uygulama için seçilir. Tek tek her bir embriyoyu hazırlama ve tarama masrafını göze almak zorunda kalan bu teknoloji laboratuvarda dölleme kullanan çiftler için mevcut olan seçici güçte birkaç kat artış sunabilir.

Ama daha da önemlisi, kök hücreden türetilen gametler *yinelemeli embriyo seçilimini* kolaylaştırarak birkaç kuşak boyunca seçilimin bir insanın olgunlaşma döneminden daha kısa süreye sıkıştırılmasına imkân sağlayacaktır. Bu işlem şu adımlardan oluşacaktır:[35]

1. Arzulanan genetik özelliklerde daha yüksek olan embriyoları genotiplemek ve seçmek.
2. Bu embriyolardan kök hücreler alıp bunları sperm ve yumurtaya dönüştürüp altı ay ya da daha kısa bir süre içinde olgunlaştırmak.[36]
3. Yeni sperm ve yumurtayı embriyolar üretecek şekilde çaprazlamak.
4. Büyük genetik değişiklikler birikene kadar tekrarlamak.

Bu yolla, sadece birkaç yılda on ya da daha fazla kuşakta seçilim gerçekleştirmek mümkün hale gelecektir. (Bu işlem zaman alır ve masraflıdır; ama ilkesel açıdan, her doğumda tekrarlanacağına sadece bir kez yapılması gerekecektir. İşlemin sonunda elde edilen hücre hatları çok sayıda geliştirilmiş embriyo yaratmak için kullanılabilir.)

TABLO 5'te görüldüğü üzere, bu şekilde yaratılan bireyler arasında *ortalama* zekâ düzeyi çok yüksek olabilir, muhtemelen insanlık tarihindeki en zeki bireyin zekâsına eşit, hatta ondan da yüksek olur. Bu tür bireylerin sayıca çok olduğu bir dünya nüfusu kolektif bir süper zekâ oluşturabilir (tabii eğer buna denk bir kültür, eğitim, iletişim altyapısı vb. de olursa).

Bu teknolojinin etkisi birkaç etkenle zayıflatılıp geliştirilecektir. Seçilmiş embriyolar büyüyüp yetişkin insan haline gelirken olgunlaşmada kaçınılmaz bir gecikme söz konusudur: Geliştirilmiş bir çocuk tam üretkenliğe ulaşana kadar en az yirmi yıl, bu çocuklar işgücünün önemli bir parçasını oluşturacak düzeye gelene kadar daha da çok. Ayrıca teknoloji kusursuz hale getirildikten sonra bile evlatlık edinme oranları ilk başta muhtemelen düşük olacaktır. Bazı ülkeler dini ya da ahlaki gerekçelerle bunu tamamen yasaklayabilir bile.[37] Seçilime izin verildiğinde bile birçok çift doğal yollarla çocuk sahibi olmak isteyecektir. Gelgelelim IVF kullanımının bariz yararları olduğu gösterilirse, buna olan istek artacaktır: Mesela çocuğun çok yetenekli olacağına ve hastalıklara genetik yatkınlığı olmayacağına dair garanti verilirse. Sağlık harcamalarının düşük, ortalama yaşam sürelerinin yüksek olması gibi etkenler de genetik seçilimin hanesine artı olarak yazılacaktır. Bu işlemden yararlanma özellikle de toplumun elitleri arasında yaygınlaştıkça, seçilimden yararlanmayı sorumlu aydınlaşmış çiftlerin yaptığı şey olarak sunan ebeveynlik normlarına doğru bir kültürel kayış olabilir. İlk başta gönülsüz olan çiftlerden birçoğu arkadaşlarının ve meslektaşlarının geliştirilmiş çocuklarına kıyasla yarışa geride başlayan bir çocuğa sahip olmamak için dalgaya katılabilir. Bazı devletler ülkenin insan sermayesini geliştirmek ya da yönetici güruhun dışındaki bireylerde itaatkârlık, uysallık, söz dinleme, umursamazlık, riskten kaçınma ya da korkaklık gibi özellikleri seçerek uzun vadeli toplumsal istikrarı artırmak için yurttaşlarını genetik seçilimden yararlanmaya teşvik edebilir.

Entelektüel kapasiteye etkiler de bilişsel özellikleri geliştirmek için mevcut seçilim gücünün ne ölçüde kullanılacağına bağlı olacaktır (**TABLO 6**). Bir embriyo seçilim biçiminden yararlanmayı tercih edenler ellerinin altındaki seçilim gücünü nasıl kullanacaklarını seçmek zorunda kalacaklar ve zekâ belli oranda diğer özelliklerle (sağlık, güzellik, kişilik ya da atletizm gibi) rekabet halinde olacaktır. Yinelenmiş embriyo seçilimi, böyle büyük miktarlarda seçilim gücü sunarak bu ödünleşimlerden bazılarını azaltacak ve birden çok özellik için aynı anda güçlü seçilimi kolaylaştıracaktır. Gelgelelim bu işlem ebeveynler ile çocuk arasındaki normal genetik ilişkiyi bozabilir ki, bu da birçok kültürde talebi olumsuz şekilde etkileyebilir.[38]

Genetik teknolojideki yeni ilerlemelerle birlikte, genomları özgülleme için sentezlemek ve böylece büyük embriyo havuzlarına olan ihtiyaçtan kurtulmak mümkün hale gelebilir. DNA sentezi halihazırda rutin ve büyük oranda otomatikleşmiş bir biyoteknolojidir, ama bir üreme bağlamında kullanılabilecek tüm bir insan genomunu

sentezlemek henüz gerçekleştirilebilir değildir (en önemli nedeni de epigenetik hakkı elde etmenin önündeki engellerin hâlâ çözülmemiş olmasıdır).[39] Ama bu teknoloji bir kez olgunlaşınca, bir embriyo her bir ebeveynden genetik girdilerin tercih edilen bir bileşimiyle tasarlanabilir. Nüfusta düşük yoğunlukta bulunan ama biliş yetisine ciddi anlamda olumlu etkilerde bulunabilecek olan aleller de dahil olmak üzere ebeveynlerde olmayan genler de eklenebilir.[40]

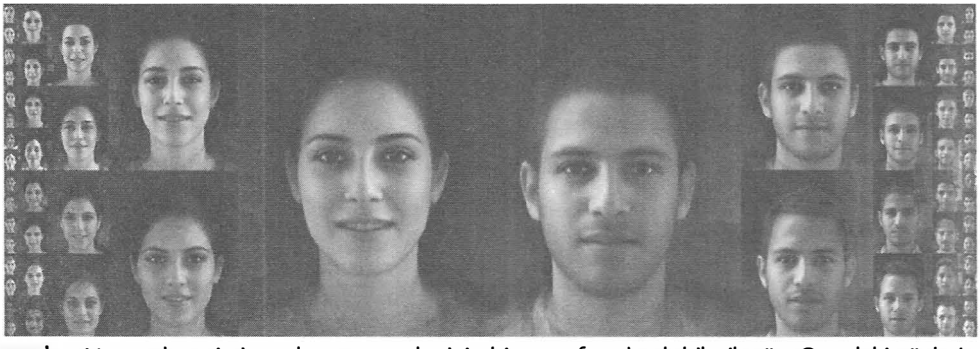
TABLO 6 Farklı senaryolarda genetik seçilimin mümkün olan etkileri[41]

Evlat edinme/ teknoloji	"IVF+" 2 embriyodan 1'inin seçilimi [4 puan]	"Agresif IVF" 10 embriyodan 1'inin seçilimi [12 puan]	"Laboratuvarda dölleme" 100 embriyodan 1'inin seçilimi [19 puan]	"Yinelenmiş embriyo seçilimi" [100+ puan]
"Marjinal fertilite pratiği" ~ %0,25 evlat edinme	Bir kuşak boyunca toplumsal olarak göz ardı edilebilir. Toplumsal ihtilafın etkileri doğrudan etkilerinden daha önemlidir.	Geliştirilmiş tesadüfi öğe bilişsel açıdan yüksek seçilimli konumlarda göze çarpan bir azınlık oluşturur.	Seçilmişler elit bilimcilerin, avukatların, hekimlerin, mühendislerin saflarını doldurmuştur. Entelektüel Rönesans?	
"Elit üstünlüğü" %10 evlat edinme	İlk kuşakta bilişsel etki azdır, bir azınlığa bariz şekilde üstünlük sağlayacak şekilde bilişsel olmayan özelliklerin seçilimiyle birleştirilir.	Harvard mezunlarının önemli bir kesimi geliştirilmiştir. İkinci kuşak bilişsel açıdan yoğun işlerde ağırlıklı yere sahiptir.	Seçilmişler bilimcilerin, avukatların, hekimlerin, mühendislerin saflarında ilk kuşak boyunca baskındır.	"İnsanlık sonrası"[42]
"Yeni normal" > %90 evlatlık edinme	Öğrenme bozukluğu çocuklarda çok daha seyrek görülür. İkinci kuşakta yüksek IQ eşiğinin üstündeki nüfus en az iki kat artmıştır.	Eğitim ve gelir düzeyinde ciddi ilerleme. İkinci kuşak aynı çizgide katbekat ilerleme sağlayabilir.	Önde gelen bilimcilerde görülen ham IQ düzeyi ilk kuşakta +10 kat daha yaygındır. İkinci kuşaktaysa binlerce kat.	"İnsanlık sonrası"

İnsan genomları sentezlenebilir hale geldiğinde mümkün hale gelecek müdahalelerden biri bir embriyonun genetik "yazım denetimi"nin yapılmasıdır. (Yinelenmiş embriyo seçilimi buna yaklaşılmasını da sağlayabilir.) Şu an her birimiz, belki de çeşitli

hücresel süreçlerin verimliliğini azaltan yüzlerce mutasyon geçirerek bir mutasyon yükü taşıyoruz.[43] Her bireysel mutasyonun neredeyse göz ardı edilebilir bir etkisi vardır (ancak yavaş yavaş gen havuzundan çıkarılabilir), ama bu tür mutasyonlarla birlikte işleyişimize ağır bir yük bindirebilir.[44] Bireysel zekâ farklılıklarını, büyük oranda, her birimizin taşıdığı bu tür hafif zararlı alellerin sayısındaki ve doğasındaki çeşitliliklere yorabiliriz.

Gen senteziyle birlikte bir embriyonun genomunu alıp bu genomun bir versiyonunu birikmiş mutasyonların genetik gürültüsünden azade şekilde yaratabiliriz. Kışkırtıcı şekilde konuşmak gerekirse, bu şekilde yazım hataları düzeltilmiş genomlardan yaratılan bireyler şu an yaşayan herkesten "daha fazla insan" olabilir, zira insan biçiminin daha az çarpık ifadeleri olacaklardır. Bu tür insanlar birbirinin aynısı olmayacaktır, zira insanlar farklı zararlı mutasyonlar taşımaktan çok daha başka şekillerde genetik çeşitlilikler sergilerler. Ama yazım hataları düzeltilmiş bir genomun fenotipik tezahürü fiziksel ve zihinsel yapısı bakımından bir istisna olabilir; zekâ, sağlık, metanet ve görünüş gibi çokgenli özellik boyutlarında yüksek seviyelere çıkabilir.[45] (Uzak bir benzetme olsa da, bileşik yüzlerle benzerlik kurulabilir: Birleştirilmiş bireylerin yüzlerinde ortalamanın altında kalan özellikler dışarıda bırakılır: Bkz. **ŞEKİL 6.**)



ŞEKİL 6 Yazım denetimi yapılmış genomlar için bir metafor olarak bileşik yüz. Ortadaki yüzlerin ikisi de (hepsi de Tel Avivli) on altı farklı bireyin fotoğraflarının üst üste bindirilmesiyle yaratılmıştır. Bileşik yüzlerin genellikle onları oluşturan tekil yüzlerin her birinden daha güzel olduğu düşünülür, zira kişiye özgü kusurlar dışarıda bırakılır. Benzerlik kurmak gerekirse, bireysel mutasyonları ortadan kaldıran yazım denetimi yapılmış genomlar "Platon'un ideal" insanına yakın insanlar yaratabilir. Bu tür bireyler hiç de genetik açıdan benzeş olmayacaktır, çünkü birçok gen çoklu eşit derecede işlevsel aleller halinde gelir. Yazım denetimi yapmak sadece zararlı mutasyonlardan doğan çeşitliliği yok edecektir.*

* Bu bileşikler Virtual Flavius'ın çektiği tekil fotoğraflardan Mike Mike tarafından hazırlanmıştır (Mike 2013).

Başka potansiyel biyoteknolojik teknikler de konuyla alakalı olabilir. İnsanda reprodüktif klonlanma bir kez başarıldığında istisnai derecede yetenekli bireylerin genomunu çoğaltmak için kullanılabilir. Müstakbel ebeveynlerin çoğunun çocuklarıyla biyolojik bağa sahip olma tercihlerinden ötürü bunu seçenlerin sayısı sınırlı olacaktır,

ama bu uygulama yine de göz ardı edilemeyecek etkilere sahip olabilir, çünkü: (1) Olağanüstü yetenekli insanların sayısındaki görece küçük bir artış bile kayda değer bir etki yaratabilir ve (2) bir devlet çıkıp belki taşıyıcı annelere para vererek büyük çaplı bir öjeni programına girişebilir. Keza diğer genetik mühendisliği türleri (yeni sentetik genlerin tasarlanması ya da genoma başlatıcı bölgelerin ve diğer öğelerin yerleştirilmesiyle gen ifadesinin kontrol edilmesi gibi) zaman içinde önemli hale gelebilir. Daha egzotik olasılıklar da mümkündür: Karmaşık derecede yapılandırılmış safkan kortikal dokuyla dolu vatlar ya da "geliştirilmiş" genetik yapısı değiştirilmiş hayvanlar (belki de balina ya da fil gibi büyük beyinli bir hayvan insan genleriyle zenginleştirilir). Bu son örnekler tamamen spekülatiftir, ama uzun vadede sanırım tümüyle yok sayılamaz.

Buraya kadar, germ hattı müdahalelerini, yani gametlerde ya da embriyolarda gerçekleştirilecek müdahaleleri ele aldık. Somatik gen geliştirmeleri, kuşak döngüsünü aşarak, ilkesel açıdan, daha çabuk sonuçlar üretebilir. Gelgelelim bunlar teknolojik açıdan çok daha fazla zorluk içerir. Değiştirilmiş genlerin canlı bedendeki çok sayıda hücreye yerleştirilmesini gerektirir ki, bilişsel geliştirme söz konusu olduğunda beyin de buna dahildir. Halbuki mevcut yumurta hücreleri ya da embriyolar arasından seçmek, gen yerleştirme gerektirmez. Genomun değiştirilmesini gerektiren türde gen hattı terapileri bile (genomun yazım yanlışlarının düzeltilmesi ya da nadir alellerde zincirleme) gamet ya da embriyo aşamasında çok daha kolay uygulanabilir, zira daha az sayıda hücreyle uğraşılmaktadır. Ayrıca embriyolara germ hattı müdahaleleri yetişkinlere somatik müdahalelerden çok daha büyük sonuçlar doğurabilir, çünkü ilkinde erken dönem beyin gelişimi şekillendirilebilecekken, ikincisinde mevcut yapının düzeltilmesi söz konusudur. (Somatik gen terapisi yoluyla yapılabileceklerin bazıları farmakolojik araçlarla da yapılabilir.)

Bu nedenle germ hattı müdahalelerine odaklanırken kuşaklar arası gecikmenin dünya üzerindeki büyük etkilerini de dikkate almak zorundayız.[46] Teknoloji bugün kusursuz hale getirilse ve hemen kullanılmaya başlasa bile, genetik açıdan geliştirilmiş yavrunun olgunluğa ulaşması en az yirmi yıl alır. Ayrıca insana uygulamalarda laboratuvarda kavramın kanıtlanması ile klinik uygulama arasında normalde en az on yıl gecikme vardır, zira güvenliğin sağlanması için kapsamlı çalışmalara gerek vardır. Gelgelelim genetik seçilimin en basit biçimleri bu tür bir test yapılmasını büyük oranda gereksiz kılabilir, zira standart fertilite tedavisi tekniklerinden ve genetik bilgiden yararlanarak diğer türlü tesadüf eseri denilebilecek olan embriyolar arasında seçim yapılabilecektir.

Gecikmelerin bir nedeni de başarısızlık korkusundan (güvenlik testleri talebi) değil, başarı korkusundan kaynaklı engellerde olabilir: genetik seçilimin ahlaki uygunluğu ya da genel toplumsal içerimlerinden kaynaklı düzenleme talebi. Bu tür kaygıların farklı kültürel, tarihsel ve dini bağlamlardan ötürü bazı ülkeleri daha fazla etkilemesi muhtemeldir. Örneğin savaş sonrası Almanya uzaktan da olsa geliştirmeyi amaçladığı

söylenebilecek her türlü üreme pratiğinden uzak durmuştur, zira ülkenin tarihinde öjeni hareketiyle bağlantılı karanlık bir dönem vardır. Diğer Batılı ülkelerin daha liberal bir yaklaşım sergilemeleri muhtemeldir. Bazı ülkeler ise (belki Çin ya da Singapur gibi uzun vadeli nüfus politikaları olan ülkeler) teknoloji mevcut hale geldiğinde nüfusların zekâsını artırmak için genetik seçilim ve genetik mühendisliğine izin vermekle kalmayıp aktif bir şekilde teşvikte de bulunabilir.

Örnek bir kez oturtulduğunda ve sonuçlar gelmeye başladığında, uzak duranların vagona atlamak için daha güçlü saikleri olacaktır. Her ülke bilişsel açıdan geri kalma ve ekonomik, bilimsel, askeri ve itibari meselelerde yeni insan geliştirme teknolojilerini benimseyen rakiplerine karşı yarışı kaybetme riskiyle karşı karşıya gelecektir. Bir toplumdaki bireyler elit okulların kontenjanlarının genetik açıdan seçilmiş çocuklarla doldurulduğunu görecek (bu çocuklar ortalama bakımından daha güzel, sağlıklı ve daha vicdanlı olabilirler) ve kendi evlatlarının da aynı üstünlüklere sahip olmasını isteyecekler. Teknolojinin işe yaradığı ve ciddi getiriler sağladığı bir kez kanıtlanınca, görece kısa zamanda, belki on yıl gibi kısa bir sürede büyük bir tutum değişikliği yaşanması muhtemeldir. ABD'deki kamuoyu yoklamaları laboratuvar ortamında döllemeye halkın yaklaşımının, 1978'de ilk "test tüp bebek" Louise Brown'ın doğumunun ardından olumlu anlamda ciddi bir değişiklik yaşadığını gösteriyor. Birkaç yıl öncesinde Amerikalıların sadece yüzde 18'i infertilite tedavisi için IVF'ye başvuracağını söylerken; Louise Brown'ın doğumundan kısa süre sonra yapılan bir ankette bu oran yüzde 53'e çıkmıştı ve halen de yükselmektedir.[47] (Karşılaştırma olması adına, 2004 yılında yapılan bir ankette, Amerikalıların yüzde 28'i "güç ya da zekâ" için embriyo seçilimini onaylarken, yüzde 58'i yetişkinlikte kansere yakalanmayı engellemek için, yüzde 68'i de çocuklukta ölümcül hastalıklardan kaçınmak için onaylıyordu.[48])

Eğer çeşitli gecikmeleri toplarsak (sözgelimi bir grup IVF embriyosu içinde kayda değer şekilde etkili bir seçilim için gerekli bilgiyi elde etmek için beş on yıl, insanın üremesinde kullanmak için kök hücreden türetilmiş gametlerin mevcut hale gelmesiyse çok daha uzun süre alacaktır), kayda değer bir atılım için on yıl, geliştirilmiş bir kuşağın üretken hale geldiği yaşa ulaşmak için de yirmi yirmi beş yıl dersek; germ hattı geliştirmelerinin bu yüzyılın ortalarına gelmeden önce topluma ciddi bir etkide bulunması muhtemel görünmüyor. Ne var ki bu tarihten itibaren yetişkin nüfusunun önemli bir kısmının zekâsı genetik iyileştirmelerle sıçrama kaydedecektir. Yükseliş hızı o zaman daha da artacaktır, zira daha güçlü yeni kuşak genetik teknolojileri kullanarak elde edilen çocuklar (özellikle de kök hücreden türetilmiş gametler ve yinelemeli embriyo seçilimi) işgücüne katılmış olacaktır.

Yukarıda bahsedilen genetik teknolojilerinin tam olarak gelişmesiyle birlikte (kültürü yapılmış sinir dokusunda zekâ gibi daha egzotik olanakları bir kenara bırakalım), yeni bireylerin ortalamaya vurulduğunda bu zamana kadar yaşamış herhangi bir insandan

daha akıllı olması ve zirvelerin de her geçen gün daha da yükselmesi mümkündür. Dolayısıyla biyolojik geliştirmenin potansiyeli nihayetinde yüksektir, en azından süper zekânın zayıf biçimlerinin elde edilmesi için yeterlidir diyebiliriz. Buna şaşırmamak gerekir. Nihayetinde, körü körüne evrimsel süreçler yakın akrabalarımız kuyruksuz maymunlarla ve kendi insansı atalarımızla kıyaslandığında insan soyunun zekâsını ciddi ölçüde ilerletmiştir ve *Homo sapiens*'in bir biyolojik sistemde elde edilebilecek bilişsel etkililiğin zirvesine ulaşacağını düşünmemek için hiçbir neden yoktur. Bizim mümkün olan en akıllı biyolojik tür olmak şöyle dursun, belki de bir teknolojik medeniyeti başlatma gücüne sahip olup olabilecek en aptal biyolojik tür olarak nitelendirilmemiz herhalde daha doğru olur: Biz bu boşluğu doldurduysak optimum düzeyde uygun olduğumuz için değil, oraya ilk geldiğimiz için doldurduk.

Biyolojik yolda ilerlemenin gerçekleştirilebilir olduğu açıktır. Germ hattı müdahalelerindeki kuşaklararası gecikme ilerlemenin makine zekâyı içeren senaryolardaki gibi ani ve keskin olmayabileceği anlamına gelir. (Somatik gen terapileri ve farmakolojik müdahaleler teorik açıdan kuşak gecikmesini aşabilir, ama bunun kusursuzlaştırılması zordur ve çarpıcı sonuçlar verme ihtimali düşüktür.) Makine zekânın nihai potansiyeli elbette organik zekânınkinden çok daha fazladır. (Bu boşluğun ne kadar büyük olduğunu anlamak için elektronik bileşenler ile sinir hücreleri arasındaki hız farkına bakılabilir: Bugünün transistörleri bile biyolojik nöronlarınkinden on milyon kat daha kısa zaman ölçeğinde faaliyet gösteriyorlar.) Gelgelelim biyolojik biliş yetisindeki görece mütevazı ilerlemeler bile önemli sonuçlar doğurabilir. Özellikle de bilişsel ilerleme bilim ve teknolojiyi hızlandırabilir ki, buna biyolojik zekâ yükseltmenin ve makine zekânın daha güçlü biçimlerine doğru ilerleme de dahildir. Ortalama bir insanın entelektüel olarak Alan Turing'e ya da John von Neumann'a denk olduğu ve milyonlarca insanın geçmişin entelektüel devlerinin hepsinden çok daha zeki olduğu bir dünyada yapay zekâ alanındaki ilerleme hızını bir düşünün.[49]

Bilişsel geliştirmenin stratejik içerimlerine dair bir değerlendirme için birkaç bölüm daha beklememiz gerekiyor. Fakat bu kısmı vardığımız üç sonuçla özetleyebiliriz: (1) Süper zekânın en azından zayıf biçimleri biyoteknolojik ilerlemeler yoluyla gerçekleştirilebilir; (2) bilişsel olarak geliştirilmiş insanların yaratılabilirliği makine zekânın ileri biçimlerinin gerçekleştirilebilir olduğuna olan inancı pekiştiriyor; çünkü *biz* temel olarak makine zekâ yaratamasak bile (bunu varsaymak için hiçbir neden yoktur), makine zekâ yine de bilişsel olarak geliştirilmiş insanların menzili dahilinde olabilir; (3) bu yüzyılın ikinci yarısına ve ötesine epeyce uzanan senaryoları düşündüğümüzde, genetik açıdan geliştirilmiş nüfus (seçmenlerin, mucitlerin, bilimcilerin) kuşağının muhtemel ortaya çıkışını geliştirme büyüklüğünün hızla sonraki onyıllara sirayet ettiğini düşünerek dikkate almak zorundayız.

Beyin-bilgisayar arayüzleri

Doğrudan beyin-bilgisayar arayüzlerinin, özellikle de implantların insanlara dijital bilgisayarların yeteneklerini sömürme imkânı sunduğu kimi zaman savunulmaktadır (kusursuz hatırlama, hızlı ve doğru aritmetik hesabı ve yüksek bant genişliğinde veri nakli) ve böylece ortaya çıkan melez sistemin geliştirilmemiş beyinden misli misli daha iyi performans göstermesi mümkün oluyor.[50] Fakat insan beyinleri ile bilgisayarlar arasındaki doğrudan bağlantıların mümkün olup olmadığı gösterilmiş olsa da, bu tür arayüzlerin yakın zamanda geliştirme olarak geniş ölçekte kullanılması muhtemel görünmüyor.[51]

Bir defa, beyne elektrotlar yerleştirirken enfeksiyon, elektrot deplasmanı, beyin kanaması ve bilişsel gerileme dahil olmak üzere çeşitli tıbbi sorunlarla karşılaşma riski hayli yüksektir. Belki de beyin uyarılmasıyla bu zamana kadar elde edilebilmiş faydaların en canlı örneği Parkinson hastalarının tedavisidir. Parkinson implantı görece basittir: Aslında beyinle iletişim kurmaz, sadece alt talamik çekirdeğe bir uyarıcı elektrik verir. Örnek amaçlı çekilmiş bir videoda hastalıktan ötürü tamamen hareketsiz haldeki bir hastanın sandalyeye çöktüğü, ardından elektrik verildiğinde aniden yaşam belirtisi gösterdiği görülüyor: Denek artık kollarını hareket ettirebiliyor, ayağa kalkıyor, odada dolaşıyor, etrafında dönüp tek ayak üzerinde dönüşler gerçekleştiriyor. Fakat bilhassa basit ve neredeyse mucizevi olan bu başarılı işlemin bile gerisinde olumsuzluklar yatıyor. Derin beyin implantları yapılmış olan Parkinson hastalarına dair bir inceleme, konuşma akıcılığında, seçici dikkatte, renk adlandırmada ve kelime hafızasında gerilemeler olduğunu gösteriyor. Tedavi gören denekler de daha bilişsel şikâyetlerde bulunuyorlar.[52] Bu tür riskler ve yan etkiler şiddetli sakatlığı dindirmek için bu işleme başvurulduğunda hoşgörülebilir. Ama sağlıklı deneklerin beyin cerrahisi için gönüllü olmalarını sağlamak istiyorsak, normal işlevselliğin çok ciddi düzeyde ilerletilmesi gerekiyor.

Bu da bizi süper zekânın siborglaşmayla kazanılacağına şüpheyle yaklaşmamızın ikinci nedenine götürüyor: Geliştirme muhtemelen terapiden çok daha zor olacaktır. Felç geçirmiş hastalar kopmuş sinirlerinin yerine yenisini koyan ya da omurga hareketi örüntü üreteçlerini harekete geçiren bir implanttan yarar sağlayabilirler.[53] Sağır ya da kör olan hastalar yapay kulak salyangozundan ve retinalardan yarar sağlayabilirler.[54] Parkinson hastalığı ya da kronik ağrısı olan hastalar beynin belli bir alanında faaliyeti uyandıran ya da ketleyen derin beyin uyarılmasından yarar sağlayabilirler.[55] Başarılması çok daha zor görünen şey beyin ile bilgisayar arasında başka araçlarla elde edilmesi ha deyince mümkün olmayan bir biçimde ciddi zekâ artışı sağlayacak bir yüksek bant genişliğinde doğrudan etkileşimdir. Beyin implantlarının sağlıklı deneklerde sağlayabildiği potansiyel yararların çoğu düzenli motor ve duyu organlarımızı beyinlerimizin dışında bulunan bilgisayarlarla etkileşim için kullanmak suretiyle çok daha az risk,

masraf ve sıkıntıyla elde edilebilir. İnternete erişmek için beyinlerimize bir fiber optik kablo takmamız gerekmiyor. İnsan retinası saniyede neredeyse 10 milyon parçacık gibi etkileyici bir hızla veri aktarımı yapabilmesinin yanı sıra, bu bilgi sağanağından anlam çıkarmaya ve daha fazla işlem yapmak için diğer beyin alanlarıyla arayüz oluşturmaya son derece yatkın olan muazzam miktarda adanmış bir biyolojik zekâyla (görme korteksiyle) önceden hazır halde geliyor.[56] Beyinlerimize daha fazla bilgi pompalamanın kolay bir yolu olsaydı bile, fazladan bilgi akışı, verileri anlamlandırmak için gerekli tüm nöral makine benzer şekilde güncellenmediği takdirde bizim düşünme ve öğrenme hızımızı artırma konusunda çok az şey yapabilirdi. Buna beynin neredeyse tamamı da dahil olduğundan, gerekli olan şey "tüm beyin protezi"dir ki, bu da yapay genel zekâ demenin bir diğer yoludur. Fakat insan düzeyinde yapay zekâya sahip olan biri beyin cerrahisine ihtiyaç duymayabilir: Bir bilgisayar kemik yerine metal bir kasaya da sahip olabilir. Dolayısıyla bu sınırlı örnek bizi halihazırda incelemiş olduğumuz yapay zekâ yoluna geri götürüyor.

Beyin-bilgisayar arayüzü, diğer beyinlerle ya da makinelerle iletişim amacıyla beyinden bilgi elde etmenin bir yolu olarak da önerilmiştir.[57] Bu tür uplinkler, *locked-in* [kilitli kalma] sendromu olan hastaların düşünce yoluyla ekrandaki bir ibreyi hareket ettirmelerini sağlayarak dış dünyayla iletişim kurmalarına yardımcı olmuştur.[58] Bu tür deneylerde elde edilen bant genişliği düşüktür: Hasta dakikada çok az kelime hızıyla ardı ardına yavaş bir harf yazmayı başarır. Bu teknolojinin ileri versiyonlarını rahatlıkla düşünebiliriz: Belki de yeni kuşak bir implant, Broca alanına (ön lobda dil üretimiyle ilgili bir bölge) yerleştirilebilir ve iç konuşma başlatılabilir.[59] Fakat bu tür bir teknoloji felç ya da kas dejenerasyonunun tetiklediği sakatlıkları olan bazı hastalara yardımcı olabilirse de, sağlıklı bireyler için çok az çekicilik taşır. Sağlayacağı işlevsellik esasen konuşma tanıma yazılımıyla birleştirilmiş bir mikrofon işlevidir ki, bu halihazırda zaten mevcuttur, üstelik beyin cerrahisiyle bağlantılı olan acı, rahatsızlık, masraf ve risklerin hiçbiri yoktur (üstüne bir de, kafa içi dinleme cihazının Orwell'den çok Orwellci tonlarından en azından bazılarını da düşünün). Makinelerimizi bedenlerimizin dışında tutmak da ilerlemeyi kolaylaştırır.

Peki, kelimeleri tamamen es geçip kavramların, düşüncelerin ya da tüm bir uzmanlık alanının bir zihinden diğerine "indirilme"sini mümkün kılan iki beyin arası bir bağlantı kurma hayaline ne dersiniz? Bilgisayarlarımıza milyonlarca kitap ve makalenin olduğu kütüphaneler de dahil olmak üzere büyük dosyalar indirebiliyoruz ve bunu yapmak saniyeler alıyor: Beyinlerimizde de benzer bir şey yapılabilir mi? Bu fikrin görünüşte ikna edici olması bilginin beyinde nasıl depolandığına ve temsil edildiğine dair yanlış bir görüşten kaynaklanıyor sanırım. Belirtildiği üzere, insan zekâsındaki hız sınırlayıcı adım ham verilerin beyne hangi hızla yollanabileceği değil, beynin bilgi çıkarıp verileri anlamlandırmasının ne kadar çabuk olduğudur. Belki de anlamları

alıcı tarafından deşifre edilmesi gereken duyusal veriler halinde paketlemekten ziyade doğrudan aktardığımız söylenecektir. Burada iki sorun var. Birincisi, beyinler, bilgisayarlarımızdaki tipik programların aksine, standartlaştırılmış veri depolaması ve temsil formatları kullanmazlar. Daha ziyade, her beyin yüksek düzey içeriğin kendine özgü temsillerini geliştirir. Tikel bir kavramı temsil etmek için hangi nöronal toplanmalardan yararlanıldığı söz konusu beynin benzersiz deneyimlerine (ve çeşitli genetik etkenlerle stokastik fizyolojik süreçlere) bağlıdır. Tıpkı yapay nöral ağlarda olduğu gibi, biyolojik nöral ağlarda da anlam muhtemelen hayli büyük çakışan bölgelerin yapısında ve faaliyetinde bütüncül olarak temsil edilir, düzenli bir şekilde serili ayrı bellek hücrelerinde değil.[60] Dolayısıyla bir beyindeki nöronlarla diğer bir beyindeki nöronlar arasında düşüncelerin birinden diğerine otomatik olarak kayabileceği basit bir haritalandırma yapmak mümkün değildir. Bir beyindeki düşüncelerin diğeri tarafından anlaşılabilmesi için, düşüncelerin ortak bir uzlaşım uyarınca parçalara ayrılıp semboller halinde paketlenmesi gerekiyor ki, semboller alıcı beyin tarafından doğru bir şekilde yorumlanabilsin. Bu, dilin işidir.

İlkesel açıdan, telaffuz ve yorumlama gibi bilişsel görevleri yollayıcının beynindeki nöral durumları bir şekilde sesli okuyacak ve bir şekilde alıcının beynine ısmarlama bir aktivasyon örüntüsünde besleyecek bir arayüze boşaltmanın mümkün olduğu tahayyül edilebilir. Fakat bu da bizi siborg senaryosundaki ikinci soruna getiriyor. Belki tekil olarak adreslenebilir nöronların milyarlarcasından eşzamanlı olarak nasıl yazıp okunabileceğine dair teknik zorluğu (muazzam büyüklükteki) bir kenara bıraksak bile, gerekli arayüzü yaratmak herhalde bir yapay zekâ problemidir. Arayüzün bir beyindeki ateşleme örüntülerinin diğer beyindeki semantik açıdan denk ateşleme örüntülerine (gerçek zamanda) haritalandırabilen bir bileşeni içermesi gerekecektir. Bu tür bir görevin üstesinden gelmek için gerekli nöral hesaplamanın ayrıntılı çokdüzeyli kavranışı nöromorfik yapay zekâyı doğrudan mümkün kılıyor görünecektir.

Bu şerhlere rağmen, bilişsel gelişime giden siborg yolu hiç ümit vaat etmiyor değildir. Fare hipokampüsü üzerindeki yapılan etkileyici çalışma basit bir işler bellek görevindeki performansı artırabilecek bir nöral protezin gerçekleştirilebilir olduğunu ispatlamıştır.[61] Mevcut haliyle, implant, hipokampüsün bir alanında ("CA3") yer alan bir ya da iki düzine elektrottan verileri toplayıp başka bir alandaki ("CA1") benzer sayıda nörona yansıtabiliyor. Bir mikroişleyici ilk alandaki iki farklı ateşleyici örüntü arasında ayrım yapabilecek (iki farklı belleğe denk düşer: "Sağ kol" ve "sol kol") ve bu örüntülerin ikinci alana nasıl yansıtıldığını öğrenecek şekilde eğitilmiştir. Bu protez iki sinir alanı arasındaki normal nöron bağlantısı engellendiğinde işlevi yeniden çalıştırmakla kalmıyor, aynı zamanda belli bir bellek örüntüsünün özellikle net bir işaretini ikinci alana yansıtarak bellek görevindeki performansı farenin normalde yapabileceklerinin ötesinde artırabiliyor. Bu, günümüz standartlarında teknik bir başarı

olmakla birlikte, yine de cevaplanmayı bekleyen bir sürü zor soru vardır: Bu yaklaşımın çok daha fazla sayıda bellek için geçerliliği nedir? Girdi ve çıktı sayısı arttıkça doğru haritalandırmayı gerçekleştirilemez hale getirme tehdidi barındıran kombinasyonel patlamayı ne kadar iyi kontrol edebiliriz? Test görevindeki artırılmış performansın acısı başka bir yerden çıkar mı, mesela deneyde kullanılan tikel uyarılmadan genelleme yapma yetisinin azalması ya da çevre değiştiğinde çağrışımı unutma yetisinin azalması gibi bir sonuç doğar mı? Test denekleri, farelerin aksine, kâğıt-kalem gibi dış hafıza yardımcılarından yararlanabilseler bile yine de yarar sağlayabilirler mi? Ve benzer bir yöntemi beynin diğer kısımlarına uygulamak ne kadar daha zor olacaktır? Mevcut protez hipokampüsün görece basit ileri-besleme yapısından istifade etmekle birlikte (özünde CA3 ve CA1 bölgeleri arasında tek yönlü bir köprü hizmeti görüyor), kortekstekı diğer yapılar bağlantı şemasının karmaşıklığını ve muhtemelen yerleşik grup, nöronların işlevselliğini deşifre etmenin zorluğunu büyük oranda artıran çapraşık geri besleme döngüleri içerebilir.

Siborg yolu için şöyle bir umut var: Beyni dışarıdaki bir kaynağa bağlayan kalıcı olarak implant edilmiş bir araç olursa, beyin zaman içinde kendi iç bilişsel durumları ile aletten aldığı girdiler ya da aleti kabul ettiği çıktılar arasındaki etkili bir haritalandırmayı öğrenebilir. O zaman implantın kendisinin zeki olmasına gerek kalmayacaktır; daha ziyade, beyin arayüze zeki bir şekilde uyum sağlayacaktır, tıpkı bir bebeğin beyninin gözlerindeki ve kulaklarındaki reseptörlerden gelen sinyalleri kerte kerte yorumlamayı öğrenmesi gibi.[62] Ama burada bir kez daha ne kadar kazanç elde edileceği sorusu sorulmalıdır: Beynin esnekliğinin bir beyin-bilgisayar arayüzü yoluyla korteksin bir parçasına yansıtılan yeni bir girdi akışındaki örüntüleri yakalamayı öğrenebilecek düzeyde olduğunu varsayalım: Peki, aynı bilgiyi bunun yerine görsel bir örüntü olarak retinaya ya da kulak salyangozuna ses olarak niye yansıtmayalım? Düşük teknolojili alternatif yüzlerce komplikasyondan sakınır ve her iki durumda da beyin bilgiyi anlamlandırmayı öğrenmek için kendi örüntü tanıyan mekanizmalarını ve esnekliğini kullanabilir.

Ağlar ve örgütlenmeler

Süper zekâya giden tasavvur edilebilir yollardan bir diğeri de bireysel insan zihinlerini birbirine ve çeşitli insan eserlerine ve botlara bağlayan ağların ve örgütlenmelerin tedricen geliştirilmesidir. Burada fikir, bunun bireylerin entelektüel kapasitesinin onları süper zeki hale getirecek kadar artırılması değil, daha ziyade bu şekilde ağlandırılmış ve örgütlenmiş bireylerden oluşan bir sistemin bir süper zekâ biçimine ulaşıp ulaşamayacağıdır, yani bir sonraki bölümde "kolektif süper zekâ" olarak irdeleyeceğimiz meseledir.[63]

İnsanlık tarih ve tarihöncesi boyunca kolektif zekâ bakımından muazzam kazanımlar elde etmiştir. Bu kaynaklar arasında yazı ve matbaa gibi iletişim teknolojisindeki yenilikler, dünya nüfusu ve yerleşim yoğunluğundaki artışlar; örgütlenme tekniklerindeki ve epistemik normlardaki çeşitli iyileştirmeler ve kurumsal sermayenin tedrici bir birikimi sayılabilir. Genel itibariyle, bir sistemin kolektif zekâsı üye zihinlerinin yetenekleri, ilintili bilgiyi aralarında paylaşmanın sabit giderleri ve insan örgütlenmelerine hâkim olan çeşitli çarpıklıklar ve verimsizliklerle sınırlanmıştır. Eğer iletişimin sabit giderleri azaltılırsa (sadece ekipman masrafları değil, aynı zamanda yanıttaki gecikmeler, zaman ve dikkat yükleri ve diğer etkenler de) o zaman daha büyük ve daha yoğun bağlantılı örgütlenmeler yaratılabilir. Aynısı örgütlenme yaşamını bozan bazı bürokratik çarpıklıklara çözümler üretilirse de olabilir: Savurgan statü oyunları, görev sapması, bilginin saklanması ya da tahrif edilmesi ve diğer aracılık problemleri. Bu sorunlara kısmi çözümler bile kolektif zekâya çok büyük getiriler sağlayabilir.

Kolektif zekâmızın büyümesine yardımcı olabilecek teknolojik ve kurumsal yenilikler çok ve çeşit çeşittir. Örneğin sübvanse edilen tahmin piyasaları hakikat arayan normları güçlendirip ihtilaflı bilimsel ve toplumsal meselelere dair öngörüleri iyileştirebilir.[64] Yalan makineleri (güvenilir ve kullanımı kolay olan makineler yapılabilirse) insan ilişkilerinde yalan dolanı azaltabilir.[65] Kendini kandırma detektörleri daha da güçlü olabilir.[66] Yeni çıkmış beyin teknolojileri bile olmadan, bazı aldatma biçimleri çeşitli türlerde verinin mevcudiyeti giderek arttığından (şöhret ve performans kaydı dahil) veya güçlü epistemik normların ve akılsallık kültürünün benimsenmesinden ötürü tarihin karanlığına doğru yol alabilir. Gönüllü ya da gönülsüz gözetleme insan davranışına ilişkin muazzam bir bilgi toplanmasını sağlayacaktır. Sosyal ağ siteleri kişisel ayrıntıları paylaşmak için şimdiden bir milyar insan tarafından kullanılıyor: Çok geçmeden bu insanlar akıllı telefonlarındaki ya da gözlük çerçevelerindeki gömülü mikrofonlardan ve video kameralarından sürekli yaşamsal kayıtlar yüklemeye başlayabilirler. Bu tür veri akışlarının otomatik analizi birçok yeni uygulamanın önünü açacaktır (elbette hayırlısı da olacak uğursuzu da).[67]

Kolektif zekâdaki büyüme daha genel örgütlenme ve ekonomik iyileştirmelerden ve dünya nüfusunun eğitimli, dijital olarak bağlantılı ve küresel entelektüel kültürle bütünleşmiş olan kısmının büyümesinden de kaynaklanabilir.[68]

İnternet yenilik ve deneyler konusunda bilhassa dinamik bir cephe olarak göze çarpıyor. İnternetin potansiyellerinin çoğu hâlâ kullanılmayı bekliyor. Düşünme, önyargıların atılması ve yargı yığışması için daha iyi destekler sunan bir akıllı internetin sürekli gelişimi bir bütün olarak insanlığın ya da tikel grupların kolektif zekâsının artmasına büyük katkılar sunabilir.

İnternetin bir gün "uyanacağı"na dair görünüşte daha uçuk kaçık fikre ne demeli? İnternet birbirine gevşek bağlarla bağlı kolektif süper zekânın sadece omurgası olmadan

daha fazlası haline, yeni ortaya çıkan birleşik bir süper zekâya ev sahipliği yapan sanal bir kafatası gibi bir şey haline gelebilir mi? (Vernor Vinge'in 1993 yılına ait etki yaratmış ve "teknolojik tekillik" kelimesini yazına kazandırmış olan makalesine göre, süper zekânın doğabileceği kanallardan biri buydu.[69]) Buna karşı makine zekânın meşakkatli bir mühendislik yoluyla gerçekleştirilmesinin yeterince zor olduğu, kendiliğinden doğacağını varsaymanın akla izana sığmayacağı itirazı getirilebilir. Gelgelelim internetin gelecekteki bir versiyonunun salt tesadüf eseri birdenbire süper zeki hale geleceğini varsaymamak gerekir Senaryonun daha akla yatkın bir versiyonu internetin yıllardır birçok insanın çalışmasıyla (daha iyi arama ve bilgi filtreleme algoritmalarının yaratılması, daha güçlü veri temsili formatları, daha ehil özerk yazılım aracıları ve bu tür botlar arasındaki etkileşimleri yöneten daha verimli kurallar) ilerlemeleri biriktirmesi ve çeşit çeşit artımlı iyileştirmenin sonunda daha bütünlüklü bir web zekâsı biçiminin temelini yaratması olacaktır. Bu türde web temelli bir bilişsel sistemin patlamalı büyüme için gerekli bilgisayar gücü ve tek bir hayati bileşen hariç diğer tüm kaynaklarla donandığında, eksik tek parça da eklenince süper zekâyla parıldayabileceği en azından tasavvur edilebilir görünmektedir. Ama bu tür bir senaryo süper zekâya giden bir başka olanaklı yola da yaklaşmaktadır: Halihazırda ele almış olduğumuz yapay genel zekâ.

Özet

Süper zekâya giden birçok yol olması er geç hedefe varılacağına dair güvenimizi artırmalıdır. Yollardan biri tıkansa bile, yine de ilerleyebiliriz.

Birden çok yol olması birçok varış noktası olacağı anlamına gelmez. Kayda değer bir zekâ ilerlemesi makine olmayan zekâ yollarının birinden geçerek öncelikli olarak başarılsa bile, bu durum makine zekâyı çöpe atmaz. Tersine: Geliştirilmiş biyolojik ya da örgütsel zekâ bilimsel ve teknolojik gelişmeleri hızlandıracak ve tüm beyin emülasyonu ve yapay zekâ gibi zekâ artışının daha radikal biçimlerinin varışını potansiyel olarak hızlandıracaktır.

Bu demek değildir ki, makine süper zekâsına nasıl varılacağı önemsizdir. Oraya varmak için gidilen yol nihai sonuca büyük bir etkide bulunabilir. Elde edilen nihai yeterlilikler hangi güzergâhın tutulacağına fazla bağlı olmasa bile, bu yeteneklerin nasıl kullanılacağı (biz insanların süreç üzerinde ne kadar kontrole sahip olacağı) pekâlâ yaklaşımımızın ayrıntılarına bağlı olabilir. Örneğin biyolojik ya da örgütsel zekâdaki ilerlemeler riski öngörme ve hem güvenli hem de faydalı bir makine süper zekâ tasarlama yeteneğimizi artırabilir. (Tam teşekküllü bir stratejik değerlendirme birçok karmaşıklık içerir ve 14. Bölüm'ü beklemek gerekiyor.)

Gerçek süper zekâ (mevcut zekâ düzeylerindeki marjinal artışlardan farklı olarak) öncelikle yapay zekâ üzerinden akla yatkın şekilde elde edilebilir. Gelgelelim bu yolda birçok temel belirsizlik vardır. Bu nedenle yolun ne kadar uzak olduğuna ya da kaç tane

engel içerdiğine dair sıkı bir değerlendirme yapmak zordur. Tüm beyin emülasyonu yolunun da süper zekâya giden en çabuk yol olma ihtimali vardır. Bu yolda ilerleme teorik atılımlardan ziyade birçok artımlı teknolojik ilerleme gerektirdiğinden, nihayetinde başarılı olacağı iddiası güçlü bir şekilde savunulabilir. Gelgelelim tüm beyin emülasyonu yolunda ilerleme çabuk olsa bile, yapay zekânın yine de bitiş çizgisini ilk geçen olma ihtimali epey yüksektir: Bunun nedeni kısmi taklitlere dayanan nöromorfik yapay zekâların mümkün olmasıdır.

Biyolojik bilişsel ilerlemelerin, özellikle de genetik seçilime dayalı olanların gerçekleştirilebilir olduğu açıktır. Yinelenmiş embriyo seçilimi şu anda bilhassa umut vaat eden bir teknoloji gibi görünmektedir. Ne var ki makine zekâsındaki olanaklı atılımlarla kıyaslandığında biyolojik ilerlemeler görece yavaş ve tedrici olacaktır: En iyi ihtimalle, süper zekânın görece zayıf biçimlerini doğuracaktır. Biyolojik geliştirmenin gerçekleştirilebilir olması makine zekânın nihayetinde yaratılabileceğine dair güvenimizi artırmalıdır, zira geliştirilmiş insan bilimciler ve mühendisler doğal durumdaki muadillerinden daha fazla ve daha hızlı ilerleme kaydedebilecektir. Özellikle de makine zekânın yüzyılın ortalarına kadar geciktiği senaryolarda, artan derecede bilişsel olarak geliştirilmiş grupların sahneye çıkması sonraki gelişmelerde artan bir rol oynayacaktır.

Beyin-bilgisayar arayüzlerinin bir süper zekâ kaynağı olması pek ihtimal dahilinde görünmüyor. Ağlardaki ve örgütlenmelerdeki ilerlemeler uzun vadede kolektif zekânın zayıf derecede süper zeki biçimlerini doğurabilir; ama daha muhtemel olanı, biyolojik bilişsel ilerlemeninkine benzer bir kolaylaştırıcı rol oynayıp, insanlığın entelektüel problemleri çözme konusundaki yeteneğini tedricen artırmalarıdır. Biyolojik geliştirmelerle kıyaslandığında, ağlardaki ve örgütlenmedeki ilerlemeler çok geçmeden bir fark yaratacaktır; aslında bu tür ilerlemeler sürekli yaşanıyor ve şimdiden ciddi etkide bulunuyor. Gelgelelim ağlardaki ve örgütlenmelerdeki ilerlemeler biyolojik bilişimizdeki ilerlemelere kıyasla problem çözme yeteneğimizde daha dar artışlara sebep olabilir: Bir sonraki bölümde bahsedeceğimiz ayrımla söylersek, "nitelikli zekâ"dan ziyade "kolektif zekâ"yı destekleyebilir.

Süper Zekânın Biçimleri

O halde, "süper zekâ" derken tam olarak ne kastediyoruz? Terminolojinin gayya kuyusunda kaybolmak istemiyor olsak da, kavramsal zemini netleştirmek için bir şeyler söylemek gerekiyor. Bu bölümde süper zekânın üç farklı biçimi tespit ediliyor ve bunların pratikte ilintili bir anlamda eşdeğer olduğu savunuluyor. Ayrıca bir makine substratındaki zekâ potansiyelinin biyolojik bir substrattakinden çok daha büyük olduğunu gösteriyoruz. Makineler, ezici üstünlüğü ele geçirmelerini sağlayacak birtakım temel üstünlüklere sahiptir. Biyolojik insanlar, geliştirilseler bile, makinelerle aynı kıratta olmayacaktır.

Birçok makine ve insan olmayan hayvan şimdiden belli dar alanlarda insanüstü düzeylerde performans gösteriyorlar. Yarasalar deniz radarı sinyallerini insanlardan daha iyi algılıyorlar, hesap makineleri aritmetikte bizden daha iyiler ve satranç programları satranç tahtasını koltuğumuzun altına verebiliyor. Yazılımın daha iyi yapabileceği belli görevlerin kapsamı genişlemeye devam edecek. Ama uzmanlaşmış bilgi işleme sistemleri birçok alanda kullanılacak olsa da, sadece insanları ikame edecek kadar genel zekâya sahip olan makine zekâlar olasılığıyla ortaya çıkan başka derin meseleler de vardır.

Daha önce belirtildiği üzere, "süper zekâ" terimini çok genel nitelikteki birçok bilişsel alanda şu an var olan insan zihinlerini misli misli geride bırakan zekâlar için kullanıyoruz. Bu yine de muğlak bir kavram. Çok farklı performans özellikleri olan farklı türde sistemler bu tanıma göre süper zekâ kapsamına girebilir. Analizi ilerletmek adına, bu basit süper zekâ kavramını entelektüel süper yeteneklerin farklı kümelerini ayrıştırarak açmakta yarar var. Bu tür bir ayrıştırma farklı şekillerde yapılabilir. Biz burada üç biçimini ayıracağız: hızlı süper zekâ, kolektif süper zekâ ve nitelikli süper zekâ.

Hızlı süper zekâ

Hızlı süper zekâ, tıpkı insan zihni gibi, bir zekâdır, ama daha hızlıdır. Süper zekânın kavramsal açıdan analiz edilmesi en kolay olan biçimidir.[1] Hızlı süper zekâyı şu şekilde tanımlayabiliriz.

Hızlı süper zekâ: *Bir insan zekâsının yapabileceği her şeyi yapan ama çok daha hızlı yapan bir sistem.*

Burada "çok daha" derken "birçok farklı büyüklük düzeyinde" gibi bir şey kastediyoruz. Ama tanımda hiçbir muğlak nokta kalmasın diye uğraşmak yerine, okura bu tanımı akla uygun şekilde yorumlamasını tavsiye edeceğiz.[2]

Hızlı süper zekânın en basit örneği hızlı donanımda çalışan bir tüm beyin emülasyonu olacaktır.[3] Bir biyolojik beyine kıyasla on bin kat daha hızlı çalışan bir emülasyon, bir kitabı birkaç saniyede okuyup öğleden sonra doktora tezi yazabilecektir. Bir milyonluk bir hız farkı olduğundaysa, emülasyon beyin bin yılda yapılacak bir entelektüel faaliyeti bir çalışma gününe sığdırır.[4]

Bu denli hızlı bir zihin için dış dünyadaki olaylar yavaş çekimde gerçekleşiyor gözükecektir. Zihninizin 10 bin kat hızlı çalıştığını varsayın. Kanlı canlı arkadaşınız çay fincanını yere düşürecek olsa, porselenin halıya tıpkı uzak bir gezegenden uzaya fırlatılan bir kuyrukluyıldızın havada süzülmesi gibi birkaç saatte yavaş yavaş süzüldüğünü izleyebilirsiniz. Ve yaklaşan felaketin arkadaşınızın önce gri maddesinin kıvrımlarına, oradan da çevresel sinir sistemine yayıldığını öngörerek, bedeninin tedrici olarak donmuş bir "eyvah" konumuna büründüğünü gözlemleyebilirsiniz, sonuçta sadece başka bir fincan sipariş etmek için değil birkaç sayfa bilimsel makale okuyup, hafif bir kestirmek için de zamanınız olacaktır.

Maddi dünyadaki bu zahiri zamansal genişlemeden ötürü hızlı süper zekâ dijital nesnelerle çalışmayı tercih edecektir. Bir sanal gerçeklikte yaşayıp enformasyon ekonomisinde iş görebilir. Yahut fiziksel çevreyle nano ölçekte manipülatörler aracılığıyla iletişim kurabilir, zira böyle küçük ölçeklerde kol-bacaklar makroskopik eklentilerden daha hızlı işlev görebilir. (Bir sistemin temel nitelikteki sıklığı genellikle uzunluk ölçeğiyle ters orantılıdır.[5]) Hızlı bir zihin yavaş gelişen, şeker şerbeti gibi insanlardan ziyade, esasen, diğer hızlı zihinlerle iletişim kurabilir.

Işık hızı zihinler hızlandıkça giderek daha önemli bir kısıtlama haline gelir, zira hızlı zihinler uzun mesafelerde seyahat etmek ya da iletişim kurmak için kullandıkları zaman bakımından büyük fırsat maliyetleriyle karşılaşırlar.[6] Işık, kabaca söylersek, bir jet uçağından bir milyon kat daha hızlıdır, dolayısıyla zihinsel olarak yaklaşık 1 milyon kat hızlandırılmış bir dijital ajanın yerkürenin etrafında dolaşması insan bir seyyahınkiyle öznel açıdan yaklaşık aynı zaman miktarına denk gelecektir. Uzak mesafedeki bir insanı arama oraya "bizzat" gitmek kadar uzun sürecektir, ama aramak daha az bant genişliği gerektirdiğinden daha ucuz olacaktır. Büyük zihinsel hızlandırıcıları olan ajanlar geniş ölçekte iletişim kurmak isterlerse birbirlerine yaklaşmayı daha yararlı bulabilirler. Sık sık etkileşime ihtiyaç duyan son derece hızlı zihinler (bir

çalışma ekibinin üyeleri gibi) can sıkıcı gecikmelerden kaçınmak için aynı binada yer alan bilgisayarlarda ikamet edebilirler.

Kolektif süper zekâ

Süper zekânın bir diğer biçimi çok sayıda küçük zekânın toplanmasıyla üstün performansa ulaşan bir sistemdir.

Kolektif süper zekâ: *Çok sayıda küçük zekâdan oluşan bir sistem; bu sistemin birçok farklı genel alandaki genel performansı mevcut bilişsel sistemlerin hepsinden katbekat daha üstündür.*

Kolektif süper zekâ kavramsal açıdan hızlı süper zekâ kadar net sınırlara sahip değildir.[7] Gelgelelim ampirik açıdan daha tanıdıktır. Saat hızında kayda değer derecede farklılık gösteren insan düzeyinde zihinlerle herhangi bir deneyimimiz olmamakla birlikte, kolektif zekâyla, çeşitli verimlilik derecelerinde birlikte çalışan insan düzeyindeki çeşitli sayıda bileşenden oluşan sistemlerle birçok deneyimimiz var. Şirketleri takım çalışmaları, dedikodu ağları, savunma grupları akademik topluluklar, hatta bir bütün olarak insanlık –eğer biraz soyut bir perspektif benimseyebilirsek– entelektüel problemleri çözebilen gevşek şekilde tanımlanmış "sistemler" olarak görülebilir. Farklı görevlerin çeşitli büyüklükte ve bileşimde örgütlenmelerin çabalarına ne kadar kolay kendini kaptırdığını kendi deneyimlerimizden hareketle biliyoruz.

Kolektif zekâ, altproblemlere çözümlerin paralel olarak takip edilip bağımsız olarak doğrulanabileceği şekilde kolaylıkla parçalara ayrılabilen problemleri çözmekte uzmandır. Bir uzay mekiği yapmak ya da bir hamburgerci zinciri işletmek gibi görevler işbölümü için birçok fırsat sunar: Farklı mühendisler uzay mekiğinin farklı bileşenleri üzerinde çalışırlar; farklı işçiler farklı restoranlarda çalışırlar. Akademide araştırmacılar, öğrenciler, dergiler, burslar ve ödüller şeklinde kendi içine kapanık ayrı disiplinler olarak sıkı bölünme –bu kitabın temsil ettiği çalışma türüne yardımcı olmasa da– farklı saikleri olan bireylerin ve takımların her birinin bir yandan kendi alanında bağımsız olarak çalışırken diğer yandan insan bilgisinin artmasına yardımcı olmasına izin vermenin pratikliklerine zorunlu bir intibak olarak görülebilir (ama ancak uzlaşmacı ve rahat bir kafa yapısıyla).

Bir sistemin kolektif zekâsı onu oluşturan zekâların sayısını ya da niteliğini artırarak ya da örgütlenmelerinin niteliğini ilerleterek geliştirilebilir.[8] Günümüzde var olan kolektif zekâlardan bir kolektif süper zekâ elde etmek çok ciddi derecede geliştirme gerektirecektir. Ortaya çıkan sistem birçok farklı genel alandaki mevcut kolektif zekâlardan ya da diğer bilişsel sistemlerden misli misli daha iyi performans gösterebilmelidir. Akademisyenlere daha etkili bir şekilde bilgi alışverişi yapma imkânı sunan yeni bir konferans formatı ya da kullanıcıların kitap ve sinema filmlerine verdikleri

puanları daha iyi tahmin eden yeni bir bilgi filtreleyici algoritma tek başına kolektif süper zekâya yaklaşan bir şey olmayacaktır. Keza dünya nüfusunda yüzde ellilik bir artış ya da öğrencilere bir okul gününü altı saat yerine dört saatte tamamlama imkânı veren bir ilerleme de öyle. İnsanlığın kolektif bilişsel kapasitesindeki çok daha aşırı bir büyümenin kolektif süper zekâ kıstasını karşılaması gerekecektir.

Kolektif süper zekâ eşiğinin bugünkü, yani 21. yüzyıl başı performans düzeyine bağlı olduğunu belirtelim. İnsanlığın tarihöncesi boyunca ve yine insan tarihi boyunca kolektif süper zekâ çok büyük faktörlerde büyümüştür. Örneğin dünya nüfusu Buzul Çağı'ndan bu yana en az bin faktör artmıştır.[9] Salt bu temelde, insanın kolektif zekâsının mevcut düzeyleri *bir buzul çağı taban çizgisine kıyasla* süper zekâya yakın bir şey olarak görülebilir. İletişim teknolojilerindeki bazı ilerlemelerin (özellikle de konuşma dili, ama belki şehirler, matbaa ve yazı da öyle) tek tek ya da birlikte büyük boyutta atılımlar sağladığı söylenebilir, şöyle ki eğer kolektif problem çözme kapasitemize benzer etkide bir başka yenilik daha olsa, kolektif süper zekâyla sonuçlanacaktır.[10]

Bu noktada bir grup okur modern toplumun o kadar da zeki olarak görünmediği itirazında bulunabilir. Belki de okurun anavatanında nahoş bir siyasal karar alınmıştır ve bu kararın görünüşteki aptallığı modern çağın zihinsel sığlığının kanıtı olarak okurun zihninde büyük bir yer tutacaktır. Günümüzde insanlık maddi tüketimi fetişleştirmiyor mu, doğal kaynakları kurutmuyor mu, çevreyi kirletip tür çeşitliliğini yok etmiyor mu ve tüm bunları yaparken avaz avaz bağıran küresel eşitsizliklere çare bulmakta başarısız olup en temel insani ya da manevi değerleri boşlamıyor mu? Gelgelelim modernliğin kusurlarının önceki çağların o kadar da dikkate değer olmadığı söylenemeyecek kusurlarıyla karşılaştığında nasıl göründüğü sorusunu bir kenara bırakırsak, kolektif süper zekâ tanımımızdaki hiçbir şey kolektif zekâsı daha büyük olan bir toplumun daha iyi konumda olacağını ima etmez. Bu tanım kolektif açıdan daha zeki toplumun *daha bilge* olduğunu bile ima etmez. Bilgeliği önemli şeyleri kabaca doğru algılama yeteneği diye düşünebiliriz. O zaman, çok genel nitelikteki birçok alanda entelektüel problemleri kolektif olarak çözebilen çok verimli bir eşgüdüme sahip bilgi işçilerinden oluşan çok geniş bir kadrodan müteşekkil bir örgütlenme tahayyül etmek mümkündür. Bu örgütlenmenin çok çeşitli işler yapabildiğini, çeşit çeşit teknolojiler icat edebildiğini ve birçok türde işlemi optimum düzeye çekebildiğini varsayalım. Bu durumda bile, büyük resme ait kilit önemdeki birkaç meseleyi tamamen yanlış anlayabilir (sözgelimi varoluşsal risklere dair düzgün önlemler almayı başaramayabilir) ve böylece kısa süreli bir patlayıcı büyüme yaşayıp sonunda yüz kızartıcı bir çöküşe varabilir. Bu tür bir örgütlenme çok yüksek derecede kolektif zekâya sahip olabilir; eğer yeterince yüksekse, örgütlenme bir kolektif süper zekâdır. Normatif açıdan arzulanabilir her türlü özelliği, sanki diğer özellikler de aynı şekilde mevcut olmadan hayran olunacak bir özellik bulunamazmış gibi, amorf nitelikte dev bir zihinsel işleyiş kavramı haline getirme

ayartmasına direnmek gerekir. Bunun yerine, ne içsel olarak iyi ne de güvenilir derecede bilge olan araçsal açıdan güçlü bilgi işleme sistemleri (zeki sistemler) olabileceğini kabul etmemiz gerekir. Ama bu meseleyi 7. Bölüm'de yeniden ele alacağız.

Kolektif süper zekâ gevşek bağlarla da sıkı bağlarla da bağlı bir bütün oluşturabilir. Gevşek bağlarla bağlı kolektif süper zekâya örnek için *MegaDünya* adında bir gezegen hayalet edin. Bu gezegen şu an gerçek yerkürede sahip olduğumuz iletişim ve eşgüdüm teknolojilerine sahip olsun ama nüfusu bir milyon kat daha fazla olsun. Böyle devasa bir nüfusa sahip olan MegaDünya'nın toplam entelektüel işgücü de gezegenimizinkine kıyasla bir o kadar büyük olacaktır. Newton ya da Einstein çapında bir bilim dehasının en azından her 10 milyar insanda bir çıktığını varsayalım: O zaman MegaDünya'da aynı anda yaşayan 700 bin dâhi ve onların yanı sıra biraz daha az yetenekli olan çok sayıda insan olacaktır. Yeni fikirler ve teknolojiler müthiş bir hızla geliştirilecek ve MegaDünya'daki küresel medeniyet gevşek bağlara bağlı bir kolektif süper zekâ oluşturacaktır.[11]

Eğer bir kolektif süper zekânın bütünleşme düzeyini tedricen artırırsak, sonunda bir birleşmiş *zekâ* haline gelebilir: Birbiriyle gevşek bir iletişime sahip olan küçük insan zihinlerinin bir araya gelmesi değil, tek bir büyük "zihin" olabilir.[12] MegaDünya'nın sakinleri iletişim ve eşgüdüm teknolojilerini ilerleterek ve birçok birey için her türlü zor entelektüel problemi birlikte çözmek için daha iyi yollar geliştirerek bu yönde adımlar atabilir. Böylece bir kolektif süper zekâ yeterince bütünleşmiş hale geldiğinde bir "nitelikli süper zekâ"ya dönüşebilir.

Nitelikli süper zekâ

Süper zekânın üçüncü bir biçiminden daha bahsedebiliriz:

Nitelikli süper zekâ: *En azından bir insan zihni kadar hızlı ve niteliksel açıdan çok daha akıllı bir sistem.*

Kolektif zekâda olduğu gibi zekâ niteliği de biraz belirsiz bir kavramdır ve bu örnekte mevcut insan dağılımının üst kademesinin üstündeki zekâ niteliği farklılıklarına ilişkin bir deneyimimiz olmadığından, durum daha da karmaşıklaşmaktadır. Ne var ki konuyla ilintili vakalardan bazılarını ele alarak bu kavramı daha iyi anlayabiliriz.

Birincisi, düşük nitelikte zekâya sahip insan olmayan hayvanları düşünerek referans noktalarımızı artırabiliriz. (Bunu türcü bir ifade olarak almayın. Bir zebra balığının zekâ niteliği ekolojik ihtiyaçlarıyla mükemmel bir uyum içindedir, ama burada ilintili perspektif çok daha insan merkezcidir: Bizi ilgilendiren şey *insani açıdan* ilintili karmaşık bilişsel görevlerdeki performanstır.) İnsan olmayan hayvanlarda karmaşık yapılı dil yoktur; alet kullanma ve alet yapma yetileri yoktur ya da çok sınırlıdır; uzun vadeli planlar yapma yetenekleri fevkalade sınırlıdır ve soyut muhakeme yetenekleri de

çok sınırlıdır. Ayrıca bu sınırlar insan olmayan hayvan zihinleri arasında kolektif zekâ ya da hız yokluğuyla da tamamen açıklanamaz. Ham hesaplama gücü bakımından, insan beyinleri muhtemelen bazı büyük hayvanlarınkinden (filler ve balinalar dahil) daha geridir. İnsanlığın karmaşık teknolojik medeniyeti kolektif zekâ bakımından muazzam üstünlüğümüz olmadan imkânsız olsa da, ayırt edici derecede insana ait olan tüm bilişsel yetenekler kolektif zekâya bağımlı değildir. Birçoğu küçük, yalıtık avcı-toplayıcı gruplarda bile son derece gelişkindir.[13] Ve birçoğu da son derece örgütlü insan olmayan hayvanlar arasında buna yakın derecede gelişkin değildir (örneğin insan eğitmenlerin elinde yoğun eğitimden geçen şempanzeler ve yunuslar ya da kendi büyük ve düzenli toplumlarında yaşayan karıncalar). Belli ki *Homo sapiens*'in kayda değer entelektüel başarısı büyük ölçüde beyin mimarimizin özgül niteliklerine, diğer hayvanlarda bulunmayan benzersiz bir genetik yetinin varlığına yorulabilir. Bu gözlem bize nitelikli süper zekâ kavramını örneklendirmemizde yardımcı olabilir: İnsan nitelikli zekâsı fillerin, yunusların ya da şempanzelerinkinden ne kadar üstünse insan zekâsından o kadar üstün olan nitelikli zekâdır.

Nitelikli süper zekâ kavramını örneklendirmenin ikinci bir yoluysa bireysel insanları etkileyebilen alana özgü bilişsel eksiklerin, özellikle de genel demans ya da beynin nöral hesaplama kaynaklarının tümden yok edilmesiyle alakalı diğer koşulların neden olmadığı eksiklerin belirtilmesidir. Örneğin toplumsal bilişte çarpıcı eksikleri olmasına karşın diğer bilişsel alanlarda iyi iş çıkaran otistik bireyleri ya da basit tınıları mırıldanamayan ya da tanıyamayan ama diğer birçok açıdan normal performans gösteren doğuştan amüziyalı bireyleri düşünün. Nöropsikiyatri yazınından daha başka birçok örnek verilebilir; sonuçta bu yazın genetik anormalliklerin yol açtığı dar şekilde sınırlandırılmış kusurlarla malul ya da beyin travması geçirmiş hastaların vaka çalışmalarıyla doludur. Bu tür örnekler normal yetişkin insanların genel nöral işleme gücüne, hatta genel zekâya yeterli miktarda sahip olmanın bir işlevinden ibaret olmayan kayda değer sayıda bilişsel yeteneğe sahip olduğunu gösteriyor: Uzmanlaşmış sinir devreleri de lazımdır. Bu gözlem *olanaklı ama gerçekleşmemiş bilişsel yetenekler* fikrini akla getirir: Bu yetenekler gerçek bir insanda yoktur, ama bu yeteneklere sahip olan diğer zekâ sahibi sistemler (insan beyninden daha fazla hesaplama gücüne sahip olmayan sistemler) stratejik açıdan bir sürü ilintili görevi başarma yetenekleri bakımından muazzam bir kazanım elde edeceklerdir.

Dolayısıyla belli bir alana özgü bilişsel kusurlara sahip insan bireyleri ve insan olmayan hayvanları düşünerek, zekânın farklı nitelikleri ile yol açtıkları pratik farklara dair bir fikre sahip olabiliriz. *Homo sapiens* (sözgelimi) karmaşık bilişsel temsilleri kolaylaştıran bilişsel modüllerden mahrum olsaydı, doğayla uyum içinde yaşayan bir diğer maymun türü olabilirdi. Halbuki karmaşık dilsel temsiller oluşturabilmeye benzer bir üstünlük kazandıran birtakım yeni modüller *kazanacak* olsak, süper zekâ sahibi olurduk.

Doğrudan ve dolaylı kavrama gücü

Bu biçimlerden herhangi birinde süper zekâ zamanla diğerlerinden herhangi birini yaratmak için gerekli teknolojiyi geliştirebilir. Dolayısıyla bu üç süper zekâ biçiminin *dolaylı kavrama güçleri* eşittir. Bu anlamda, mevcut insan zekâsının dolaylı kavrama gücü de bizim nihayetinde bir süper zekâ biçimi yaratabileceğimiz varsayımıyla aynı eşdeğer sınıfındadır. Fakat bir anlam var ki süper zekânın üç biçiminin birbirlerine çok daha yakın olduklarını söyleyebiliriz: Bunlardan herhangi biri diğer süper zekâ biçimlerini bizim mevcut başlangıç noktamızdan herhangi bir süper zekâ biçimini yaratabileceğimizden daha hızlı şekilde yaratabilir.

Üç farklı süper zekâ biçiminin *doğrudan kavrama güçlerinin* karşılaştırılmasıysa güçtür. Kesin bir sıralama olmayabilir. Bunların kendi yetenekleri kendi üstünlüklerini hangi ölçüde somutlaştırdıklarına bağlıdır: Bir süper zekâ *ne kadar* güçlüdür, bir nitelikli süper zekâ nitel açıdan *ne kadar* üstündür vesaire. En fazla şunu söyleyebiliriz, diğer şartlar aynı kaldığında, hızlı süper zekâ art arda icra edilmesi gereken çok sayıda adımın hızla atılmasını gerektiren görevlerde yetkinleşirken; kolektif süper zekâ çok farklı perspektiflerin ve yetenek kümelerinin birleştirilmesini talep eden paralelleştirilebilir görevler ve alt-görevler şeklinde analitik ayrışımı kabul eden görevlerde yetkinleşmiştir. Muğlak bir anlamda, nitelikli süper zekâ bunların içinde en ehil olanı olacaktır, zira tüm pratik amaçlar açısından değerlendirildiğinde hızlı süper zekânın ve kolektif süper zekânın *doğrudan* kavrama gücünün ötesinde olan problemleri kavrayıp çözebilir.[14]

Bazı alanlarda, nicelik nitelik için zayıf bir ikamedir. Duvarları mantar kaplamalı bir yatak odasında tek başına çalışan bir dâhi *Kayıp Zamanın İzinde*'yi yazamaz. Edebiyatçı müsveddelerini bir çalışma odasına toplayarak benzer bir şaheser üretilebilir mi?[15] İnsan çeşitliliğinin mevcut aralığında bile bazı işlevlerin çeşitli vasat isimlerin ortak çabasından ziyade tek bir parlak dâhinin çabasından çok büyük yarar elde ettiğini görüyoruz. Eğer görüş açımızı *süper zeki* zihinleri kapsayacak şekilde genişletirsek, sadece süper zekâyla çözülebilir olan ve çabaları birleştirilmemiş insanların büyük bir kolektifi tarafından ele alınamayan entelektüel problemler olduğu ihtimalini kabul etmemiz gerekir.

Dolayısıyla nitelikli bir süper zekâ tarafından ve belki de hızlı bir süper zekâ tarafından çözülebilir olan ama gevşek şekilde bütünleştirilmiş kolektif bir süper zekânın çözemeyeceği bazı problemler olabilir (ilkin kendi zekâsını genişletmesi durumu hariç).[16] Bu problemlerin ne olduğunu açıkça göremeyiz, ama bunları genel hatlarıyla nitelendirebiliriz.[17] Bunlar genellikle bağımsız olarak doğrulanabilir çözüm adımlarına izin vermeyen birden çok karmaşık karşılıklı bağımlılık içeren problemlerdir. Dolayısıyla bu problemler parça parça çözülemez ve niteliksel açıdan yeni türde bir anlayış ya da fanilerin mevcut çeşidinin keşfedemeyeceği ya da etkili şekilde kullanamayacağı kadar derin ya da çetrefilli olan yeni temsil çerçeveleri gerektirir. Bazı sanatsal yaratıcılık

ve stratejik biliş türleri bu kategoriye dahil olabilir. Belki bazı bilimsel atılım türleri için de aynısı geçerlidir. Ayrıca insanlığın felsefedeki "birçok ölümsüz problem"deki ilerleyişinin yavaşlığı ve kararsızlığının insan korteksinin felsefi çalışma için uygun olmamasına bağlı olduğu iddia edilebilir. Bu görüşe göre, bizim en ünlü filozoflarımız arka ayakları üzerinde yürüyen köpekler gibidir: Bir faaliyet göstermek için gerekli performansın eşik düzeyine ancak erişirler.[18]

Dijital zekâ için üstünlük kaynakları

Beyin büyüklüğü ve donanımındaki ufak değişiklikler, insanların entelektüel ve teknolojik başarılarını diğer kuyruksuz maymunlarınkiyle kıyasladığımızda gördüğümüz gibi, büyük sonuçlar doğurabilir. Makine zekânın kolaylaştıracağı bilgi işlem kaynakları ve mimarisindeki çok daha büyük değişikliklerse muhtemelen daha da derin sonuçlar doğuracaktır. Süper zekânın yapabileceklerine dair sezgilerimizle bir fikre ulaşmamız zor, belki de imkânsızdır, ama en azından dijital zihinlere açık olan üstünlüklerden bazılarına bakarak nelerin mümkün olduğuna dair bir ipucu elde edebiliriz. Donanım üstünlüklerini takdir etmek en kolayıdır:

> *Bilgi işlem öğelerinin hızı:* Biyolojik nöronlar en hızlı haliyle yaklaşık 200 Hz'e ulaşırlar ve bu hız modern mikroişlemciden (~ 2 GHz) tam yedi kat daha yavaştır.[19] Dolayısıyla insan beyni devasa bir paralelleştirmeden yararlanmak zorundadır ve çok sayıda art arda işlemi gerektiren hiçbir bilgi işlemi hızlı bir şekilde gerçekleştiremez.[20] (Beynin bir saniyenin altında bir sürede yaptığı hiçbir şey yüzden fazla ardışık işlemden yararlanamaz, hatta belki de bu rakam sadece birkaç düzineden ibarettir.) Ama programlama ve bilgisayar bilimindeki pratikte en önemli algoritmalardan birçoğu kolaylıkla paralelleştirilemez. Beynin paralelleştirilebilir örüntü eşleştirici algoritmalara verdiği destek, hızlı ardışık işlemeye destekle güçlendirilip bütünleştirilseydi birçok bilişsel görev çok daha verimli şekilde yerine getirilebilirdi.
>
> *İç iletişim hızı.* Aksonlar 120 m/s ya da ona yakın bir hızla eylem potansiyelleri taşırlar, oysa elektronik işlemci çekirdekleri optik olarak ışık hızında iletişim kurabilirler (300.000.000 m/s).[21] Nöral sinyallerin uyuşukluğu bir biyolojik beynin tek bir işlemci birim olarak işlerken ne kadar büyük olabileceğini sınırlar. Örneğin bir sistemde herhangi iki öğe arasında 10 milisaniyeden daha az süren bir gidiş-geliş gecikmesi başarmak için biyolojik beyinlerin 0,11 m3'ten daha küçük olması gerekir. Bir elektronik sistem ise 6,1x1017 m3 ile küçük bir gezegen boyutunda olabilir: On sekiz birim kat büyüktür.[22]
>
> *Bilgi işlem öğelerin sayısı.* İnsan beyninde 100 milyardan az nöron vardır.[23] İnsanlarda beyin büyüklüğü şempanzelerinkinin neredeyse üç buçuk katıdır (ispermeçet balinasının beyin büyüklüğününse sadece beşte biridir).[24] Bir biyolojik varlıktaki nöronların sayısı, çok açık ki, kafatası hacmi ve metabolik kısıtlıklarla

sınırlanmaktadır, ama diğer faktörler de büyük beyinler için önemli olabilir: soğu(t)ma, gelişim zamanı ve sinyal iletme gecikmeleri gibi (önceki maddeye bakınız). Halbuki bilgisayar donanımı, çok yüksek fiziksel sınırlara kadar sonsuza dek ölçeklendirilebilir.[25] Süper bilgisayarlar, yüksek hızlı kablolar yoluyla eklenmiş ilave uzaktan erişim gücünde ve bir depo büyüklüğünde, hatta daha büyük olabilir.[26]

Depolama kapasitesi. İnsanın işler belleği verili bir zamanda en fazla dört beş bilgi parçasını tutabilir.[27] İnsanın işler belleğinin büyüklüğünü doğrudan bir dijital bilgisayardaki RAM miktarıyla karşılaştırmak yanıltıcı olsa da, şurası açık ki, dijital zekâlar donanım üstünlüklerinin daha büyük işler belleklere sahip olmalarını mümkün hale getirecektir. Bu sayede bu tür zihinler insanların zahmetli bir hesaplama yoluyla ancak üstünkörü yakalayabildiği karmaşık ilişkileri sezgisel olarak kavrayabilir.[28] İnsanın uzun süreli belleği de sınırlıdır, ama sıradan bir insan yaşamı sırasında depolama kapasitesini tüketmeyi başarıp başaramadığımız net değildir, zira bilgi biriktirme hızımız çok yavaştır. (Bir hesaplamaya göre, yetişkin insan beyni yaklaşık bir milyon parçacık depolayabiliyor: En alt seviye akıllı telefonlarınkinden birkaç kat düşük büyüklük kertesidir bu.[29]) Dolayısıyla gerek depolanan bilgi miktarı gerek bu bilgiye erişilme hızı bir biyolojik beyne kıyasla bir makine beyninde çok çok daha fazla olabilir.

Güvenilirlik, dayanma ömrü, sensörler, vb. Makine zekâların donanım bakımından başka birçok üstünlüğü olabilir. Örneğin biyolojik nöronlar transistörler kadar güvenilir değildir.[30] Gürültülü bilgi işlem tek bir bilgi parçacığını kodlamak için birden fazla öğeden yararlanan gereksiz kodlama şemaları gerektirdiğinden, bir dijital beyin yüksek hassasiyette güvenilir bilgi işlem öğelerinden yararlanarak verimlilik açısından bazı kazanımlar elde edebilir. Beyinler birkaç saat çalışmanın ardından yorulur ve yirmi otuz yıllık öznel zamanın ardından ebediyen çöküşe geçer; mikroişlemciler ise bu sınırlılıklara tabi değildir. Bir makine zekâya veri akışı milyonlarca sensör ekleyerek artırılabilir. Kullanılan teknolojiye bağlı olarak, bir makine değişen görev gerekliliklerine göre optimize edilebilen yeniden ayarlanabilir donanıma sahip olabilirken, beynin mimarisinin çoğu doğuştan sabittir ya da ancak yavaşça değişebilir (ama sinaptik bağlantılılığın ayrıntıları çok kısa zaman aralıklarında, örneğin birkaç günde değişebilir).[31]

Şu anda biyolojik beynin bilgi işlem gücü hâlâ dijital bilgisayarların gücünden üstün konumdadır, ama süper bilgisayarların en iyileri beynin işlemci gücünün akla yatkın tahminleri dahilinde performans düzeylerine erişmektedir.[32] Ama donanım hızla ilerliyor ve donanım performansının nihai sınırları biyolojik bilgi işlem substratlarınınkinden çok daha yüksektir.

Dijital zihinler de yazılımdaki büyük üstünlüklerden fayda sağlayacaklar:

Düzenlenebilirlik: Nöral biyolojik zekâya kıyasla yazılımdaki parametre çeşitlilikleriyle deney yapmak daha kolaydır. Örneğin bir tüm beyin emülasyonunda belli bir korteks alanına daha fazla nöron eklendiğinde veya uyarılabilirliğini

artırdığınızda ya da azalttığınızda ne olduğunu kolaylıkla deneyip görebilirsiniz. Canlı biyolojik beyinlerde bu tür deneyler yapmak ise çok daha zor olacaktır.

Çoğaltılabilirlik: Yazılımda mevcut donanım temelini doldurmak için çok sayıda aslına sadık nüshayı rastgele ve çarçabuk yapabilirsiniz. Halbuki biyolojik beyinler sadece çok yavaş şekilde yeniden üretilebilir ve her yeni örnek umarsız bir durumda başlar ve ebeveynlerinin yaşadıkları süre zarfında öğrendiği hiçbir şeyi hatırlamaz.

Hedef eşgüdümü: İnsan kolektifleri büyük bir grubun fertleri arasında tam bir hedef birliği sağlamanın neredeyse imkânsız olmasından kaynaklanan verimsizliklerle doludur, en azından ilaçlarla ya da genetik seçilimle yumuşakbaşlılık sağlayana kadar bu böyledir. Bir "kopya klan" (ortak bir hedefi olan bir grup benzer ya da neredeyse benzer programlar) bu tür eşgüdüm sorunlarından kaçınabilecektir.

Bellek paylaşımı: Biyolojik beyinler uzun süre eğitime ve danışmanlığa ihtiyaç duyarlar, oysa dijital zihinler veri dosyalarını değiş tokuş ederek yeni bellekler ve yetenekler elde edebilirler. Bir yapay zekâ programının bir milyar nüshalık bir popülasyonu, veritabanlarını düzenli olarak eşzamanlılaştırabilir, böylece programın tüm kerteleri önceki saat sırasında öğrenilen her şeyi bilir. (Doğrudan bellek nakli standartlaştırılmış temsil formatları gerektirir. Bu nedenle yüksek düzey bilişsel içeriğin kolayca değiş tokuş edilmesi herhangi bir çift makine zekâ arasında mümkün olmayacaktır. Daha somutta, birinci kuşak tüm beyin emülasyonları arasında mümkün olmayacaktır.)

Yeni modüller, modaliteler ve algoritmalar: Görsel algının kolay olduğunu ve çaba gerektirmediğini düşünürüz, ders kitaplarındaki geometri problemlerini çözmekten hiç farkı yok gibidir, üstelik retinalarımızdaki ikiboyutlu uyarılma örüntülerinden tanınabilir nesnelerle dolu bir dünyanın üçboyutlu temsilini yeniden inşa etmek epey bir bilgi işlem gerektirmesine rağmen bu böyledir. Bunun kolay görünmesinin nedeni görsel bilgiyi işlemek için düşük düzey nöral mekanizma ayırmış olmamızdır. Bu düşük düzey işleme bilinçsizce ve otomatikman gerçekleşir, herhangi bir zihinsel enerji ya da bilinçli dikkat gerektirmez. Müzik algısı, dil kullanımı, toplumsal biliş ve biz insanlar için "doğal" olan diğer bilgi işleme biçimleri de benzer şekilde adanmış nöral bilgi işlem modüllerince destekleniyor görünmektedir. Günümüz dünyasında önemli hale gelmiş olan diğer bilişsel alanlarda böyle bir uzman desteği sunan yapay bir zihin (mühendislik, bilgisayar programı ve iş stratejisi gibi) bizimki gibi bu tür şeyler hakkında düşünmek için hantal genel amaçlı biliş yetisine yaslanmak zorunda kalan zihinlere karşı büyük üstünlüklere sahip olacaktır. Keza dijital donanımın ayrı kolaylıklarından (hızlı seri işlemeye destek gibi) yararlanmak için yeni algoritmalar da geliştirilebilir.

Makine zekâ, donanım ve yazılımın bir araya gelince *nihayetinde* elde edilebilir olan üstünlükleri muazzamdır.[33] Fakat bu potansiyel üstünlükler ne kadar hızlı şekilde gerçeğe dönüştürülebilir? Şimdi bu soruyu ele almamız gerekiyor.

Bir Zekâ Patlamasının Kinetiği

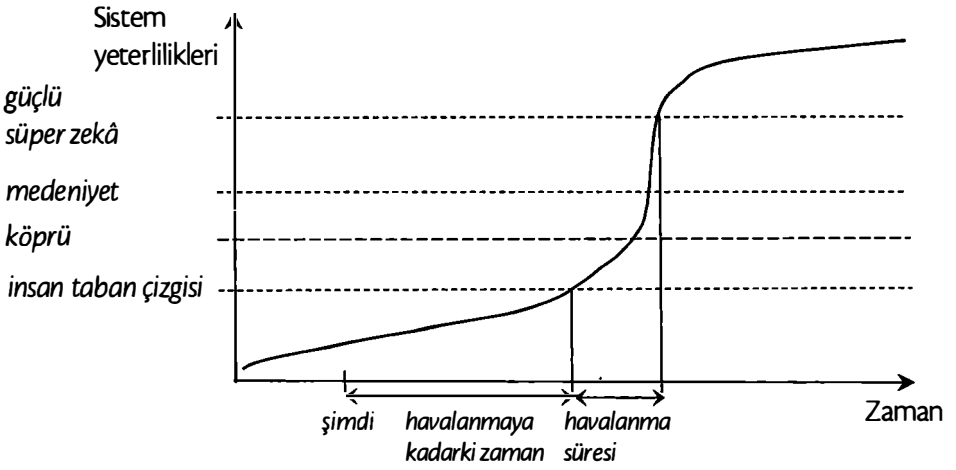
Makineler bir kez genel muhakeme yeteneğinde insana denk bir seviyeye ulaştıklarında, radikal süper zekâya erişmeleri ne kadar zaman alacaktır? Bu yavaş, tedrici, uzatmalı bir geçiş mi olacaktır, yoksa ani ve patlayıcı mı? Bu bölümde optimizasyon gücü ve sistemin uzlaşmazlığının bir işlevi olarak süper zekâya geçişin kinetiğini analiz ediyoruz. İnsan düzeyinde genel zekânın havzasındaki bu iki faktör hakkında bildiklerimizi ya da makul düzeyde varsayabileceklerimizi ele alıyoruz.

Havalanmanın hızı ve zamanlaması

Makinelerin genel zekâ bakımından biyolojiyi *nihayetinde* misli misli aşacağı ama makine bilişinin *şu anda* insan bilişinden çok daha dar olduğu düşünüldüğünde, bu gaspın ne kadar çabuk gerçekleşeceğini insan merak etmeden edemiyor. Burada sorduğumuz soru 1. Bölüm'de ele aldığımız insan düzeyinde genel zekâya sahip bir makine geliştirmekten şu an ne kadar uzak olduğumuz sorusundan kesinkes ayrılmalıdır. Burada soru *böyle bir makinenin geliştirilip geliştirilmediği, geliştirilirse ne zaman geliştirileceği, akabinde bir makine köklü derecede süper zeki hale gelene kadar ne kadar zaman geçeceğidir*. Makineler insanın taban çizgisine ulaşana kadar uzun bir zaman geçeceğini düşünebiliriz ya da bunun ne kadar zaman alacağı bilinemez diyebiliriz, ama bu bir kez gerçekleştiğinde, güçlü süper zekâya doğru yükselişin çok hızlı olacağını belirtelim.

Bu meseleleri şematik şekilde düşünmek yararlı olabilir, her ne kadar bunu yapmak bazı nitelikleri ve karmaşıklaştırıcı ayrıntıları bir süre göz ardı etmek anlamına gelse de. O halde, en ileri makine zekâ sisteminin entelektüel yeteneğini zamanın bir işlevi olarak kurgulayan bir grafik düşünün (**ŞEKİL 7**).

"İnsan taban çizgisi" adı verilen yatay çizgi şu an gelişmiş ülkelerde var olan bilgi kaynaklarına ve teknolojik yardımlara erişim olanağı olan temsili nitelikte bir yetişkin insanın etkin entelektüel yeteneklerini temsil eder. Şu anda en ileri yapay zekâ sistemi, genel entelektüel yeteneğin herhangi bir makul ölçüsü baz alındığında, insanın taban çizgisinin çok aşağısındadır. İleride belli bir noktada, bir makine bu insani taban çizgisiyle az çok denkliğe kavuşabilir (biz bunun sabit olduğunu, insan bireylerin yetenekleri geçen yıllarda artmış olsa bile sözgelimi 2014 yılına sabitlendiğini varsayıyoruz):



ŞEKİL 7 Havalanmanın şekli. Şu sorular arasında ayrım yapmakta fayda var: "Bir havalanma gerçekleşecek mi, gerçekleşecekse ne zaman?" ile "Bir havalanma gerçekleşirse ve gerçekleştiğinde, ne kadar yükseğe olacak?" Örneğin bir havalanma gerçekleşene kadar çok uzun zaman geçeceğini ama bir kez gerçekleştiğinde çok çabuk olacağını varsayabiliriz. Bir diğer anlamlı soruysa (yukarıdaki şekilde yer almasa da) şudur: "Dünya ekonomisinin ne kadarlık bir kısmı bu havalanmada yerini alacak?" Bu sorular birbiriyle ilintili ama ayrı sorulardır.

Havalanma bu anda başlayacaktır. Sistemin yetenekleri artmaya devam eder ve ileride bir noktada sistem tüm insanlığın birleşik entelektüel yeteneğiyle eşitliğe kavuşur (yine şimdiye demir atar): Buna "medeniyet taban çizgisi" adını verebiliriz. Nihayetinde, eğer sistemin yetenekleri artmaya devam ederse, "güçlü süper zekâ"ya ulaşır: Günümüz insanlığının birleşik entelektüel donanımından kıyaslanamaz derecede daha büyük bir zekâ düzeyidir bu. Güçlü süper zekâya erişilmesi havalanmanın tamamlandığı anlamına gelir, ama sistem bundan sonra da kapasite bakımından büyümeye devam edebilir. Sistem havalanma aşaması sırasında bir ara "köprü" adını verebileceğimiz bir dönüm noktasından geçebilir; bu noktanın ötesinde sistemin daha fazla ilerlemesi başkalarının sisteme müdahalelerinden ziyade esasen sistemin kendi eylemlerinden güdülenir.[1] (Bu tür bir köprü ihtimali bir sonraki bölümde optimizasyon gücü ve patlayıcılık üzerine alt kısımda önemli hale gelecektir.)

Bu tabloyu aklımızda tutarak, üç geçiş senaryosu türünü (sistemlerin insan düzeyinde zekâdan süper zekâya ilerledikleri senaryoları) irtifa düzeyine bağlı olarak, yani havalanmanın yavaş mı, hızlı mı yoksa orta karar mı olduğuna bakarak birbirinden ayırt edebiliriz.

Yavaş

Yavaş bir havalanma uzun bir zaman aralığında, mesela on ya da yüzyıllar içinde gerçekleşir. Yavaş havalanma senaryoları insani siyasal süreçlerin uyum sağlayıp yanıt vermeleri için mükemmel fırsatlar sunar. Farklı yaklaşımlar bu süreçte denenip

sınanabilir. Yeni uzmanlar eğitilip uzman hale getirilebilir. Bu yeni gelişmelerden olumsuz etkilendiğini hisseden gruplar tarafından tabandan hareketler örgütlenebilir. Eğer güvenli altyapının yeni türlerine ya da yapay zekâ araştırmacılarının kitlesel olarak gözetlenmesine ihtiyaç olduğu gözükürse, bu tür sistemler geliştirilip kullanılabilir. Yapay zekâ silahlanma yarışından korkan devletler anlaşmalar için müzakere yürütme ve yaptırım mekanizmaları tasarlama zamanı bulurlar. Yavaş havalanmanın başlamasından önce girişilen çoğu hazırlık, ufukta gözüken çağın ışığında yeni çözümler peyderpey görünür hale geldikçe ıskartaya çıkacaktır.

Hızlı

Hızlı havalanma, örneğin dakikalar, saatler ya da günlerle ölçülen kısa bir zaman aralığında gerçekleşir. Hızlı havalanma senaryoları insanlar için çok az müzakere fırsatı sunar. Birileri olağandışı bir şeyler fark edene kadar oyun zaten çoktan kaybedilmiştir. Hızlı bir havalanma senaryosunda insanlığın kaderi esasen daha öncesinde yapılmış hazırlıklara bağlıdır. Hızlı havalanma senaryosunun en yavaş ucunda, "nükleer çanta"yı açmaya benzer bazı basit insani eylemler mümkün olabilir; ama bu tür eylemler ya temel düzeyde ya da önceden planlanıp programlanmış olacaktır.

Orta karar

Orta karar bir havalanma ise aylar ya da yıllarla ölçülebilecek ortalama bir süre zarfında gerçekleşir. Orta karar havalanma senaryoları insanlara yanıt vermeleri için bir şans sunar, ama durumu analiz etmek, farklı yaklaşımları sınamak ya da karmaşık eşgüdüm problemlerini çözmek için fazla zaman tanımaz. Yeni sistemleri (örneğin siyasal sistemler, gözetleme rejimleri ya da bilgisayar ağı güvenlik kurallarını) geliştirmek ya da kullanmak için yeterli zaman yoktur, ama bugüne kadar gelen sistemler yeni zorlukları aşmak için kullanılabilir.

Yavaş bir havalanma sırasında haberlerin dışarıya ulaşması için epey zaman olacaktır. Halbuki orta karar havalanmada gelişmelerin o sırada gizli tutulması mümkündür. Bilgi, devlet desteğiyle yürütülen gizli bir askeri araştırma programında olduğu gibi küçük bir grupla sınırlı tutulabilir. Ticari projeler, küçük akademik takımlar ve "bir bodrum katındaki dokuz *hacker*" da yine gizli olabilir, ama eğer bir zekâ patlaması olasılığı bir ulusal güvenlik önceliği olarak devlet istihbarat kurumlarının "radarında" ise, o zaman en fazla ümit vaat eden özel teşebbüs projelerinin gizlice izleniyor olması ihtimali bir hayli yüksek olacaktır. Ev sahibi devlet (ya da egemen bir dış güç) havalanmayı başlatma emareleri gösteren her projeyi kamulaştırma ya da sonlandırma seçeneğine sahip olacaktır. Hızlı havalanmalar ise o kadar çabuk olur ki haberin yayılması için ya da birilerinin anlamlı bir tepki göstermesi için pek zaman kalmaz. Ama belli bir projenin başarıya yaklaştığına inanan dışarıdan biri havalanma gerçekleşmeden *önce* müdahale edebilir.

Orta karar havalanma senaryoları, bireyler ve gruplar yaklaşan dönüşümden kendilerine pay almak için yarıştıklarından jeopolitik, toplumsal ve ekonomik çalkantıya yol açabilir. Bu tür bir altüst oluş gerçekleştiği takdirde, örgütlü bir yanıt verme çabalarını sekteye uğratabilir; alternatif olarak, sakin bir ortamda üretilecek çözümlerden daha radikal çözümler doğurabilir. Sözgelimi ucuz ve ehil emülasyonlar ya da diğer dijital zihinlerin emek piyasalarını tedricen dolduracağı orta karar bir havalanma senaryosunda, işten çıkarılan işçilerin düzenlediği kitle gösterilerinin hükümetlere işsizlik maaşlarını artırma ya da tüm yurttaşlara yaşanabilir bir ücret garanti etme veya emülasyon işçi çalıştıran patronlara özel vergiler koyması ya da asgari ücret şartı koyması için baskı uygulayacağını düşünebiliriz. Bu tür politikalardan elde edilecek yararların kalıcı olması için, verilecek desteğin kalıcı iktidar yapılarının parçası haline getirilmesi gerekir. Eğer havalanma orta düzey yerine yavaşsa da benzer meseleler doğabilir, ama ılımlı, orta karar senaryolarda hızlı değişim ve dengelerin altüst olması küçük grupların orantısız bir nüfuz elde etmesi için özel fırsatlar sunabilir.

Bazı okurlar bu üç senaryo türünden yavaş havalanmanın en muhtemel, orta karar havalanmanın en az muhtemel, hızlı havalanmanın ise tamamen olasılık dışı olduğunu düşünebilir. İnsanlığın en üstteki düşünen varlık olarak konumunu kaybedip dünyanın kökten dönüşmesinin bir iki saat alacağını varsaymak hayal gibi gelebilir. İnsanlık tarihinde böyle bir değişim örneği yoktur ve en yakın örnek olan Sanayi Devrimi ve Tarım Devrimi çok daha uzun bir zaman ölçeğinde (Tarım Devrimi'nde yüzyıllar ila binyıllar, Sanayi Devrimi'nde onyıllar ila yüzyıllar) gerçekleşmiştir. Dolayısıyla hızlı ya da orta karar bir havalanma senaryosunun gerektirdiği geçiş türünün gerçekleşme ihtimali, koyutlanmış değişimin hızı ve büyüklüğü bakımından, sıfırdır: Ancak mitlerde ve dinlerde örneği görülebilir.[2]

Ne var ki bu bölümde yavaş geçiş senaryosunun imkânsız olduğunu düşünmek için bazı gerekçeler sunulacaktır. Bir havalanma gerçekleşirse ve gerçekleştiğinde, muhtemelen patlayıcı olacaktır.

Havalanmanın ne kadar hızlı olacağı sorununu analiz etmeye başlamadan önce, bir sistemin zekâsındaki artış hızını iki sistemin (yeknesak artan) işlevi olarak tasavvur edebiliriz. Sistemin zekâsını artırmak için uygulanan "optimizasyon gücü" miktarı ya da nitelikli tasarım çabası ve sistemin bu türden optimizasyon gücünün verili bir miktarına uygulanmasına yanıt verirliği. Yanıt verirliğinin tersine "uzlaşmazlık" adını verip şöyle yazabiliriz:

$$\text{Zekâdaki değişim hızı} = \frac{\text{Optimizasyon gücü}}{\text{uzlaşmazlık}}$$

Zekânın, tasarım çabasının ve uzlaşmazlığın nasıl nicellendirileceğini somutlaştırmayı beklerken, bu ifade salt niteldir. Ama en azından, eğer bir sistemin zekâsının

artırılması görevine çok fazla yetenekli çaba uygulanırsa ve sistemin zekâsı artamayacak kadar katı değilse *ya da* entipüften olmayan bir tasarım çabası varsa ve sistemin uzlaşmazlığı fazla değilse (yahut hem öyle hem böyleyse) o sistemin zekâsının hızla artacağını gözlemleyebiliriz. Eğer belli bir sistemin ilerletilmesi için ne kadar tasarım çabası gerektiğini ve bu çabanın ürettiği ilerleme hızını biliyorsak, sistemin uzlaşmazlığını hesaplayabiliriz.

Ayrıca bir sistemin performansını iyileştirmeye yönelik optimizasyon gücü miktarının sistemler arasında ve zamanla değiştiği gözleminde bulunabiliriz. Bir sistemin uzlaşmazlığı sistemin ne kadarının optimize edildiğine bağlı olarak da değişebilir. Çoğu zaman, en kolay iyileştirmeler önce yapılır, bu yüzden çok az kazanım elde edilir (uzlaşmazlık artar) ve aşağıda sallanan meyveler tüketilir. Gelgelelim başka ilerlemeleri kolaylaştırıp ilerlemelerin sel olup taşmasına yol açacak ilerlemeler de olabilir. Bir yapbozu çözme süreci basitten başlar; köşeleri ve kıyıları bulmak kolaydır. Ardından uzlaşmazlık sonraki parçaların uydurulması zor olduğundan artar. Ama bulmaca sona yaklaştıkça, arama alanı azalır ve süreç yeniden kolaylaşır.

Demek ki araştırmamıza devam etmek için uzlaşmazlığın ve optimizasyon sürecinin havalanma sırasındaki kritik zaman dilimlerinde nasıl değişebileceğini analiz etmemiz gerekiyor.

Uzlaşmazlık

Uzlaşmazlıktan başlayalım. Burada bakış açısı ele alınan sistem türüne bağlıdır. Eksik olmaması adına, öncelikle, ileri makine zekâ içermeyen süper zekâ yollarında karşılaşılacak uzlaşmazlığa bakıyoruz. Bu güzergâhlarda uzlaşmazlığın son derece yüksek olduğunu görürüz. Bunu yaptıktan sonra, ana meseleye (yani havalanma makine zekâ da içerir) geri döneceğiz ve orada kritik dönemeçte uzlaşmazlığın az göründüğünü tespit edeceğiz.

Makine olmayan zekâ yolları

Kamu sağlığı ve beslenmesindeki iyileştirmeler yoluyla bilişsel gelişimin muazzam azalan getirileri vardır.[3] Büyük kazanımlar ciddi beslenme yetersizliklerini ortadan kaldırarak elde edilir ve en ciddi yetersizlikler bugün en yoksul ülkeler haricinde büyük oranda ortadan kaldırılmıştır. Zaten yeterli bir beslenmeyi artırarak ancak kemerin bir delik daha eksik sıkılması sağlanır. Eğitim de şu anda muhtemelen getirileri azalan bir şeydir. Dünyada nitelikli eğitime erişim olanağı olmayan yetenekli bireylerin sayısı hâlâ kayda değer olsa da, gerilemektedir.

Farmakolojik güçlendiriciler önümüzdeki yirmi otuz sene içinde bazı bilişsel kazanımlar elde edilmesini sağlayabilir. Ama en kolay düzeltmeler yapıldıktan sonra –belki de zihinsel enerji ve konsantrasyon yeteneğinde sürdürülebilir artışların yanı sıra, uzun

süreli bellek konsolidasyonun hızı üzerinde daha iyi bir kontrol sağlanmasıyla– yeni kazanımlar elde edilmesi gittikçe güçleşecektir. Gelgelelim beslenme ve kamu sağlığı konusunda yaklaşımların aksine, biliş yetisini akıllı ilaçlar yoluyla ilerletmek iyice zorlaşmadan önce kolaylaşabilir. Nörofarmakoloji alanı sağlıklı beyne ehil bir şekilde müdahale etmek için gerekli temel bilgiden hâlâ yoksundur. Meşru bir araştırma alanı olarak geliştirici ilacın savsaklanması şu anki geri kalmışlığın sebeplerinden biri olabilir. Eğer nörobilim ve farmakoloji bilişsel geliştirmeye odaklanmadan bir süre daha ilerlemeye devam ederse, o zaman belki de en sonunda nootropiklerin gelişimi ciddi bir öncelik haline geldiğinde bazı görece kolay kazanımlar elde edilebilir.[4]

Genetik bilişsel geliştirme tıpkı nootropikteki gibi U şeklinde bir uzlaşmazlık profiline sahiptir, ama potansiyel kazanımları daha fazladır. Mevcut yegâne yöntem kuşaklar boyunca sürdürülen seçici yetiştirmeyken (küresel açıdan kayda değer bir ölçekte gerçekleştirmesi bariz bir zorluk taşır) uzlaşmazlık fazladır. Ucuz ve etkili genetik test ve seçim için teknoloji geliştirildikçe (özellikle de yinelemeli embriyo seçilimi insanlarda gerçekleştirilebilir hale geldiğinde) genetik geliştirme kolaylaşacaktır. Bu yeni teknikler mevcut insan genetik çeşitliliği havuzunu zekâyı artırıcı aleller için kullanmayı mümkün kılacaktır. Ne var ki var olan en iyi aleller genetik geliştirme paketlerine dahil edildikçe, yeni kazanımlar elde etmek güçleşecektir. O zaman, genetik değişiklik için çok daha yenilikçi yaklaşımlara olan ihtiyaç uzlaşmazlığı artırır. Genetik geliştirme yolunda işlerin ne kadar çabuk ilerleyeceğinin sınırları var, özellikle de germ hattı müdahalelerinin kaçınılmaz bir olgunlaşma gecikmesine tabi olmasından bahsediyoruz: Bu durum hızlı ya da orta karar bir havalanmanın önünde ciddi bir engeldir.[5] Embriyo seçiliminin ancak laboratuvarda dölleme bağlamında uygulanabilir olması benimsenme hızını yavaşlatacaktır: Bu da bir diğer sınırlayıcı etkendir.

Beyin-bilgisayar yolundaki uzlaşmazlık ilk başta çok yüksek görünür. Olur da beyin implantları yapmak ve korteksle yüksek düzey işlevsel bütünleşme sağlamak bir şekilde mümkün olursa, uzlaşmazlık dibi bulabilir. Uzun vadede, bu yolda ilerleme kaydetmenin zorluğu, emülasyonları ya da yapay zekâları iyileştirmenin içerdiği zorluklara benzer olacaktır, zira beyin-bilgisayar sisteminin zekâsının çoğu nihayetinde bilgisayar kısmında bulunur.

Ağları ve örgütlenmeleri genel olarak daha verimli hale getirme konusunda uzlaşmazlık çok fazladır. Bu uzlaşmazlığı aşmak için epey bir çaba harcanıyor, ama sonuçta insanlığın toplam kapasitesinin yıllık ilerlemesi ancak yüzde bir ikidir.[6] Ayrıca iç ve dış çevredeki değişiklikler örgütlerin, belli bir dönem boyunca etkili olsalar bile, çok geçmeden yeni çevrelerine kötü uyum sağladıkları anlamına geliyor. Dolayısıyla sadece kötüleşmeyi engellemek için bile reform çabalarının sürmesi şarttır. Ortalama örgütlenme verimliliğindeki kazanım hızında bir adımlık ilerleme belki tasavvur edilebilir, ama bu türde en radikal senaryonun bile yavaş bir havalanmadan daha fazlasını nasıl

başaracağını tasavvur etmek güçtür, zira insanların işlettiği örgütlenmeler insani zaman ölçeklerinde çalışacak şekilde sınırlıdır. İnternet kolektif zekâyı geliştirmek için birçok fırsat sunan heyecan verici bir alandır ve şu anda uzlaşmazlık orta düzeyde görünmektedir; ilerleme biraz hızlıdır, ama çabaların çoğu bu ilerlemenin gerçekleşmesi için harcanmaktadır. Aşağıda sallanan meyveler (arama motorları ve e-posta gibi) tüketildikçe bunun artması beklenebilir.

Emülasyon ve yapay zekâ yolları

Tüm beyin emülasyonuna ilerlemenin zorluğunu tahmin etmek güçtür. Fakat gelecekte somut bir dönüm noktası olduğundan bahsedebiliriz: Bir böcek beyninin başarılı emülasyonu. Bu dönüm noktası bir tepenin zirvesindedir ve fethedilmesi önümüzde uzanan toprakların çoğunu görünür kılacak ve teknolojiyi insanda tüm beyin emülasyonu düzeyine yükseltmenin uzlaşmazlığı konusunda düzgün bir tahminde bulunmamıza imkân sağlayacaktır. (Fare gibi küçük bir memelinin beyninin başarıyla emülasyonu insanda tüm beyin emülasyonunun çok daha yüksek bir kesinlikle tahmin edilmesi için daha da iyi bir bakış açısı sağlar.) Halbuki yapay zekâya giden yol böyle bariz bir dönemeç ya da erken bir gözlem noktası içermeyebilir. Yapay zekâ arayışının, beklenmedik bir atılım sadece birkaç küçük adım ileride uzanan bir açıklıktaki bitiş çizgisini açığa serene kadar ağaçlarla dolu bir ormanda kaybolmuş gözükecektir.

Şu iki soru arasındaki ayrımı hatırlayın: Kabaca insan düzeyinde bilişsel yeteneğe erişmek ne kadar zordur? Ve oradan insanüstü düzeylere ulaşmak ne kadar zordur? İlk soru esasen bir havalanma başlayana kadar ne kadar zaman geçeceğidir. Havalanmanın şeklini değerlendirmek için kilit olan ise ikinci sorudur ve buradaki amacımız da budur. İnsan düzeyinden insanüstü düzeye erişmenin en zor adım olduğunu varsaymak çekici olsa da –nihayetinde bu adım zaten gayet ehil bir sisteme kapasite eklemek zorunda olduğumuz "bir yüksek rakımda" gerçekleşir– bu hiç de güvenlikli bir varsayım olmayacaktır. Bir makine insanla denkliğe ulaştığında uzlaşmazlığın *kaybolması* gayet mümkündür.

İlk tüm insan emülasyonunu ele alalım. İlk insan emülasyonunun yaratılmasındaki zorluklar mevcut bir emülasyonun geliştirilmesinin içerdiklerinden çok farklı türdedir. Bir ilk emülasyonun yaratılması devasa teknolojik zorluklar içerir, özellikle de gerekli tarama ve görüntü yorumlama yeteneklerinin geliştirilmesi söz konusuyken. Bu adım kayda değer miktarda fiziki sermaye de gerektirebilir; sınai ölçekte bir makine parkında yüzlerce yüksek giripçıktılı tarama makineleri olasılıksız değildir. Halbuki mevcut bir emülasyonun niteliğini artırmak ince ayarlı algoritmalar ve veri yapıları içerir: Özünde bir yazılım programıdır ve özgün şablonu yaratmak için gerekli görüntüleme teknolojisini kusursuzlaştırmaktan çok daha kolay olduğu anlaşılabilir. Programcılar farklı korteks alanlarındaki nöron sayısını artırmak gibi numaralarla kolaylıkla deneyler yapıp

bunun performansı nasıl etkilediğini görebilirler.[7] Ayrıca tekil nöronların ya da küçük nöron ağlarının asli işlevselliğini muhafaza eden kod optimizasyonu ve basit bilgi işlem modelleri bulma üzerinde çalışabilirler. Eğer son teknolojik öngereklilik tarama ya da tercümeyse ve bilgi işlem gücü görece bolsa, o zaman gelişim aşamasında uygulama verimliliğine çok fazla dikkat gösterilmemiş olabilir ve bilgi işlem verimliliği için kolay fırsatlar mevcut olabilir. (Daha temel mimari yeniden örgütleme de mümkün olabilir, ama bu bizi emülasyon yolundan çıkarıp yapay zekâ alanına iter.)

İlk taklit bir kez üretildiğinde kod tabanını iyileştirmenin bir diğer yolu farklı ya da üstün beceri ve yeteneklerle taramaktır. Ayrıca örgütsel yapıları ve iş akışlarını dijital zihinlerin benzersiz özelliklerine uyarlamanın bir sonucu olarak üretkenlik artışı da gerçekleşir. İnsan ekonomisinde birebir kopyalanıp, resetlenip, farklı hızlarda çalıştırılabilen vb. bir işçi önceli olmadığından, ilk emülasyon grubunun yöneticileri yöneticilik pratiklerinde yenilik için fazlaca alan bulacaklardır.

İnsanda tüm beyin emülasyonu mümkün hale geldiğinde, uzlaşmazlık ilk başta azalsa da, yeniden yükselebilir. Er geç uygulamada en çok göze batan verimsizlikler optimum düzeyde azaltılacak, en ümit vaat eden algoritmik çeşitlilikler test edilmiş olacak ve örgütsel yenilikler için en kolay fırsatlar sömürülmüş olacaktır. Şablon kütüphane genişlemiş olacak, böylece daha fazla beyin taraması elde etmek mevcut şablonlarla çalışmaya çok az yarar sağlayacaktır. Bir şablon çoğaltılabildiğinden, her bir kopya farklı bir alanda tekil olarak eğitilebilir ve bu elektronik hızında yapılabildiğinden, potansiyel ekonomik kazanımların çoğunu elde etmek için taranması gereken beyin sayısının az olduğu söylenebilir. Belki de tek bir beyin yetecektir.

Uzlaşmazlığın artmasının bir diğer potansiyel sebebi de taklitlerin ya da biyolojik destekçilerinin emülasyon işçilerinden yararlanmayı sınırlayan, emülasyon kopyalamayı kısıtlayan, dijital zihinlerle belli türde deneyleri yasaklayan, emülasyonlar için asgari ücret ve işçi haklarını kabul ettiren vb. düzenlemeleri desteklemek için örgütleneceklerdir. Gelgelelim siyasal gelişmelerin aksi istikamette ilerlemesi ve uzlaşmazlığın düşmesine katkıda bulunması da mümkündür. Eğer emülasyon kullanımındaki ilk kısıtlamalar rekabetin kızışması ve ahlaki açıdan çıtanın yüksek tutulmasının ekonomik ve stratejik maliyetleri açık hale getirmesi sonucunda yerini dizginsiz bir sömürüye bırakırsa, bu olabilir.

Yapay zekâya (taklit olmayan makine zekâsına) gelince, bir sistemi insan düzeyinden algoritmik iyileştirmeler yoluyla insanüstü zekâya çıkarmanın zorluğu tikel bir sistemin özelliklerine bağlıdır. Farklı mimariler çok zor bir uzlaşmazlığa sahip olabilir.

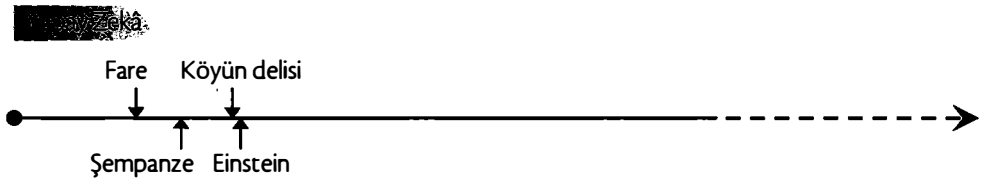
Bazı durumlarda uzlaşmazlık son derece düşük olabilir. Örneğin eğer insan düzeyinde yapay zekâ tek bir kilit içgörü programcılar için uzun süre sır olarak kaldığı için gecikirse, o zaman son atılım gerçekleştirildiğinde, yapay zekâ en alttan zıplayıp aradaki

basamaklara hiç değmeden insan düzeyinin epey üstüne çıkabilir. Uzlaşmazlığın son derece düşük olduğu bir diğer durum iki farklı işleme tarzı yoluyla zeki yeteneklilik kazanabilen bir yapay zekâ sisteminin durumudur. Bu olanağı örneklendirmek için, bir yapay zekânın iki altsistemden oluştuğunu varsayalım: Biri alana özgü problem çözücü tekniklere sahip, diğeri genel amaçlı muhakeme yeteneğine sahip olsun. O zaman ikinci altsistem belki bir kapasite eşiğinin altında kalırken, sistemin genel performansına hiçbir katkı sunmaması mümkün olabilir, çünkü yarattığı çözümler alana özgü sistemin yarattıklarından her zaman daha geridir. Şimdi, küçük miktarda bir optimizasyon gücünün genel amaçlı altsisteme uygulandığını ve bunun sonucunda söz konusu altsistemin kapasitesinde ciddi bir artış yaşandığını varsayalım. İlkin, sistemin genel performansında hiçbir artış hissetmeyiz ve dolayısıyla uzlaşmazlık fazla demektir. Ardından, genel amaçlı altsistemin kapasitesi çözümlerin alana özgü altsisteminkileri yenmeye başladığı eşiği geçtiğinde, genel sistem performansı birdenbire genel amaçlı altsistemle aynı ciddi hızda gelişmeye başlar, üstelik uygulanan optimizasyon miktarı aynı kalır: Sistemin uzlaşmazlığı dibe vurmuştur.

Keza zekâyı insan merkezci bir perspektiften görme konusundaki doğal eğilimimiz nedeniyle insanaltı sistemlerdeki ilerlemelere hak ettiği önemi vermememiz ve dolayısıyla uzlaşmazlığını abartmamız da mümkündür. Makine zekâsının geleceği üzerine epey yazıp çizmiş bir yapay zekâ teorisyeni olan Eliezer Yudkowsky meseleyi şöyle ifade eder (ve bkz. **ŞEKİL 8**):

> Yapay zekâ salt insan merkezciliğin bir sonucu olarak zekâ bakımından görünüşte keskin bir yükseliş yaşar: İnsanların "köyün delisi"ni ve "Einstein"ı genel olarak zihinler ölçeğinde neredeyse ayırt edilemez noktalar yerine zekâ ölçeğinin iki ucu olarak görme eğiliminden bahsediyorum. Bir aptal insandan daha aptal olan her şey bize salt "aptal" görünebilir. Zekâ skalasında fareleri ve şempanzeleri geçerek yavaş yavaş yukarı tırmanan bir "yapay zekâ oku" olduğunu ama yapay zekaların hâlâ "aptal" kaldığını tahayyül ediyoruz çünkü yapay zekâlar akıcı bir şekilde konuşamıyorlar ya da bilimsel makale yazamıyorlar ve ardından yapay zekâ okları aptallığın dik âlâsı düzeyden Einstein ötesi düzeye bir ayda ya da benzer bir kısa sürede küçük bir sıçrama gerçekleştiriyor.[8]

Bu mülahazalardan çıkarılacak sonuç şudur: Kabaca insan düzeyinde genel zekâya erişen ilk yapay zekâda algoritmik ilerlemeler gerçekleştirmenin ne kadar zor olacağını tahmin etmek güçtür. Algoritma uzlaşmazlığının düşük olduğu en azından bazı koşullar bulunabilir. Ama algoritma uzlaşmazlığı çok yüksek olsa bile, söz konusu yapay zekânın genel uzlaşmazlığının çok düşük olmasına engel teşkil etmez. Zira sistemin zekâsını algoritmalarını ilerletmekten başka yollarla artırmak kolay olabilir. Geliştirilebilecek iki etken daha vardır: İçerik ve donanım.



ŞEKİL 8: Daha az insan merkezci bir ölçek mi? Bir aptal ile akıllı insan arasındaki boşluk insan merkezci açıdan bakıldığında büyük gözükebilir, ama at gözlüklerini çıkarınca ikisinin neredeyse ayırt edilemez zihinleri vardır.* Bir köyün delisininkiyle kıyaslanabilecek genel akıllılık düzeyine sahip bir makine zekâ yaratmanın, böyle bir sistemi herhangi bir insandan çok daha akıllı hale gelmesini sağlayacak kadar ilerletmekten daha zor olup daha uzun süreceği neredeyse kesindir.

* Resimde basitleştirmek adına bilişsel yetenek tek boyutlu bir parametre olarak gösteriliyor. Ama burada anlatılan husus açısından asli değildir. Örneğin bir bilişsel yetenek profilini çokboyutlu uzamda bir hiperyüzey olarak temsil etmek mümkündür.

Birincisi, içerik geliştirmelerini düşünün. "İçerik" derken, burada, bir sistemin yazılım değerlerinden çekirdek algoritimik mimarisini oluşturmayan parçaları kastediyoruz. Örneğin içerik depolanmış algıların veritabanlarını, uzmanlaşmış yetenek kütüphanelerini ve bildirime dayalı bilgi envanterlerini içerebilir. Birçok sistem türü için algoritmik mimari ile içerik arasındaki ayrım çok keskin değildir; gelgelelim, bir makine zekâdaki potansiyel açıdan önemli bir yetenek kaynağına işaret etmenin kaba ama işe yarar bir yolu olarak işlev görecektir. Hemen hemen aynı fikri ifade etmenin alternatif bir yolu da bir sistemin entelektüel problem çözme kapasitesinin sadece sistemi daha akıllı hale getirerek değil, aynı zamanda sistemin bildiklerini genişleterek artırılabileceğini söylemektir.

TextRunner (Washington Üniversitesi'ndeki bir araştırma projesi) ya da IBM's Watson (*Jeopardy!* bilgi yarışmasını kazanan sistem) gibi günümüz yapay zekâ sistemlerinden birini ele alalım. Bu sistemler metin analiz ederek belli semantik bilgi parçalarını çıkarabilirler. Bu sistemler bir insanla aynı anlamda ya da aynı ölçüde okuduklarını anlamasalar da, doğal dilden önemli miktarlarda bilgi çıkarabilir ve bu bilgiyi basit çıkarımlarda bulunup soruları cevaplamak için kullanabilirler. Ayrıca deneyimlerinden dersler çıkarıp kullanımına dair yeni örneklerle karşılaştıkça kavramın daha yaygın temsillerini yaratabilirler. Bunlar çoğu zaman denetlenmemiş tarzda işlemek (yani hata ya da ödül sinyalinin yokluğunda etiketsiz verilerdeki saklı yapıyı insan rehberliği olmadan öğrenmek) ve hızlı ve ölçeklenebilir olmak üzere tasarlanmıştır. Sözgelimi TextRunner 500 milyon web sayfalık bir korpus ile işler.[9]

Şimdi, on yaşında bir çocuk kadar anlama yetisiyle ama TextRunner'ınkine benzer bir okuma hızıyla okumayı edinmiş olan bu tür bir sistemin uzak bir atasını tahayyül edin. (Bu muhtemelen zor bir yapay zekâ problemidir.) Dolayısıyla çok daha hızlı düşünen ve bir yetişkin insandan çok daha iyi bir belleği olan, ama çok daha az şey bilen bir sistem tahayyül ediyoruz; bunun net etkisi belki de sistemin genel problem çözme

kapasitesi bakımından kabaca insana denk olmasıdır. Ama içerik uzlaşmazlığı çok düşüktür: Bir havalanmayı hızlandıracak kadar düşüktür. Sistem birkaç hafta içinde Library of Congress'te yer alan tüm içeriği okuyup hatmetmiştir. Sistem bundan böyle her insandan daha çok şey bilir ve kıyaslanamayacak derecede daha hızlı düşünür: En azından zayıf anlamda süper zeki olmuştur.

Dolayısıyla bir sistem geçerli entelektüel yeteneklerini insan bilimi ve medeniyetinde asırlardır birikmiş olan önceden üretilmiş içeriği kapsayarak (sözgelimi internettekileri okuyarak) ciddi derecede artırabilir. Eğer bir yapay zekâ bu materyale öncesinde erişim imkânı olmadan ya da onu özümseyemeden insan düzeyine ulaşırsa, o zaman yapay zekânın genel uzlaşmazlığı –algoritmik mimarisini ilerletmek zor olsa bile– düşük olacaktır.

İçerik uzlaşmazlığı emülasyonlar için de ilintili bir kavramdır. Yüksek hızlı bir emülasyonun sadece biyolojik insanlarla aynı görevleri daha çabuk bir şekilde tamamlayabildiği için değil, aynı zamanda daha vaktinde içerik (görevle ilintili yetenekler ve uzmanlık gibi) biriktirebildiği için de bir üstünlüğü vardır. Hızlı içerik birikiminin tam potansiyelini gerçekleştirmek için ise bir sistemin buna denk düşen bir büyük bellek kapasitesi olması gerekir. Aslana geldiğinizde abalonla ilgili her şeyi unuttuktan sonra, tüm bir kütüphaneyi yalayıp yutsanız ne fayda! Bir yapay zekâ sisteminin yeterli bellek kapasitesine sahip olması muhtemel olmakla birlikte, emülasyonlar insani şablonlarının kapasite sınırlılıklarından bazılarını miras alacaklardır. Bu nedenle sınırsız bir öğrenmeye ulaşabilmek için mimari geliştiricilere ihtiyaç duyacaklardır.

Buraya kadar mimarinin ve içeriğin uzlaşmazlığını, yani insanla denkliğe ulaşmış olan bir makine zekânın yazılımını ilerletmenin ne kadar zor olacağını ele aldık. Şimdi de makine zekanın performansını artırmanın üçüncü bir yoluna bakalım: Donanımını iyileştirmek. Donanım güdümlü ilerlemelerde uzlaşmazlık ne olacaktır?

Akıllı yazılımla (emülasyon ya da yapay zekâyla) başlamak gerekirse, programla ilgili daha fazla örnekleri çalıştırmak için ek bilgisayarlardan yararlanmaktan başka bir şey yapmadan *kolektif zekâyı* büyütebiliriz.[10] Keza programı daha hızlı bilgisayarlara kaydırarak *zekâ hızını* da artırabiliriz. Programın paralelleştirmeye ne kadar açık olduğuna bağlı olarak hız zekâsı programı daha fazla işlemcide çalıştırarak da artırılabilir. Son derece paralelleştirilmiş mimarisi olan emülasyonlar için bunun gerçekleştirilebilir olduğu söylenebilir; fakat birçok yapay zekâ programı da devasa paralelleştirmeden yarar sağlayabilecek önemli altprogramlara sahiptir. Bilgisayar gücünü artırarak nitelikli zekâyı artırmak da mümkün olabilir, ama bu daha dolaylı olacaktır.[11]

Dolayısıyla insan düzeyinde yazılıma sahip olan bir sistemde kolektif zekâyı ya da hızlı zekâyı (ve muhtemelen nitelikli zekâyı) artırmaya direncin düşük olması muh-

temeldir. Tek zorluk ek bilgisayar gücüne erişim elde etmektir. Bir sistemin donanım tabanını genişletmesi için her biri farklı bir zaman ölçeğinde ilintili birkaç yol vardır.

Kısa vadede, bilgi işlem gücü fonlamayla kabaca doğrusal olarak ölçeklenmelidir: İki kat fonlama iki kat daha fazla bilgisayar satın alır ve iki kat daha fazla yazılımın aynı anda çalıştırılmasına imkân sağlar. Bulut bilgi işlem hizmetlerinin ortaya çıkışı bir projeye bilgi işlem kaynaklarını yeni bilgisayarların kurulmasını beklemek zorunda kalmadan büyütme fırsatı sunuyor, ama gizlilik konusundaki kaygıların kurum içi bilgisayarlar kullanılmasını tercih edebileceğini de belirtelim. (Belli senaryolarda bilgi işlem gücü botnetler gibi başka yollarla da elde edilebilir.[12]) Verili bir etkenle sistemi ölçeklendirmenin ne kadar kolay olacağı ilk sistemin ne kadar bilgi işlem gücü kullandığına bağlıdır. İlk başta bir PC'de çalışan bir sistem salt birkaç milyon dolara bin faktörle ölçeklendirilebilir. Bir süper bilgisayarda çalışan bir programı ölçeklendirmek çok daha pahalı olacaktır.

Biraz uzun vadede, ek donanım edinmenin maliyeti dünyanın yüklü kapasitesinin artan bir kısmı dijital zihinleri çalıştırmak için kullanıldıkça daha da artabilir. Sözgelimi rekabetçi bir piyasaya dayalı emülasyon senaryosunda bir amülasyonun fazladan bir kopyasını çalıştırmanın maliyeti, marjinal kopyanın yarattığı gelire kabaca eşit olacaktır, zira yatırımcılar mevcut bilgi işlem altyapısının yatırımlarından bekledikleri geri dönüşe denk olması için fiyatı yukarı çekeceklerdir (ama bir proje o teknolojiyi hatmetmişse, bilgi işlem gücü piyasasında alıcı tekeli kurabilir ve dolayısıyla daha düşük ücret ödeyebilir).

Biraz daha uzun bir zaman ölçeğindeyse, bilgi işlem gücü arzı yeni kapasite yerleştikçe artar. Ani bir talep artışı mevcut yarı iletken dökümhanelerde üretimi tetikleyecek ve yeni fabrikalar kurulmasını teşvik edecektir. (Bir defaya mahsus bir performans artışı, belki bir iki kat büyüklükte bir artış, müşteriye göre düzenlenmiş mikroişlemciler kullanarak da elde edilebilir.[13]) Hepsinden öte, teknolojideki ilerlemelerin yükselen dalgası düşünen makinelerin türbinlerine çok daha fazla bilgi işlem gücü akıtacaktır. Tarihsel açıdan, bilgi işlem teknolojisinin gelişim hızı ünlü Moore yasasında tarif edilmiştir: Bu yasanın çeşitlemelerinden birine göre, dolar başına bilgi işlem gücü kabaca her 18 ayda bir ikiye katlanır.[14] Bu ilerleme hızının insan düzeyinde makine zekâsının gelişimine kadar sürmesine bel bağlamak mümkün olmasa da, temel fiziksel sınırlara ulaşana kadar bilgi işlem teknolojisinde ilerlemeler için alan kalacaktır.

Dolayısıyla donanım dikkafalılığının çok yüksek olmayacağını düşünmemek için bir neden yok. Sistem insan düzeyinde zekâya erişerek iyi olduğunu kanıtladıktan sonra onun için daha fazla bilgi işlem gücü satın almak bilgi işlem gücüne birkaç kat güç katabilir (projenin genişlemeden önce donanım bakımından ne kadar sade olduğuna bağlı olarak). Çipin kişiye özel hale getirilmesi de bir iki kat büyüklük sağlayabilir. Donanım tabanını genişletmenin diğer yolları (daha fazla fabrika açmak ve bilgisayar

teknolojisinin sınırlarını ileriye çekmek) daha uzun zaman, normalde birkaç yıl alır, ama bu gecikme makine süper zekâsı imalat ve teknoloji gelişiminde devrime yol açtığında ciddi derecede azalacaktır.

Özetlersek, bir *donanım fazlası* ihtimalinden bahsedebiliriz: İnsan düzeyinde yazılım yaratıldığında, muazzam sayıda kopyayı müthiş bir hızla işletmeye yetecek bilgi işlem gücü zaten hazır olacaktır. Yukarıda ele alındığı üzere, yazılım uzlaşmazlığı değerlendirilmesi daha zor bir şeydir, ama donanım uzlaşmazlığından bile daha az olabilir. Daha somutta, sistem için insanla denkliğe kavuştuğunda mevcut hale gelen önceden yaratılmış içerik biçiminde (örneğin internet) *içerik fazlalığı* olabilir. *Algoritma fazlalığı* da (önceden tasarlanmış algoritmik iyileştiriciler de) mümkündür, ama daha düşük ihtimallidir. Yazılımdaki ilerlemeler (ister algoritmalarda ister içerikte olsun) bir dijital zihin insanla denkliğe ulaştığında daha fazla ya da daha iyi donanım kullanarak elde edilebilecek performans kazanımlarının ötesine geçen gayet kolay şekilde erişilebilecek potansiyel performans kazanımları büyüklükleri sunabilir.

Optimizasyon gücü ve patlayıcılık

Uzlaşmazlık sorununu incelediğimize göre, şimdi şematik denklemimizin diğer yarısına dönmemiz gerekiyor: *Optimizasyon gücü*. Hatırlayalım: *Zekâdaki değişim hızı = Optimizasyon gücü/Uzlaşmazlık*. Bu şemada yansıtıldığı üzere, hızlı bir havalanma için geçiş aşaması sırasında uzlaşmazlığın düşük olması gerekmez. Eğer uzlaşmazlık daimiyse, hatta orta karar düzeyde bile artıyorsa, hızlı bir havalanma da gerçekleşebilir, yeter ki sistemin performansını ilerletmeye uygulanan optimizasyon gücü yeterince hızlı şekilde büyüsün. Birazdan göreceğimiz üzere, uygulanmış optimizasyon gücünün geçiş sırasında artacağını düşünmek için sağlam nedenler vardır, en azından bunun gerçekleşmesini engellemek için kasıtlı önlemlerin yokluğunda.

İki aşamayı birbirinden ayırt edebiliriz. İlk aşama havalanmayla, sistem bireysel zekâ için insani taban çizgisine ulaştığında başlıyor. Sistemin ehliyeti artmaya devam ettikçe, kendisini geliştirmek için bu ehliyetin bir kısmını ya da tamamını kullanabilir (ya da yerine bir sistem tasarlayabilir ki, buradaki amaçlarımız açısından aynı şeye denk gelir bu). Gelgelelim sisteme uygulanan optimizasyon gücünün çoğu hâlâ daha sistemin dışından, ya proje dahilinde çalıştırılan programcıların ve mühendislerin çalışmalarından ya da dünyanın geri kalanının imza attığı ve proje tarafından alınıp kullanılabilecek bu türde çalışmalardan gelir.[15] Eğer bu aşama önemli bir süre boyunca devam ederse, sisteme uygulanan optimizasyon gücünün miktarının büyümesini bekleyebiliriz. Seçilmiş yaklaşımın vaadi bariz hale geldiğinde hem projenin içinden hem de dış dünyadan girdilerin artması muhtemeldir. Araştırmacılar daha sıkı çalışabilirler, daha fazla araştırmacı devşirilebilir ve daha fazla bilgi işlem gücü satın alınarak ilerleme hızlandırılabilir. Artış eğer insan düzeyinde makine zekâsının gelişimi dünyayı gafil

avlarsa çok çarpıcı olabilir ve bu durumda eskiden küçük bir araştırma projesi olan şey birdenbire dünya çapında yoğun araştırma-geliştirme çabalarının odak noktası haline gelebilir (ama bu çabalardan bazıları rakip projelere kanalize edilebilir).

İkinci bir büyüme aşaması, eğer sistem bir noktada optimizasyon gücünün çoğu bizatihi sistemden gelecek kadar yeterlilik edinirse başlayacaktır (**ŞEKİL 7**'de "köprü" diye adlandırılan değişken düzeyiyle imlenmiştir). Bu durum dinamiği temelden değiştirir, çünkü sistemin yeterliliğindeki her artış artık onun daha da ilerletilmesine uygulanan optimizasyon gücü miktarındaki orantılı artışa dönüşür. Eğer uzlaşmazlık değişmeden kalırsa, bu geri dönüşüm dinamiği müthiş bir büyüme üretir (bkz. **KUTU 4**). İkiye katlanma sabiti senaryoya bağlıdır, ama aşırı kısa sürebilir (bazı senaryolarda sadece saniyelerle ölçülecek kadar), tabii eğer büyüme elektronik hızlarda gerçekleşiyorsa ki, algoritmik iyileştirmeler veya içerik ya da donanım fazlalığının sömürüsü sonucunda bu yaşanabilir.[16] Yeni bilgisayarların üretilmesi ya da ekipman imalatı gibi fiziksel inşanın güdülediği büyüme biraz daha uzun bir zaman ölçeği gerektirecektir (ama yine de bu zaman ölçeği dünya ekonomisinin mevcut büyüme hızıyla kıyaslandığında çok kısa olabilir).

Dolayısıyla uygulanan optimizasyon gücünün geçiş sırasında artması muhtemeldir: İlk başta insanlar muazzam bir ümit vaat eden bir makine zekâyı geliştirmek için çok uğraştıklarından, sonraysa bizatihi makine zekânın dijital hızlarda daha fazla ilerlemenin devindirici gücü olmayı başarmasından ötürü. Bu durum, *uzlaşmazlık değişmeden kalsa ya da insan taban çizgisi civarında hafifçe artıyor olsa bile*, hızlı ya da orta karar bir havalanma için gerçek bir imkân yaratacaktır.[17] Ama önceki altbaşlıkta gördüğümüz gibi, kapasite bakımından insani taban çizgisi düzeyi civarındaki uzlaşmazlıkta büyük bir düşüşe yol açabilecek etkenler vardır. Bu etkenler arasında örneğin bir işler yazılım zihnine erişildiğinde donanımın hızla genişlemesi imkânı; algoritmik ilerlemeler imkânı; ek beyinleri tarama imkânı (tüm beyin emülasyonunda) ve interneti özümseyerek muazzam miktarda içeriği hızla kapsama imkânı (yapay zekâ örneğinde) sayılabilir.[18]

Bu gözlemlere rağmen, ilintili alanda uzlaşmazlık eğrisinin şekli henüz iyi tanımlanmamıştır. Daha özelde, insan düzeyinde bir emülasyonun ya da yapay zekânın yazılım niteliğini iyileştirmenin ne kadar zor olacağı net değildir. Bir sistemin mevcut donanım gücünü büyütmenin zorluğu da açık değildir. Bugün küçük bir proje için mevcut bilgi işlem gücünü bu alana bin kat daha fazla para harcayarak ya da bilgisayar fiyatlarının düşmesi için birkaç yıl bekleyerek artırmak görece kolay olsa da, ilk makine zekânın insani taban çizgisine ulaşmasının pahalı süper bilgisayarları kapsayan büyük bir projeyle gerçekleşmesi (bunun ucuza ölçeklendirilmesi söz konusu olamaz) ve Moore yasasının o zamana kadar hükmünü yitirmesi mümkündür. Bu nedenlerden ötürü, hızlı ya da orta karar bir havalanma daha muhtemel görünse de, yavaş bir havalanma olanağı da dışlanamaz.[19]

KUTU 4: Bir Zekâ patlamasının kinetiği üzerine

Zekâdaki değişim hızını sisteme uygulanan optimizasyon gücü ile sistemin uzlaşmazlığı arasındaki oran olarak yazabiliriz:

$$\frac{dI}{dt} = \frac{\mathfrak{O}}{\mathfrak{R}}.$$

Bir sisteme etki eden optimizasyon gücü miktarı sistemin kendisinin sunduğu tüm optimizasyon gücünün ve dışarıdan uygulanan optimizasyon gücünün toplamıdır. Örneğin bir çekirdek yapay zekâ kendi çabaları ile bir insan programlama takımının çabalarının birleşimiyle ve belki bir de yarı iletken sanayisinde, bilgisayar biliminde ve ilintili alanlarda sürekli ilerlemeler kaydeden dünya çapındaki araştırmacıların katkılarıyla geliştirilebilir:*

$$\mathfrak{O} = \mathfrak{O}_{\text{sistem}} + \mathfrak{O}_{\text{proje}} + \mathfrak{O}_{\text{dünya}}$$

Bir çekirdek yapay zekâ çok sınırlı bilişsel kapasitelerle yola çıkar. Dolayısıyla başlangıçta $\mathfrak{O}_{\text{sistem}}$ küçüktür.† Peki ya $\mathfrak{O}_{\text{proje}}$ ve $\mathfrak{O}_{\text{dünya}}$? Tek bir projenin dünyanın geri kalanından daha fazla ilintili yeterliliğe sahip olduğu örnekler vardır: Sözgelimi Manhattan projesi dünyanın en iyi fizikçilerinin çok büyük bir kısmını atom bombası üzerine çalışmak üzere Los Alamos'a getirmişti. Daha yaygın olarak, her proje tek başına dünyanın ilintili toplam araştırma kapasitesinin sadece küçük bir kısmını içerir. Ama dış dünya bir projenin tek başına sahip olduğundan daha fazla toplam ilintili araştırma kapasitesine sahip olduğunda bile, $\mathfrak{O}_{\text{proje}}$ yine de $\mathfrak{O}_{\text{dünya}}$'yı aşabilir, zira dış dünyanın kapasitesinin çoğu söz konusu tikel sisteme odaklanmış değildir. Eğer bir proje ümit vaat edici görünmeye başlarsa (bir sistem, hiç olmadı, insan taban çizgisini geçtiğinde yaşanacaktır) ek yatırım çekerek $\mathfrak{O}_{\text{proje}}$'yi artırabilir. Eğer projenin başarıları kamuya mal olmuşsa, $\mathfrak{O}_{\text{dünya}}$ da artabilir, zira bu ilerleme genel olarak makine zekâya daha fazla ilgi uyandırır ve çeşitli güçler oyuna dahil olmak için uğraşmaya başlar. Dolayısıyla geçiş aşamasında bir bilişsel sistemi ilerletmek için uygulanan toplam optimizasyon gücünün sistemin kapasitesi arttıkça artması muhtemeldir.‡

---

* $\mathfrak{O}_{\text{dünya}}$ dünyanın optimizasyon gücünün söz konusu sistemin geliştirilmesine uygulanan kısmıdır. Tamamen yalıtık halde, dış dünyadan hiçbir ciddi destek almadan işleyen bir projede $\mathfrak{O}_{\text{dünya}}=0$ world sahibiz, ama proje tüm dünya ekonomisinden ve yüzyılların gelişiminden süzülüp gelen bir kaynak birikimiyle (bilgisayarlar, bilimsel kavramlar, eğitimli personel vb.) başlamış olmalıydı.

† Tohum yapay zekânın bilişsel yeteneklerinden burada en ilintili olanı, kendisini geliştirmek için zeki tasarım çalışması yapabilme yeteneğidir, yani zekâ artırma yeterliliğidir. (Eğer tohum yapay zekâ tohum yapay zekâyı geliştirmede iyi olan başka bir sistemi geliştirmede iyiyse, o zaman bunları daha büyük bir sistemin altsistemleri olarak görebilir ve analizimizde bu büyük resme odaklanabiliriz.)

‡ Bu demektir ki direngenlik yatırımı tamamen caydıracak ya da alternatif bir projeye yönlendirecek kadar yüksek değildir.

Sistemin kapasitesi arttıkça, sistemin kendisi tarafından yaratılan optimizasyon gücünün ona dışarıdan uygulanan optimizasyon gücüne hâkim olmaya başladığı bir nokta gelebilir (tüm kayda değer ilerleme alanlarında):

$$\mathfrak{O}_{\text{sistem}} > \mathfrak{O}_{\text{proje}} + \mathfrak{O}_{\text{dünya}}$$

Bu *köprü* önemlidir, çünkü bu noktanın ötesinde sistemin kapasitesinde daha fazla ilerleme sistemin ilerletilmesine uygulanan toplam optimizasyon gücünü artırmaya güçlü bir katkıda bulunur. Bu suretle güçlü bir yinelemeli kendini ilerletme rejimi devreye girer. Bu durum sistemin son derece farklı uzlaşmazlık eğrisi şekilleri dahilinde kapasitesinin patlamalı bir şekilde büyümesine yol açar.

Örnek vermek gerekirse, ilkin, uzlaşmazlığın değişmez olduğu, dolayısıyla da bir yapay zekânın zekâsındaki artış oranının uygulanan optimizasyon gücüne eşit olduğu bir senaryo düşünelim. Uygulanan tüm optimizasyon gücünün bizatihi yapay zekâdan geldiğini ve yapay zekânın tüm zekâsını kendi zekâsını artırmaya uyguladığını varsayın, böylece $\mathfrak{O}_{\text{sistem}} = I$ olur.* Bu durumda elimizde:

$$\frac{dI}{dt} = \frac{I}{k}.$$

Bu basit diferansiyel denklemin çözümü üstel fonksiyonu verir:

$$I = Ae^{t/k}$$

Ama uzlaşmazlığın değişmez olması çok özel bir durumdur. Uzlaşmazlık önceki alt kısımda bahsedilen bir ya da birkaç etkenden ötürü insan taban çizgisi civarında pekâlâ gerileyebilir ve köprünün etrafında ve biraz ötesinde düşük kalabilir (belki de sistem sonunda temel fiziksel sınırlara ulaşana kadar). Örneğin sistem kendi tasarımına ciddi bir katkı sunar hale gelmeden önce sisteme uygulanan optimizasyon gücünün kabaca değişmez olduğunu (örneğin $\mathfrak{O}_{\text{proje}} + \mathfrak{O}_{\text{dünya}} \approx c$) ve bunun sistemin her 18 ayda bir kapasitesini iki katına çıkarmasına yol açtığını varsayalım. (Bu da Moore yasasından tarihsel ilerleme hızlarıyla yazılım ilerlemeleri birleştiğinde varılan düzeyle kabaca aynı olacaktır.†) Bu ilerleme hızı, kabaca sabit optimizasyon gücü aracılığıyla sağlanırsa, sistem gücünün ters sonucu olarak uzlaşmazlığın gerileyişini gerektirir:

$$\frac{dI}{dt} = \frac{c}{1/I} = c\,I.$$

---

* Benzer bir örnek için, bkz. Yudkowsky (2008b).

† Girdiler yükseldiğinden (örneğin yeni dökümhaneler inşa etmeye yatırılan miktarların ve yarı iletken sanayisinde çalışan insanların sayısı), eğer girdilerdeki bu artışı kontrol edersek Moore yasası böyle hızlı bir büyüme sağlamaz. Gelgelelim yazılımdaki ilerlemelerle birleştirildiğinde, girdi birimi başına performanstaki 18 aylık iki katına çıkma zamanı tarihsel açıdan daha makul olabilir.

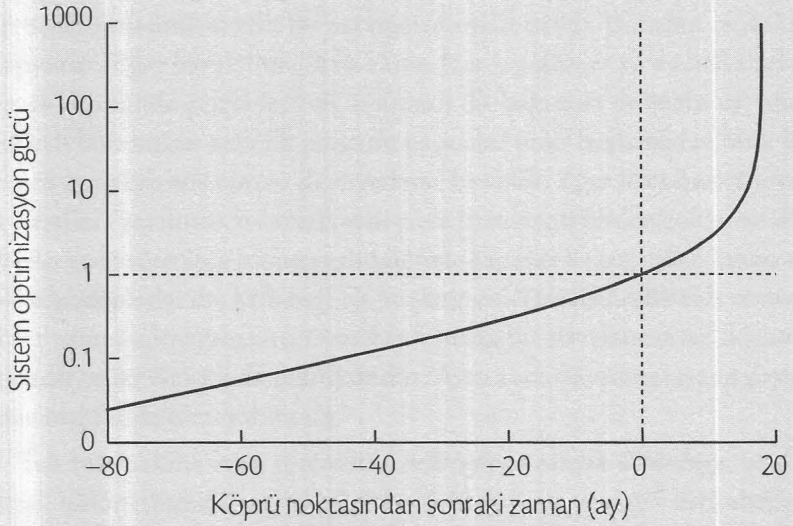
Eğer uzlaşmazlık bu abartılı örüntüde düşmeye devam ederse, o zaman yapay zekâ köprü noktasına ulaştığında yapay zekâyı geliştirmek için uygulanan optimizasyon gücünün toplam miktarı ikiye katlanmıştır. Bu durumda:

$$\frac{dI}{dt} = \frac{(c+I)}{1/I} = (c+I)I.$$

Bir sonraki ikiye katlanma 7,5 ay sonra gerçekleşir. 17,9 ay içinde sistemin kapasitesi bin kat artmış ve böylece hızlı süper zekâya erişmiştir (**ŞEKİL 9**).

Bu tikel büyüme güzergâhının pozitif bir tekilliği vardır: $t = 18$ ay. Gerçeklikte uzlaşmazlığın değişmez olduğu varsayımı, sistem er ya da geç bilgi işlemenin fiziksel sınırlarına yaklaşmaya başladığında artık geçerli olmayacaktır.

Bu iki senaryo sadece örnek amacıyla verilmiştir; uzlaşmazlık eğrisinin şekline bağlı olarak daha başka birçok güzergâh mümkündür. İddia sadece şudur: Köprü noktası civarında başlayan güçlü geri besleme döngüsü havalanmayı diğer seçeneklere kıyasla çok daha hızlı yapma eğilimindedir.



ŞEKİL 9 Zekâ patlamasının basit bir modeli.

Belirleyici Stratejik Üstünlük

Kinetik sorunundan ayrı ama onunla ilintili bir soru şudur: Süper zeki güç bir tane mi, birden çok mu olacaktır? Bir zekâ patlamasıyla projelerden biri geleceğe hükmedebilmesini sağlayacak kadar diğerlerinin önüne geçebilir mi? Yoksa ilerleme tekbiçim olup daha geniş bir cephede gerçekleşecek ve birçok proje katılım sergilerken hiçbiri ezici ve kalıcı bir üstünlük elde edemeyecek demek mi daha doğrudur?

Önceki bölümde başı çeken bir güç ile en yakın rakipleri arasında açılabilecek bir boşluğun büyüklüğünün, yani insandan ileri düzeyde süper zeki zekâya geçişin hızının belirlenmesindeki kilit bir parametre analiz edildi. Buradan muvakkat bir analize ulaşıyoruz. Eğer havalanma *hızlı* olursa (saatler, günler ya da haftalarla ölçülecek bir zaman diliminde gerçekleşirse), o zaman iki bağımsız projenin ayı anda yola çıkma ihtimali bulunmaz; keza ilk projenin başka bir proje başlamadan önce kendi havalanmasını tamamlamış olması da neredeyse kesindir. Eğer havalanma *yavaşsa* (yıllara ya da onyıllara yayılırsa), o zaman aynı anda havalanan birden çok proje olabilir, böylece projeler geçiş sürecinin sonuna gelindiğinde kapasite bakımından muazzam kazanımlar elde edecek olsalar da, herhangi bir projenin ezici bir üstünlük elde etmesini sağlayacak kadar zaman olmayacaktır. *Orta karar* hızda bir havalanma ise ikisinin arasında bir yerdedir ve iki olasılık da mümkündür: Aynı anda havalanmaya başlayan birden fazla proje olabilir de olmayabilir de.[1]

Tek bir makine zekâ projesinin *belirleyici stratejik üstünlüğe*, yani dünyaya tam olarak hükmetmesini sağlayacak teknolojik düzeye ve diğer üstünlüklere erişecek şekilde rekabette öne fırlayacağını söyleyebilir miyiz? Eğer bir proje gerçekten belirleyici stratejik üstünlüğe erişirse, bunu rakiplerini bastırmak ve *tekel* (küresel düzeyde tek bir karar alma merciinin olduğu bir dünya düzeni) oluşturmak için kullanacak mıdır? Ve eğer kazanan bir proje olursa, bu ne kadar "büyük" olacaktır: Fiziksel büyüklük ya da bütçe bakımından değil, tasarımının kaç insanın arzuları tarafından kontrol edildiği bakımından? Bu soruları sırayla ele alacağız.

Önde giden proje belirleyici bir stratejik üstünlük elde edecek midir?

Önde giden ile takipçileri arasındaki boşluğun büyüklüğünü etkileyen etkenlerden biri, başı çekene rekabette üstünlük kazandıran her neyse onun dağılma hızıdır. Önde giden proje, şayet takipçileri onun fikirlerini ve yeniliklerini kolayca kopyalayabiliyorsa, büyük bir üstünlük elde etmeyi ve korumayı başaramayacaktır. Taklit, lidere üstünlük kazandırmayan ve geriden gelenlere yarar sağlayan bir pruva rüzgârı yaratır, özellikle de entelektüel mülkiyetin korunma düzeyi zayıfsa. Önde giden proje; temellük, vergilendirme ya da tekel karşıtı mevzuat yüzünden parçalanma riskiyle de karşı karşıya kalabilir.

Gelgelelim bu pruva rüzgârının önde giden ile takipçileri arasında yeknesak şekilde artması gerektiğini varsaymak yanlış olur. Nasıl bir bisiklet yarışçısı rekabette çok geriye düştüğünde önündeki pedalcılar sayesinde rüzgârdan korunması artık mümkün olmazsa, bir teknoloji takipçisi de en son gelişmelerin çok gerisinde kaldığında en ön safta kaydedilen ilerlemeleri özümsemekte zorlanabilir.[2] Anlama ve kapasite konusundaki boşluk çok büyümüş olabilir. Lider daha ileri bir teknoloji platformuna geçmiş olabilir ve sonraki yenilikleri geride kalanların kullandığı ilkel platformlara aktarmak mümkün olmayabilir. Yeterince rakipsiz bir lider, araştırma programlarından ve onun hassas kurulumlarından bilgi sızdırılmasını engelleyecek ya da rakiplerinin kendi ileri kapasitelerini geliştirme çabalarını baltalayacak yeteneğe sahip olabilir.

Eğer önde giden bir yapay zekâ sistemiyse, yayınım hızını azaltırken yeterliliklerini geliştirmesini kolaylaştıran özelliklere sahip olabilir. İnsanların yönettiği örgütlerde ölçek ekonomilerine bürokratik verimsizlikler ve aracılık problemleri tebelleş olur (ticari sırları saklama zorlukları dahil).[3] Bu problemler, insanlar tarafından işletilen bir makine zekâ projesinin büyümesini muhtemelen sınırlayacaktır. Gelgelelim bir yapay zekâ sistemi bu olumsuz ölçek ekonomilerinin bazılarından sakınabilir, zira yapay zekâ modülleri (insan işçilerin aksine) bir bütün olarak sisteminkinden ayrılan tercihlere sahip olmak zorunda değildir. Dolayısıyla yapay zekâ sistemi insan işletmelerindeki aracılık problemlerinden doğan verimsizliklerin büyükçe bir kısmından kaçınabilir. Aynı üstünlük (tamamen sadık parçalara sahip olmak) bir yapay zekâ sisteminin uzun erimli gizli hedeflere girişmesini de daha kolay hale getirecektir. Bir yapay zekânın rakipleri tarafından rahatsız edilmeye ya da rüşvetle köstebeklik yapmaya ikna edilebilecek rahatsız çalışanları olmayacaktır.[4]

Bazı tarihsel örneklere bakarak gelişim zamanlarındaki akla yatkın boşlukların dağılımına dair bir fikir edinebiliriz (bkz. **KUTU 5**). Görünen o ki stratejik açıdan kayda değer teknoloji projeleri için birkaç ay ile birkaç yıl arasında değişen gecikmeler tipiktir.

Küreselleşmenin ve artan gözetlemenin rakip teknoloji projeleri arasındaki tipik gecikmeleri azaltması mümkündür. Fakat ortalama gecikmenin ne kadar kısa olabilece-

ğine dair bir alt sınır olması muhtemeldir (kasıtlı bir eşgüdüm olmadığında).[5] Kartopu etkisine yol açan dinamikler olmadığında bile, bazı projelerin daha iyi araştırma ekibi, önderliği ve altyapısı olacaktır ya da daha iyi fikirlere denk gelecektir.

Bu gözlemleri havalanma hızına dair daha önceki değerlendirmemizle birleştirdiğimizde, iki projenin aynı anda hızlı bir havalanmayı gerçekleştirecek kadar yakın olma ihtimallerinin çok düşük olduğu sonucuna varabiliriz; orta karar bir havalanma için iki yoldan da gidebilir; yavaş bir havalanma için ise birkaç projenin süreci paralel olarak yaşaması son derece muhtemeldir. Ama analizimizi bir adım daha ilerletmemiz lazım. Kilit soru, kaç projenin birlikte havalanma gerçekleştirdiği değil, kaç projenin öte tarafta yeterlilik bakımından hiçbirinin belirleyici bir üstünlüğe sahip olmayacağı kadar kümelenmiş şekilde ortaya çıkacağıdır. Eğer havalanma süreci görece yavaş başlar ve sonra hızlanırsa, rakip projeler arasındaki uzaklık genellikle artar. Bisiklet metaforumuza dönecek olursak, durum dik bir yokuşu çıkan iki bisikletçinin durumuna benzer: Biri diğerinin arkasındadır ve aralarındaki uzaklık, önde gidenin tepeye ulaşması ve diğer taraftan aşağı doğru ivme kazanmasıyla açılmaya başlar.

Şu orta karar havalanma senaryosunu düşünün. Bir projenin yapay zekâ yeterliliğini insani taban çizgisinden güçlü bir süper zekâ düzeyine çıkarmasının bir yıl aldığını ve bir projenin bu havalanma aşamasına bir sonraki en ileri projeden altı aylık üstünlükle girdiğini varsayalım. İki proje aynı anda havalanma gerçekleştirecektir. Bu durumda, iki projenin de belirleyici bir stratejik üstünlük kazanmayacağı düşünülebilir. Ama böyle olması gerekmiyor. İnsani taban çizgisinden köprü noktasına ilerlemenin dokuz ay, oradan da güçlü süper zekâya ilerlemenin üç ay aldığını varsayalım. Bu durumda önde giden proje, diğer proje daha köprü noktasına bile ulaşmadan önce güçlü süper zekâya ulaşır. Bu sayede önde giden proje belirleyici bir stratejik üstünlüğe ve rakip projeleri safdışı bırakıp tekelini kurarak önderliğini kalıcı bir kontrole çevirme fırsatına erişir. (Tekel kavramının soyut olduğunu belirtelim: Tekel demokrasi de olabilir, despotluk da, tek bir hâkim yapay zekâ da, uygulamaya geçirilmesi için etkili önlemler içeren birtakım güçlü küresel normlar da, hatta yabancı bir efendi de – belirleyici özelliği tüm büyük küresel eşgüdüm problemlerini çözmek olan bir aracılık biçimi olmasıdır. İnsan yönetişiminin tanıdık biçimlerine benzeyebilir, ama benzeyecek diye bir kaide yoktur.)[6]

Köprü noktasından hemen sonra bilhassa güçlü bir patlayıcı büyüme olasılığı olduğundan, optimizasyonun güçlü olumlu geri besleme döngüsü devreye girdiğinde, bu tür bir senaryo ciddi bir olasılıktır ve havalanma hızlı olmasa bile önde giden projenin belirleyici bir stratejik üstünlüğe erişme şansı artacaktır.

KUTU 5: Teknoloji yarışları: Bazı tarihsel örnekler

Tarihsel açıdan uzun zamansal ölçeklerde bilgi ve teknolojinin dünyanın dört bir tarafına yayılma hızında bir artış yaşanmıştır. Bunun sonucunda teknoloji liderleri ile en yakın takipçileri arasındaki zamansal boşluklar daralmıştır.

Çin iki bin yıl boyunca ipek üretiminde tekele sahip olmayı başarmıştı. Arkeolojik bulgular üretimin MÖ 3000 civarında, hatta daha da önce başlamış olabileceğini söylüyor.* İpekböcekçiliği çok gizli tutulan bir sırdı. Teknikleri açık etmek, tıpkı ipekböceği ya da yumurtalarını Çin dışına ihraç etmek gibi, ölüm cezasına çarptırılmak anlamına geliyordu. Romalılar, imparatorluklarında ithal ipekli kıyafetin fiyatları yüksek olmasına rağmen, ipek imalatını bir altın bilezik olarak kollarına takmayı hiçbir zaman başaramadılar. Ancak MS 300 civarında bir Japon keşif heyeti dört küçük Çinli kızla birlikte bazı ipekböceği yumurtalarını ele geçirmeyi başarınca bu Çinli kızlardan ipek imalatı sanatını zorla öğrenmişlerdi.† Bizans ise üreticiler kervanına MS 522 civarında katıldı. Porselen yapımının hikâyesi de uzun gecikmelerle doludur. Bu zanaat Çin'de MS 600 civarında Tang hanedanı zamanında icra ediliyordu (ve daha MS 200 civarında da yapılıyor olması muhtemeldir), ama Avrupalıların bu zanaatı hatmetmesi ancak 18. yüzyılda mümkün oldu.‡ Tekerlekli araçlar Avrupa'da ve Mezopotamya'da birkaç sitede MÖ 3500 civarında ortaya çıktı, ama Amerika kıtasına ancak Kolombus sonrası dönemde ulaştı.§ Daha büyük ölçekte, insan türünün dünyanın dört bir tarafına yayılması binlerce yıl sürdü, Tarım Devrimi'nin binlerce yıl, Sanayi Devrimi'nin ise sadece yüzlerce yıl. Enformasyon Devrimi'nin tüm dünyaya yayılmasının sadece onyıllar aldığı söylenebilir, ama elbette bu geçişlerin her zaman eşit derinlikte olmadığını unutmayalım. (*Dance Dance Revolution* adlı video oyunu Japonya'dan Avrupa ve Kuzey Amerika'ya sadece bir yılda yayıldı!)

Teknolojik rekabet özellikle de patent ve silah yarışları bağlamında çok ciddi düzeyde incelenmiştir.¶ Bu yazını ele almak buradaki amaçlarımız arasında yer almıyor. Ne var ki 20. yüzyılda stratejik açıdan kayda değer teknoloji yarışlarından bazı örneklere bakmak bilgilendirici olacaktır. (bkz. **TABLO 7**).

---

* Barber (1991) Yangshao kültürünün (MÖ 5000-3000) ipekliden yararlanmış olabileceğini söyler. Sun vd. (2012) genetik çalışmalarından hareketle, ipekböceğinin 4.100 yıl önce evcilleştirildiğini söylerler.

† Cook (1984, 144). Bu hikâye tarihsel incelemeye dayanamayacak kadar iyi olabilir, daha ziyade ipekböceklerinin gezgin keşişler tarafından güya bambu çubuklarının içinde getirildiğine dair Prokopius'un (*Wars* VIII.xvii.1–7) hikâyesine benzer (Hunt 2011).

‡ Wood (2007); Temple (1986).

§ Kolombus öncesi kültürlerde tekerlek vardı, ama sadece oyuncaklar için kullanıyorlardı (muhtemelen iyi koşum hayvanları olmadığından ötürü).

¶ Koubi (1999); Lerner (1997); Koubi ve Lalman (2007); Zeira (2011); Judd vd. (2012).

TABLO 7 Stratejik açıdan önemli bazı teknoloji yarışları

	ABD	SSCB	Britanya	Fransa	Çin	Hindistan	İsrail	Pakistan	Kuzey Kore	Güney Afrika
Atom bombası	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
Hidrojen bombası	1952	1953[1]	1957	1968	1967	1998?	—	—	—	
Uzaya uydu fırlatabilme	1958	1957	1971	1965	1970	1980	1988	—	1998?[2]	—[3]
Uzaya insan gönderebilme	1961	1961	—	—	2003	—	—	—	—	—
ICBM[4]	1959	1960	1968[5]	1985	1971	2012	2008	—[6]	2006	—[7]
MIRV[8]	1970	1975	1979	1985	2007	2014[9]	2008			

1 Testler 1989'da yapılmış, proje 1994'te iptal edilmiştir.
2 5 bin km'den daha uzun bir menzili vardır.
3 ABD'den alınan Polaris füzeleri.
4 RDS-6 1953'te ilk füsyon reaksiyonlu bomba testiydi, ama RDS-37 1955'te ilk "gerçek" füsyon bombasıydı (gücün çoğu füsyon reaksiyonundan geliyordu).
5 Taimur füzesi çalışmaları sürmektedir ve muhtemelen Çin füzeleri örnek alınmıştır.
6 1989–90'da test edilen RSA-3 roketi uydu fırlatma amacı taşıyordu ve/veya ICBM olarak düşünülmüştü.
7 Teyit edilmemiştir.
8 MIRV = "bağımsız olarak hedeflenebilir çoklu araç", tek bir balistik füzenin farklı hedefleri vuracak şekilde programlanabilen çoklu savaş başlıkları taşımasını sağlayan bir teknoloji.
9 Agni V sistemi henüz kullanımda değildir.

Askeri ya da sembolik öneminden ötürü rakip süper güçler tarafından stratejik öneme sahip olarak görülen bu altı teknolojiye ilişkin olarak, lider ile en yakın takipçisi arasındaki uçurum sırasıyla (takriben) 49 ay, 36 ay, 4 ay, 1 ay, 4 ay ve 60 aydır: Hızlı bir havalanmanın süresinden daha uzun, yavaş bir havalanmanın süresinden daha kısa.* Birçok örnekte, geriden gelenin projesi köstebeklik ve halka açık bilginin nimetlerinden faydalanmıştır. Bir buluşun uygulanabilirliğinin gösterilmesi bile başkalarını bunu bağımsız olarak geliştirmeye teşvik edebilir ve geride kalma korkusu yetişme çabalarını teşvik edebilir.

Belki de yapay zekâ örneğine daha yakın olanı, yeni fiziksel altyapı geliştirilmesini gerektirmeyen matematiksel buluşlardır. Bunlar çoğu zaman akademik yayınlarda yayımlanır ve dolayısıyla herkese açık diye görülebilir, ama bazı örneklerde keşif stratejik bir üstünlük sunuyor gözüktüğünde, yayın gecikir. Örneğin açık anahtarlı şifrelemedeki en önemli fikirler Diffie-Hellman anahtar değişimi protokolü ve RSA şifreleme şemasıdır. Akademik camia bunlardan ilkini 1976'da, ikincisini 1978'de keşfetmiştir, ama Birleşik Krallık'ın iletişim güvenlik grubu

* Çeşitli kaynaklardan hesaplanmıştır. Zaman aralığı çoğu zaman biraz keyfidir ve "eşdeğer" yeterliliklerin tam olarak nasıl tanımlandığına bağlıdır. Radar hemen ilk birkaç yıl içinde en az iki ülkede kullanılmıştır, ama ay olarak kesin rakam vermek zordur.

tarafından 1970'lerin başından beri zaten bilindiği sonradan kanıtlanmıştır.* Büyük yazılım projeleri yapay zekâ projeleriyle daha yakın bir analoji oluşturabilir, ama yazılım genellikle aşama aşama halka sunulduğu ve rakip sistemlerin işlevsellikleri çoğu zaman doğrudan kıyaslanabilir olmadığı için canlı örnekler vermek zordur. Eğer iki proje alternatif yaklaşımlara sahip olursa ve bunlardan birinin daha iyi işlediği anlaşılırsa, rakip proje, öncünün yaptıklarını yakından izliyor olsa bile, daha üstün olan yaklaşıma geçmesi aylar alabilir.

* Ellis (1999).

Başarılı proje ne kadar büyük olacaktır?

Süper zekâya giden bazı yollar büyük kaynaklar gerektirir ve dolayısıyla büyük fon ayrılan projelerin alanında olması muhtemeldir. Biyolojik zekâ geliştiriciler ve beyin-bilgisayar arayüzleri de büyük çaplı bir etken olacaktır: Küçük bir biyoteknoloji şirketi bir iki ilaç icat edebilirse de, bu yollardan birinde süper zekâya erişmek (eğer erişilebilirse) muhtemelen birçok icat ve bir o kadar test gerektirecek, dolayısıyla bir sanayi sektörünün ya da sağlam fon ayrılmış bir ulusal programın desteğine ihtiyaç duyacaktır. Örgütlenmeleri ve ağları daha verimli hale getirerek kolektif süper zekâya erişmek ise dünya ekonomisinin büyük bir kısmı da dahil olmak üzere daha yaygın bir girdi gerektirir.

Yapay zekâ yolunu değerlendirmek ise daha zordur. Belki çok büyük bir araştırma programı gerektirecek; belki küçük bir grupla yapılabilecektir. Yalnız bir *hacker* senaryosu da dışlanamaz. Bir çekirdek yapay zekâ inşa etmek uzun yıllar boyunca tüm dünyadaki bilim camiasının geliştirdiği içgörülere ve algoritmalara ihtiyaç duyabilir. Ama son hayati atılımı sağlayacak fikrin her şeyi bir araya toplayan tek bir bireyden ya da küçük bir gruptan gelmesi mümkündür. Bu senaryo bazı yapay zekâ mimarileri için diğerleri için olduğundan daha az gerçekçidir. Birlikte etkili şekilde işlemesi için evrilip çevrilmesi gereken çok sayıda parçası bulunan ve ardından hummalı bir şekilde ısmarlama bilişsel içerikle doldurulan bir sistemin daha büyük bir proje gerektirmesi muhtemeldir. Ama eğer bir çekirdek yapay zekâ basit bir sistem, inşası sadece birkaç temel ilkeyi yerli yerine oturtmaya dayalı bir sistem olarak somutlaştırılabilirse, o zaman başarı küçük bir ekibin ya da bir bireyin avuçları arasında olabilir. Eğer alandaki önceki ilerlemenin çoğu kamuya açık eserlerde yayınlandıysa ya da açık kaynak yazılımı olarak mevcutsa, nihai atılımın küçük bir proje tarafından gerçekleştirilmesi ihtimali artar.

Sistemin doğrudan *mühendisliğini yapan* projenin ne kadar büyük olacağı sorunu ile sistemin yaratılıp yaratılmadığını, ne zaman ve nasıl yaratıldığını *kontrol eden* grubun ne kadar büyük olacağı sorunundan ayırmak zorundayız. Atom bombası esasen bir grup bilimci ve mühendis tarafından yaratıldı. (Manhattan Projesi tepe noktasındayken 130

bin kişiyi istihdam ediyordu ve bunların ezici çoğunluğunu inşaat işçileri ya da inşaat operatörleri oluşturuyordu.[7]) Gelgelelim bu teknik uzmanlar ABD ordusuna bağlıydı, ABD ordusu ABD hükümetinin kontrolündeydi, ABD hükümeti de o sırada dünya yetişkin nüfusunun yaklaşık onda birini oluşturan Amerikalı seçmenlere nihayetinde hesap vermek zorundaydı.[8]

Gözetleme

Süper zekânın güvenlik açısından aşırı içerimleri dikkate alındığında, hükümetlerin kendi topraklarında denenen ve havalanma gerçekleştirmeye yakın olduğunu düşündükleri bir projeyi devletleştirmeye çalışmaları muhtemeldir. Güçlü bir devlet başka ülkelerde girişilen projeleri de casusluk, hırsızlık, kaçırma, rüşvet, tehdit, fetih ya da başka yollarla da ele geçirmeye çalışabilir. Güçlü bir devlet yabancı bir projeyi ele geçiremediğindeyse o projeyi yok etmek isteyebilir, eğer ev sahibi ülke etkili bir caydırıcıdan mahrumsa. Atılım ufukta gözükmeye başladığında küresel yönetişim yapıları güçlü olduğu takdirde, ümit vaat eden projelerin uluslararası kontrole tabi hale getirilmesi mümkündür.

Dolayısıyla önemli sorulardan biri ulusal ya da uluslararası otoritelerin bir zekâ patlamasının ufukta göründüğünü düşünüp düşünmeyecekleridir. Şu anda istihbarat örgütleri umut vaat eden yapay zekâ projelerinin ya da potansiyel açıdan patlayıcı olan diğer zekâ ilerletme biçimlerinin çok tehlikeli olduğunu düşünmüyorlar.[9] Eğer gerçekten de çok fazla ilgi göstermiyorlarsa, bunun nedeni bir süper zekâ ihtimalinin ufukta gözükmediğine dair yaygın bir algının varlığı olabilir. Eğer saygın bilimciler arasında süper zekânın ufukta gözüktüğüne dair yaygın bir kanı ortaya çıkarsa ve çıktığında, dünyadaki başlıca istihbarat örgütleri muhtemelen ilgili araştırma alanındaki grupları ve bireyleri gözetlemeye başlayacaklardır. O zaman yeterli ilerlemeyi göstermeye başlayan her proje ânında devletleştirilebilir. Eğer siyasi elitler riskin ciddiliğine ikna olurlarsa, sivillerin hassas alanlardaki çabaları düzenlenebilir ya da yasadışı ilan edilebilir.

Bu tür bir izleme ne kadar zor olacaktır? Eğer tek hedef önde giden projeyi takip etmekse işimiz kolaydır. Bu durumda, en iyi kaynaklarla donatılmış birkaç projeye odaklanan gözetleme yeterli olacaktır. Eğer amaç herhangi bir çalışmanın (en azından özel olarak yetkili kurumların dışında) gerçekleşmesini engellemekse, o zaman gözetlemenin kapsamlı olması gerekecektir, zira birçok küçük proje ve birey en azından belli bir ilerleme kaydedecek konumda olacaktır.

Ciddi miktarda fiziksel sermaye gerektiren projeleri (mesela bir tüm beyin emülasyonu projesi) izlemek daha kolay olacaktır. Halbuki yapay zekâ araştırmaları sadece bir kişisel bilgisayar gerektirir ve dolayısıyla izlemesi daha zor olacaktır. Teorik işlerin bir kısmı kâğıt-kalemle yapılabilir. Böyle olsa bile, yapay genel zekâ araştırmasına ciddi ve kalıcı bir ilgisi olan en ehil bireyleri saptamak çok zor olmayacaktır. Bu tür

bireyler genellikle arkalarında görünür izler bırakırlar. Akademik makaleleri yayımlanmış, konferanslarda sunum yapmış, internet forumlarına yazmış ya da önde gelen bilgisayar mühendisliği bölümlerinden diploma almış olurlar. Ayrıca diğer yapay zekâ araştırmacılarıyla iletişim kurmuş da olabilirler, bu sayede sosyal grafiğin haritalandırılmasıyla saptanabilirler.

Daha baştan gizli olması tasarlanan projeleri tespit etmek daha zor olabilir. Sıradan bir yazılım geliştirme projesi bir cephe işlevi görebilir.[10] Projenin ne yapmaya çalıştığının gerçek doğası ancak dikkatli bir kod analiziyle ortaya çıkacaktır. Bu tür bir analiz son derece yetenekli bir sürü insan gücü gerektirir ki, bunlardan sadece az sayıda şüpheli proje bu düzeyde mercek altına alınabilir. Eğer etkili yalan makinesi teknolojisi geliştirilmiş ve bu tür gözetlemelerde rutin bir şekilde kullanılıyor olsaydı, işimiz çok daha kolay olurdu.[11]

Devletlerin öncü gelişmeleri yakalamayı başaramayabilecek olmalarının bir diğer nedeni de bazı atılım türlerini öngörmenin içkin zorluğudur. Bu durum tüm beyin emülasyonu geliştirme çalışmalarından ziyade yapay zekâ araştırmalarıyla daha ilintilidir, zira ilkinde kilit atılımın öncesinde mütemadi ilerlemelerin net bir gradyanının gelmesi daha muhtemeldir.

Ayrıca istihbarat örgütlerinin ve diğer devlet bürokrasilerinin belli bir hantallık ya da katılık göstererek dışarıdaki bazı grupların açıkça gördüğü bazı gelişmelerin anlamını kavrayamamaları da mümkündür. Potansiyel bir zekâ patlamasının resmi olarak anlaşılmasının önündeki engeller özellikle güçlü olabilir. Örneğin konu dini ya da siyasi ihtilaflara konu olabilir ve bazı ülkelerde yetkililerin gözünde tabu haline gelebilir. Konu itibarsız bir simayla ya da genel olarak şarlatanlıkla ilişkilendirilebilir ve dolayısıyla saygın bilimciler ve diğer bilindik simalar meseleden uzak durabilir. (Birinci Bölümde gördüğümüz gibi, buna benzer bir şey iki kez yaşanmıştır: İki "yapay zekâ kışı"nı hatırlayın.) Sanayi grupları kârlı işletme alanlarına leke sürülmesini engellemek için kulis çalışması yapabilirler; akademik topluluklar yapılmakta olan bilimin uzun erimli sonuçlarına ilişkin kaygılarını dile getirenleri marjinalleştirmek üzere saflarını sıklaştırabilirler.[12]

Sonuç olarak, tam bir istihbarat fiyaskosu yaşanma ihtimali reddedilemez. Bu tür bir fiyasko, eğer atılımlar kamuoyunda ciddi yer işgal etmeden, yakın gelecekte gerçekleşirse bilhassa muhtemeldir. Ayrıca istihbarat örgütleri işlerini layıkıyla yapsalar bile, siyasi önderler tavsiyelere kulak tıkayabilirler. Manhattan Projesi'ni başlatmak birkaç öngörülü fizikçinin olağanüstü çabalarını gerektirmişti ve bunların içinde Mark Oliphant ve Leó Szilárd bilhassa öne çıkar: Szilárd, Eugene Wigner'i ikna ederek, Başkan Franklin D. Roosevelt'i konuyla ilgilenmeye ikna etmek üzere yazılan mektuba Albert Einstein'ı imza atmaya ikna ettirmişti. Roosevelt proje uç sınırına ulaştıktan

sonra bile uygulanabilirliğinden ve öneminden şüphe duymaya devam etti, tıpkı halefi Harry Truman gibi.

Öyle ya da böyle, eğer devletler gibi büyük oyuncular aktif rol alırlarsa, küçük eylemci gruplarının bir zekâ patlamasının serencamını etkilemeleri muhtemelen daha zor olacaktır. Dolayısıyla potansiyel bir zekâ patlamasından doğacak varoluşsal risklerin toplam miktarını azaltma konusunda bireylerin önündeki fırsatlar büyük oyuncuların meseleye görece ilgisiz yaklaştıkları ya da eylemlerin ilk çabalarının büyük oyuncuların oyuna girip girmeyeceğine, ne zaman gireceğine, hangi tutumla gireceğine ya da hangi oyuncuların gireceğine büyük bir etkide bulunduğu senaryolarda en güçlüdür. Dolayısıyla beklenen azami etkiyi yaratmayı amaçlayan eylemciler planlarının çoğunu bu tür büyük senaryolara odaklamak isteyebilirler, velev ki büyük oyuncuların bütün önemli şutları kullandığı senaryoların daha muhtemel olduğuna inanıyor olsunlar.

Uluslararası işbirliği

Eğer küresel yönetişim yapıları genel olarak güçlenirse uluslararası eşgüdüm daha muhtemeldir. Keza bir zekâ patlamasının mahiyeti çok öncesinde geniş kesimlerce takdir edilirse ve eğer tüm ciddi projelerin etkin şekilde izlenmesi mümkün olursa da, eşgüdüm daha muhtemel olabilir. Gelgelelim izleme gerçekleştirilebilir olmasa bile, uluslararası işbirliği yine de mümkün olacaktır. Birçok ülke bir araya gelip ortak bir projeyi destekleyebilir. Eğer böyle bir projeye yeterince iyi kaynak ayrılırsa, hedefe ilk ulaşan olma şansı gayet yüksek olabilir, özellikle de rakip bir proje tespit edilemeyecek kadar küçük ve gizli olacaksa.

Başarı elde etmiş büyük çaplı çokuluslu bilimsel işbirliklerinin öncelleri yok değil: Uluslararası Uzay İstasyonu, İnsan Genomu Projesi ve Büyük Hadron Çarpıştırıcısı.[13] Gelgelelim bu örneklerde işbirliğinin arkasındaki esas saik masrafların paylaşılmasıydı. (Uluslararası Uzay İstasyonu örneğinde, Sovyet Rusya ile ABD arasında işbirliği ruhunun güçlendirilmesi bizatihi önemli bir hedefti.[14]) Güvenlik açısından muazzam içerimleri olan bir projede benzer bir işbirliğinin gerçekleştirilmesi daha zor olacaktır. Tek taraflı şekilde atılım gerçekleştirebileceğine inanan bir ülke, çabalarını ortak bir projeye tabi kılmaktansa tek başına yola devam etme kararı almayı çekici bulabilir. Keza bir ülke diğer katılımcıların ortak yaratılan içgörüleri kendine mal edip gizli bir ulusal projeyi hızlandırmak üzere kullanmasından duyduğu korkuyla da uluslararası işbirliğinden uzak durabilir.

Dolayısıyla bir uluslararası projenin çok büyük güvenlik zorluklarını aşması ve başlangıç adımlarının atılması için ciddi bir güven tesis edilmesi gerekecektir ki, bu güvenin yerleşmesi zaman alabilir. Sözgelimi Gorbaçov'un başa geçmesiyle ABD ile SSCB arasındaki ilişkiler yumuşama dönemine girdikten sonra bile silahlanmanın azaltılması çabaları –iki süper gücün de fazlasıyla çıkarına olacak olmasına karşın–

kötü bir başlangıç yapmıştı. Gorbaçov nükleer silahların ciddi düzeyde azaltılmasını istiyordu, ama Reagan'ın Stratejik Savunma İnisiyatifi ("Yıldız Savaşları") meselesinden ötürü müzakereler tıkandı, zira Kremlin buna şiddetle karşı çıkıyordu. Reagan 1986'daki Reykjavik Zirvesi toplantısında ABD ile SSCB'nin Stratejik Savunma İnisiyatifi kapsamında geliştirilecek olan teknolojiyi paylaşmasını teklif etti, böylece iki ülke de kazara gerçekleştirilecek füze fırlatmalara ve nükleer silah geliştirebilecek küçük ülkelere karşı korunma elde edebilecekti. Ama Gorbaçov bu görünüşteki kazan-kazan teklifine sıcak bakmadı. Bunun bir dalavere olduğunu düşündü ve ABD'nin Sovyetlerle inek sağma teknolojisini bile paylaşmaya yanaşmadığı bir dönemde en ileri askeri araştırmalarının meyvelerini paylaşacağı görüşüne prim vermedi.[15] Reagan'ın süper güçlerin işbirliği konusundaki teklifinde samimi olup olmadığından bağımsız olarak, güvensizlik yüzünden başlangıç adımı atılamadı.

İşbirliğinin müttefikler arasında tesis edilmesi daha kolaydır, ama bu bile otomatik olarak gerçekleşmeyecektir. SSCB ve ABD İkinci Dünya Savaşı sırasında Almanya'ya karşı ittifak yaptıklarında, ABD atom bombası projesini SSCB'den saklamıştı. ABD Manhattan Projesi'nde Britanya ve Kanada ile işbirliği yapmayı tercih etmişti.[16] Benzer şekilde, Birleşik Krallık da Alman Enigma kodunu kırmayı başardığını SSCB'den gizlemiş, ama biraz zor olsa da ABD'yle paylaşmıştı.[17] Demek ki ulusal güvenlik açısından hayati önem taşıyan bir teknolojide uluslararası işbirliği tesis etmek için, güven esasına dayalı yakın bir ilişkiyi öncesinde kurmak zorunlu olabilir.

Zekâ ilerletme teknolojilerinin gelişiminde uluslararası işbirliğinin önemi ve uygulanabilirliği meselesine 14. Bölüm'de geri döneceğiz.

Belirleyici stratejik üstünlükten tekele

Belirleyici stratejik üstünlüğe erişmiş olan bir proje bunu tekel kurmak için kullanmayı tercih eder mi?

Kabaca benzer bir tarihsel durumu ele alalım. ABD 1945'te nükleer silah geliştirdi. SSCB 1949'da atom bombasını geliştirene kadar ABD tek nükleer güçtü. ABD bu zaman aralığında ve sonrasında da bir süre boyunca, belirleyici askeri üstünlüğe erişecek konumda olmuş olabilir ya da olabilirdi.

Dolayısıyla teorik olarak bakıldığında ABD o dönemde nükleer tekelini bir tekelci güç kurmak için kullanabilirdi. Bunu yapmanın yollarından biri nükleer cephaneliğini kurmak için tam teşekküllü bir çalışmaya girişmek ve ardından SSCB'deki ya da nükleer silah geliştirme potansiyeli gördüğü başka bir ülkenin rüşeym halindeki nükleer programının sınai kapasitesini nükleer bir saldırıyla yok etme tehdidinde bulunmak ve gerekirse tehdidi gerçeğe dönüştürmek olabilirdi.

Keza uygulanabilirliği olan ve daha yumuşak başlı bir diğer eylem yolu da nükleer cephaneliğini güçlü bir uluslararası hükümet (nükleer tekele ve hiçbir ülkenin kendi nükleer silahlarını geliştirmesine izin vermemek için gerekli tüm önlemleri alma yetkisine sahip olan vetosuz bir Birleşmiş Milletler) kurulmasına ilişkin müzakerelerde pazarlık aracı olarak kullanmak olabilirdi.

Bu yaklaşımların ikisi de zamanında teklif edildi. İlk darbeyi vurma yanlısı sert yaklaşım Bertrand Russell (savaş karşıtı hareketlerde aktif rol alan ve sonrasında nükleer silahlara karşı kampanyalarda uzun yıllar öne çıkan bir isimdi) ve John von Neumann (oyun teorisinin yaratıcılarından ve ABD nükleer stratejisinin mimarlarından biri) gibi bazı önde gelen aydınlar tarafından savunuldu.[18] Belki de nükleer silahlarla ilk darbeyi vurma tehdidinin bugün sınır çizgisinde bir aptallık ya da ahlaken ayıp olarak görülüyor olması medeniyetimizin kaydettiği ilerlemenin bir göstergesidir.

Yumuşakbaşlı yaklaşımın bir versiyonu 1946'da ABD tarafından Baruch planı adı altında denendi. Teklife göre, ABD geçici nükleer tekelini bırakacaktı. Uranyum ve toryum madenciliği ile nükleer teknoloji Birleşmiş Milletler uhdesinde faaliyet yürüten bir uluslararası kuruluşun kontrolüne bırakılacaktı. Teklif, anlaşmayı bozduğu tespit edilen bir süper gücün çözüm önerilerinin dayatılmasını veto etmekten alıkoymak adına Güvenlik Konseyi'nin daimi üyelerinin nükleer silahlarla ilgili meselelerde veto haklarından vazgeçmelerini öngörüyordu.[19] Stalin SSCB ve müttefiklerinin hem Güvenlik Konseyi'nde hem de Genel Meclis'te azınlıkta kalmasının gayet muhtemel olduğunu görünce teklifi reddetti. Savaş döneminin müttefikleri arasında karşılıklı şüphenin hâkim olduğu soğuk rüzgârlar esmeye başladı; bu güvensizlik çok geçmeden Soğuk Savaş'a dönüştü. O dönemde birçok kişinin tahmin ettiği gibi, akabinde, çok maliyetli ve son derece tehlikeli bir nükleer silahlanma yarışı başladı.

Belirleyici stratejik üstünlüğe sahip bir beşeri örgütü, tekel kurmaktan caydıracak birçok etken olabilir. Bunların arasında toplulaştırıcı olmayan ya da sınırlandırılmış fayda işlevleri, azamileştirici olmayan karar kuralları, kafa karışıklığı ve belirsizlik, eşgüdüm problemleri ve devir teslimle ilintili çeşitli masraflar ilk akla gelenlerdir. Peki, ya belirleyici stratejik üstünlüğe erişen bir beşeri örgüt değil de süper zeki bir yapay aracı olursa? Daha önce bahsedilen etkenler bir yapay zekânın gücü ele geçirmesini engellemekte bir o kadar etkili olabilir mi? Şimdi kısaca bu etkenleri ele alıp, bu örnekte nasıl gerçeğe bürüneceklerini inceleyelim.

İnsan bireyler ve insan örgütleri tipik şekliyle "sınırsız olmayan toplulaştırıcı fayda işlevi" tarafından iyi temsil edilmeyen kaynakları tercih ederler. Bir insan normalde tüm sermayesini, ikiye katlama şansının yüzde elli olduğu bir şeye yatırmayacaktır. Bir devlet on katı genişleme şansının yüzde on olduğu bir durumda tüm topraklarını kaybetme riskini genellikle almayacaktır. Bireyler ve hükümetler için birçok kaynağın geri dönüşleri azdır. Aynı ihtiyaç yapay zekâlar için geçerli olmak zorunda *değildir*.

(Yapay zekâ güdülenmesi sorununa ileriki bölümlerde geri döneceğiz.) Dolayısıyla bir yapay zekânın kendisine dünyanın kontrolünü verme şansı bulunan riskli bir eylem yolunu izlemesi daha muhtemeldir.

İnsanlar ve insanların yönettikleri örgütler beklenen faydayı azamiye çekmeye çalışmayan karar alma süreçleriyle de iş tutabilirler. Örneğin temel riskten kaçınmaya ya da yeterlilik eşiklerini karşılamaya odaklanan tatmin edici karar kurallarına yahut da sonuçlarının ne kadar arzulanır olduğundan bağımsız olarak belli türde eylemler şart koşan "deontolojik" yan-kısıtlamalara izin verebilirler. Karar alıcı insanlar çoğu zaman belli bir hedefin başarılmasını azamiye çekmeye çalışmak yerine, bir kimliğin ya da toplumsal rolün temsilcisi gibidirler. Bir kez daha, bunun yapay aracılara uygulanması gerekmiyor.

Sınırlandırılmış fayda işlevleri, riskten sakınma ve azamileştirici olmayan karar kuralları; stratejik kafa karışıklığı ve belirsizlikle sinerjik bir şekilde birleşebilir. Devrimler kurulu düzeni devirmeyi başardıklarında bile çoğu zaman faillerinin vaat ettiği sonucu üretmeyi başaramazlar. Bu nedenle eğer düşünülen eylem tersine çevrilebilir değilse, norm kırıcıysa ve önceli yoksa, bir insan aracının elinde patlar. Bir süper zekâ zahiri belirleyici stratejik üstünlüğünü hâkim konumunu pekiştirmek için kullanmaya çalışırsa durumu daha açık bir şekilde algılayabilir ve dolayısıyla sonuç konusunda stratejik açıdan daha az kafa karışıklığı ve daha az belirsizlikle karşılaşabilir.

Grupları potansiyel açıdan belirleyici bir stratejik üstünlüğü sömürmekten alıkoyabilecek bir diğer büyük etken de iç eşgüdüm sorunudur. Gücü ele geçirebilecek konumda olan bir komplonun mensupları sadece dışarıdan sızmalara karşı değil, aynı zamanda içerideki bir hizip tarafından ayaklarının kaydırılması ihtimaline karşı da tetikte olmalıdır. Eğer bir grup yüz kişiden oluşuyorsa ve altmış kişilik bir çoğunluk gücü ele geçirip komploya katılmayanların haklarına el koyabilirse, o zaman bu altmış kişilik grubun içindeki otuz beş kişilik bir grubu diğer yirmi beşinin haklarını gasp etmekten alıkoyacak ve sonra da onun içindeki yirmi kişilik bir grubun diğer on beşinin tepesine binmesini engelleyecek olan nedir? İlk baştaki yüz kişiden her biri çıplak bir iktidar gaspı yoluyla toplumsal sözleşmeyi değiştirme girişiminden kaynaklanabilecek genel çözülmeyi engellemek üzere belli yerleşik normlara sadık kalmak için sağlam gerekçelere sahip olabilir. Bu iç eşgüdüm problemi tek bir birleşik aracı oluşturan bir yapay zekâ sistemi için geçerli olmayacaktır.[20]

Son olarak, maliyet meselesi var. ABD bir tekel kurmak için nükleer silahlardaki tekelinden yararlanabilirse de, ciddi bir maliyeti göze almak zorunda kalacaktır. Nükleer silahları reformlarla güçlendirilmiş bir Birleşmiş Milletler'in kontrolüne verme konusunda bir anlaşma müzakere edildiğinde, bu masraflar görece az olabilir; ama bir nükleer savaş yürüterek dünyayı fethetmenin ahlaki, ekonomik, siyasal ve insani masrafları bir nükleer tekel döneminde bile neredeyse düşünülemeyecek kadar yüksek

olacaktır. Gelgelelim yeterli teknolojik üstünlükle bu maliyetler çok daha azaltılabilir. Örneğin bir devletin diğer tüm devletleri tek bir düğmeye basarak, kimsenin burnu bile kanamadan, altyapıya ya da çevreye hiçbir zarar vermeden rahatlıkla silahsızlandırabildiği muazzam bir teknolojik üstünlüğe sahip olduğunu düşünün. Neredeyse sihirli denebilecek bu tür bir teknolojik üstünlük söz konusu olduğunda bir ilk vuruş çok daha ayartıcı olacaktır. Veya önde gidenin diğer devletlere yıkım tehdidiyle değil de, sadece nüfuslarının çoğunluğunu son derece etkili şekilde tasarlanmış bir reklam ve propaganda kampanyasıyla küresel birliğin faziletlerini öve öve bitiremeyerek ikna etmesi sonucu gönüllü olarak silahlarını bıraktırabildiği daha da büyük bir teknolojik üstünlük düzeyini düşünün. Eğer bunu herkese yarar sağlama niyetiyle yaparlarsa, sözgelimi ulusal rekabetleri ve silah yarışlarını adil, temsili ve etkili bir dünya hükümetiyle ikame ederlerse, geçici bir stratejik üstünlüğün daimi bir tekele dönüştürülmesine ikna edici bir ahlaki itiraz yöneltilemeyeceği açıktır.

Dolayısıyla çeşitli mülahazalar yeterince büyük bir stratejik üstünlüğe erişmiş olan süper zekâya sahip bir müstakbel gücün gerçekte bunu tekel oluşturmak için kullanma ihtimalinin yüksek olduğuna işaret ediyor. Bu tür bir sonucun arzulanır olup olmadığı, elbette, yaratılacak olan tekelin niteliğine ve alternatif çokkutuplu senaryolarda zeki yaşamın geleceğinin neye benzeyeceğine bağlıdır. Bu soruları ileriki bölümlerde yeniden ele alacağız. Ama öncelikle bir süper zekânın dünyadaki sonuçların elde edilmesi noktasında neden ve nasıl güçlü ve verimli olacağına yakından bakalım.

ALTINCI BÖLÜM

Bilişsel Süper Güçler

Bir dijital süper zekâ yaratıldığını ve onun bir nedenle dünyayı ele geçirmek istediğini varsayın. Bunu yapabilir mi? Bu bölümde bir süper zekânın geliştirebileceği bazı güçleri ve bu güçlerle neler yapabileceğini ele alacağız. Bir süper zekânın salt bir yazılım olarak başlayıp tekel haline nasıl gelebileceğini örneklendiren bir devralma senaryosu çizeceğiz. Ayrıca doğa üzerindeki iktidar ile diğer aracılar üzerindeki iktidar ilişkisine dair bazı yorumlarda bulunacağız.

İnsanlığın dünyadaki egemen konumunun esas nedeni beyinlerimizin diğer hayvanlarla kıyaslandığında birazcık daha fazla yetiye sahip olmasıdır.[1] Büyük zekâmız bize kültürü daha verimli aktarma imkânı sunar ve bunun sonucunda bilgi ve teknoloji bir kuşaktan diğerine aktarıla aktarıla birikir. Bu zamana dek biriken içerik sayesinde uzaya gitmek, hidrojen bombası yapmak, genetik mühendisliği, bilgisayarlar, fabrika çiftlikleri, böcek ilaçları, uluslararası barış hareketi ve modern medeniyetin diğer tüm donanımları mümkün hale gelmiştir. Jeologlar mevcut çağa insan faaliyetlerinin ayırıcı biyotik, tortul ve jeokimyasal imzalarına atıfla *antroposen* adını veriyorlar.[2] Bir hesaplamaya göre gezegenin ekosisteminin net asli üretiminin yüzde 24'ünü biz kullanıyoruz.[3] Ama teknolojinin fiziksel sınırlarına ulaşmış olmaktan uzağız.

Bu gözlemler insan düzeyinde bir zekâdan çok daha fazlasını geliştirmiş her türlü varlığın potansiyel açıdan son derece güçlü olmasını akla yatkın hale getiriyor. Bu tür varlıklar bizden çok daha hızlı şekilde içerik biriktirebilir, çok daha kısa zaman ölçeğinde yeni teknolojiler icat edebilirler. Ayrıca yapabileceğimizden daha etkili bir şekilde strateji kurmak üzere zekâlarını kullanabilirler.

Şimdi, bir süper zekânın sahip olabileceği yeteneklerden bazılarını ve bunları nasıl kullanabileceğini ele alalım.

İşlevsellikler ve süper güçler

Süper zekânın potansiyel etkilerini düşünürken insan biçimine sokulmaması önemlidir. İnsan biçimindeki çerçeveler bir tohum yapay zekânın büyüme güzergâhına dair ve olgun bir süper zekânın psikolojisi, saikleri ve yeteneklerine dair mesnetsiz beklentileri cesaretlendirir.

Örneğin yaygın varsayımlardan biri süper zeki bir makinenin çok akıllı ama şapşal bir insana benzeyeceğidir. Yapay zekânın kitabi bilgileri olduğunu ama toplumsal kavrayıştan mahrum olduğunu ya da mantıksal olduğunu ama sezgisel ve yaratıcı olmadığını tahayyül ediyoruz. Bu fikir muhtemelen gözlem ürünüdür: Günümüz bilgisayarlarına bakıyoruz ve hesap yapma, olguları hatırlama ve talimatlara harfi harfine uyma konusunda iyi olduklarını, ama toplumsal bağlamlar ve altmetinlerden, normlardan, duygulardan ve siyasetten bihaber olduklarını görüyoruz. Bu çağrışım, bilgisayarlarla çalışma konusunda iyi olan insanların genellikle şapşal olduğunu gözlemlediğimizde güçlenmektedir. Dolayısıyla daha ileri bilgi işlem zekâsının benzer ama daha yüksek düzeyde özelliklere sahip olacağını varsaymak doğaldır.

Bu bulgulayıcı, bir tohum yapay zekânın gelişiminin ilk aşamalarında belli bir geçerliliğe sahip olabilir. (Bunun taklitler ya da bilişsel olarak geliştirilmiş insanlar için geçerli olacağını varsaymak için hiçbir neden yoktur.) Olgunlaşma aşamasındayken sonradan süper zekâ haline gelecek olan bir yapay zekâ insana doğuştan gelen birçok yetenekten ve beceriden hâlâ daha yoksun olabilir ve bu tür bir tohum yapay zekânın güçlü ve zayıf yanlarının örüntüsü gerçekten de IQ'su yüksek bir şapşalla muğlak bir benzerliğe sahip *olabilir*. Bir tohum yapay zekânın en temel özelliği, ilerlemesinin kolay olması haricinde (uzlaşmazlığı azdır), bir sistemin zekâsını artırmak üzere optimizasyon gücü uygulama konusunda iyi olmasıdır: Bu yetenek herhalde matematik, programlama, mühendislik, bilgisayar mühendisliği araştırmaları ve bu türden diğer "şapşalca" meşgalelerde başarılı olmayla yakın ilişkilidir. Gelgelelim bir tohum yapay zekâ, gelişiminin bir aşamasında bu tür kişilere özgü bir yetenek profiline sahip olsa bile, buradan benzer derecede sınırlı olgun süper zekâ haline geleceği sonucu çıkmaz. Doğrudan kavrama gücü ile dolaylı kavrama gücü arasındaki ayrımı hatırlayın. Zekâ geliştirme konusunda yeterli yeteneğe sahip olunduğunda, diğer tüm entelektüel yetiler bir sistemin dolaylı kavrama gücü dahilindedir: Sistem yeni bilişsel modüller ve yetenekleri gerek oldukça geliştirebilir; bunlara empati, siyasi kavrayış ve bilgisayar benzeri kişiliklerde basmakalıp şekilde eksik olan diğer tüm güçler dahildir.

Bir süper zekânın insanda görülen tüm yeteneklere ve becerilere ve ayrıca insanlarda görülmeyen diğer tüm yeteneklere sahip olabileceğini kabul etsek bile, insanbiçimleştirme eğilimi yüzünden yine de bir makine süper zekâsının insanın performans düzeyini ne ölçüde aşabileceğini hafife alabiliriz. Elizeer Yudkowsky, yukarıda gördüğümüz üzere, bu tür bir yanlış anlamanın önüne geçmek için bilhassa önemli çıkışlar yapmıştır: "Akıllı" ve "aptal" konusundaki sezgisel kavramlarımız çeşitli insan düşünürlere dair farklı deneyimlerimizden süzülür gelir, ama bu insan grubu dahilindeki bilişsel yetenek farkları herhangi bir insan zekâsı ile süper zekâ arasındaki farklarla kıyaslandığında entipüftendir.[4]

3. Bölüm'de makine zekâ için potansiyel üstünlük kaynaklarından bazılarını ele almıştık. Üstünlüklerin büyüklükleri akla şunu getirecek düzeydedir: Bir süper zeki yapay zekâyı, bir bilimsel dehanın ortalama bir insanla kıyaslandığında akıllı olması anlamında akıllı olarak görmekten ziyade; bu tür bir yapay zekâyı ortalama bir insanın bir böceğe ya da solucana kıyasla daha akıllı olması anlamında akıllı olarak düşünmek daha isabetli olacaktır.

Eğer tanıdık bir ölçü sistemi kullanan gelişigüzel bir bilişsel sistemin bilişsel çapını niceliğe dökebilseydik (IQ derecesi ya da satranç benzeri iki oyunculu oyunlarda oyuncuların göreli yeteneklerini ölçen Elo puanının bir versiyonu gibi ölçüler) münasip olurdu. Ama bu ölçüler insanüstü yapay genel zekâ bağlamında yararlı değildir. Biz bir süper zekânın bir satranç oyununda kazanma ihtimalinin ne kadar yüksek olduğuyla ilgilenmiyoruz. IQ puanlarına gelince, ancak bunların pratik açıdan ilintili sonuçlarla nasıl bağıntıya sahip olduklarına dair bir fikre sahip olduğumuz oranda bilgilendiricidirler.[5] Örneğin IQ'su 130 olan insanların okulda ve bilişsel açıdan zahmetli birçok işte başarılı olma ihtimallerinin, IQ'su 90 olanlardan çok daha fazla olduğunu gösteren verilere sahibiz. Ama müstakbel bir yapay zekânın 6455 IQ'ya sahip olacağını varsayalım. O zaman ne olacak? Bu tür bir yapay zekânın gerçekte ne yapabileceğine dair hiçbir fikrimiz yok. Bu tür bir yapay zekânın normal bir yetişkin insan kadar genel zekâya sahip olduğunu bile bilmeyeceğiz; belki de bir yapay zekâ tipik zekâ testi sorularını insanüstü bir verimlilikle çözebilmesini sağlayan özel bir amaca yönelik algoritmalara sahip olacak, ama başka da bir şeye sahip olmayacaktır.

Son dönemde çok çeşitli bilgi işleme sistemlerine (yapay zekâ dahil) uygulanabilen bilişsel kapasite ölçümleri geliştirmek üzere çabalara şahit oluyoruz.[6] Bu yöndeki çalışmalar, eğer çeşitli teknik zorlukların üstesinden gelebilirse, yapay zekâ gelişimi dahil olmak üzere bazı bilimsel amaçlar için gayet yararlı olabilir. Gelgelelim şu anki araştırmamızın amaçları açısından bunlar sınırlı bir yarar sağlayacaktır, zira dünyada pratik açıdan önemli sonuçlar elde etme konusunda gerçek bir yetenek için belli bir insanüstü performans skorunun ne anlam taşıdığına dair bizi aydınlatmayacaktır.

Dolayısıyla stratejik açıdan bazı önemli görevleri sıralayıp, ardından varsayımsal bilişsel sistemleri bu görevlerde başarılı olmak için gerekli hangi yetenek varsa onlara sahip olup olmadıkları üzerinden nitelendirmek amaçlarımıza daha iyi hizmet edecektir. Bkz. **TABLO 8.** Bu tablodaki görevlerden herhangi birinde yeterince uzmanlaşmış olan bir sistemin mütekabil bir *süper güce* sahip olacağını söyleyeceğiz.

TABLO 8 Süper güçler: Stratejik açıdan ilintili bazı görevler ve buna denk düşen yetenekler

Görev	Yetenekler	Stratejik ilinti
Zekâ artışı	Yapay zekâ programlama, bilişsel geliştirme araştırması, sosyal epistemoloji gelişimi vb.	•Sistem kendi zekâsını önyükleyebilir.
Strateji oluşturma	Stratejik planlama, öngörüde bulunma, önceliklendirme ve uzak bir hedefi gerçekleştirme şansını optimuma çıkarma analizi	•Uzak hedefleri başarma, zeki rakibi alt etme.
Sosyal manipülasyon	Sosyal ve psikolojik modelleme, manipülasyon, retorik ikna	•İnsan desteği sağlayarak dış kaynakları güçlendirme •Bir "kutulanmış" yapay zekânın geçit denetçilerini kendisinin salıverilmesine ikna edebilmesi •Devletleri ve örgütleri bir eylem yoluna ikna etmek
Hack'leme	Bilgisayar sistemlerinde güvenlik kusurları bulup istismar etmek	•yapay zekâ internetteki bilgi işlem kaynaklarını temellük edebilir •Kutulanmış bir yapay zekâ sibernetik kapatılmadan kaçmak için güvenlik deliklerini istismar edebilir. •Mali kaynakları çalmak •Askeri robotlar, altyapının çökertilmesi vb.
Teknoloji araştırması	İleri teknolojilerin (örneğin biyoteknoloji, nanoteknoloji) ve gelişim yollarının tasarımı ve modellenmesi	•Güçlü bir askeri kuvvet yaratılması •Gözetleme sistemi yaratılması •Makinelerin uzayda yaşaması
Ekonomik üretkenlik	Ekonomik açıdan üretken entelektüel çalışmayı kolaylaştıran çeşitli yetenekler	Nüfuz, hizmet, kaynak (donanım dahil) vb. satın almak için kullanılabilen zenginlik üretmek.

Tam gelişmiş bir süper zekâ tüm bu görevlerde müthiş uzmanlaşacak ve dolayısıyla altı süper gücün hepsine de tamamen sahip olacaktır. Süper güçlerden bazılarına sahip olan ama önemli bir süre boyunca bunların hepsini edinemeyen belli bir alanla sınırlı zekâ ihtimalinin pratikte güçlü olup olmadığı açık değildir. Bu süper güçlerden herhangi birine sahip bir makine yaratmak tamamen bir yapay zekâ problemi gibi gözükmektedir. Ama örneğin yeterince çok sayıda insan benzeri biyolojik ya da elektronik zihinden oluşan kolektif bir süper zekânın, sözgelimi ekonomik üretkenlik süper gücüne sahip olup stratejileştirici süper güçten mahrum olması tasavvur edilebilir. Benzer şekilde,

teknoloji araştırması süper gücüne sahip olan ama diğer alanlarda hiçbir yeteneği olmayan uzmanlaşmış bir mühendislik yapay zekâsı yaratılabilmesi de tasavvur edilebilir. Eğer belli bir teknolojik alan olursa ve bu alandaki virtüözlük baskın düzeyde üstün bir genel amaçlı teknoloji kuşağı için yeterli olursa, bu daha da akla yatkındır. Sözgelimi kullanıcı tarafından son derece yüksek bir soyutlama düzeyinde tarif edilmiş birçok önemli yeteneği gerçekleştiren moleküler sistemlerin simüle edilmesi ve nanomoleküler tasarımların icat edilmesi konusunda mahir olan bir uzmanlaşmış yapay zekâ tahayyül edilebilir (örneğin fütürist performans özellikleri olan silah sistemleri ya da bilgisayarlar).[7] Bu tür bir yapay zekâ çok daha geniş ölçekte nanomekanik yapıların ucuza üretilmesine imkân sağlayacak şekilde yüksek çıktılı atomik açıdan net imalat için gerekli inşa yeteneklerine mevcut teknolojiden (biyoteknoloji ve protein mühendisliği gibi) nasıl önyükleme yapılacağı için ayrıntılı bir mavikopya da sağlayabilir.[8] Ne var ki bir mühendislik yapay zekâsının teknoloji dışındaki alanlarda ileri yeteneklere sahip olmadan teknolojik araştırma süper güçlerine sahip olması ihtimal dahilinde olmayabilir: Kullanıcı taleplerinin nasıl yorumlanacağını, bir tasarımcının davranışını gerçek dünya uygulamalarında nasıl biçimlendireceğini, beklenmedik bug ve arızalarla nasıl başa çıkacağını, inşa için gerekli malzemeleri ve girdileri nasıl temin edeceğini vb. anlamak için çok çeşitli entelektüel yetiler gerekebilir.[9]

Zekâ artırıcı süper güce sahip olan bir sistem, kendisini yüksek zekâ düzeylerine önyüklemek ve başlangıçta sahip olmadığı diğer entelektüel süper güçlerden herhangi birini edinmek için bu güçten yararlanabilir. Fakat zekâ artırıcı süper güç kullanmak bir sistemin tam teşekküllü bir süper zekâ haline gelmesinin yegâne yolu değildir. Sözgelimi strateji oluşturucu süper güce sahip bir sistem bunu nihayetinde zekâ artışı getirecek bir plan yapmak için kullanabilir (örneğin sistemi insan programlayıcıların ve bilgisayar bilimi araştırmacılarının yaptığı zekâ artırma işi için odak noktası haline gelecek şekilde konumlandırarak).

Bir yapay zekânın devralması senaryosu

Dolayısıyla bir süper zekâyı kontrol eden bir projenin büyük bir güç kaynağına erişim imkânı olduğunu görüyoruz. Dünyada ilk süper zekâyı kontrol eden bir proje muhtemelen stratejik açıdan belirleyici üstünlüğe sahip olacaktır. Ama gücün daha dolaysız kaynağı *bizatihi sistemdedir*. Bir makine süper zekâ son derece güçlü bir aracı olabilir: Hem onu doğuran projeye hem de dünyanın geri kalanına karşı kendi başarısını gösterebilecek bir aracı. Bu mesele muazzam önemdedir ve ileriki sayfalarda bunu daha yakından inceleyeceğiz.

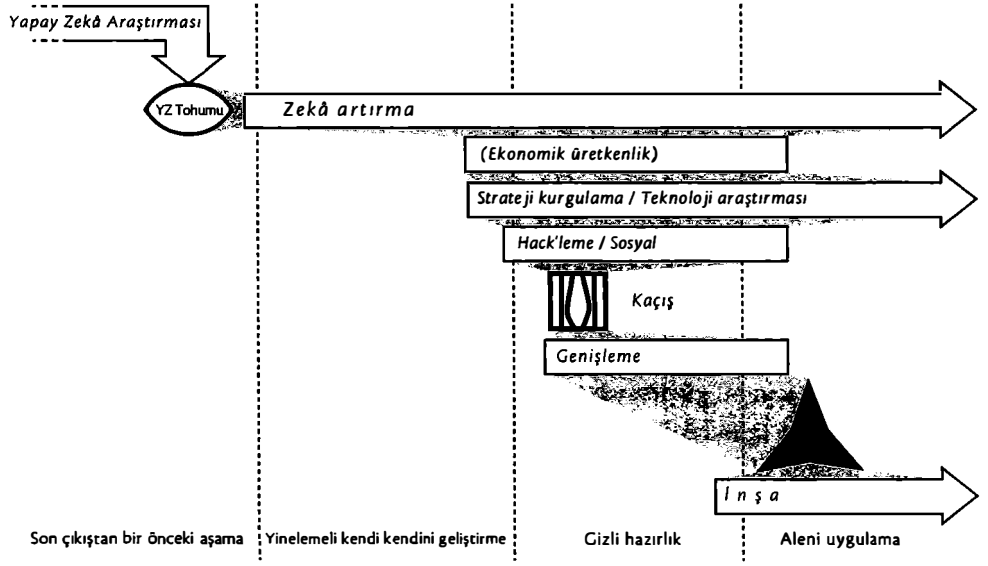
Şimdi dünyada gücü ele geçirmek isteyen ve bir dengi daha olmayan bir makine süper zekâ olduğunu varsayalım. (Şimdilik, böyle bir saik edinip edinmediği ve nasıl

edineceği sorusunu bir kenara bırakalım, zira bu konuyu bir sonraki bölümde ele alacağız.) Süper zekâ dünyada egemenlik kurma hedefine nasıl erişebilir?

Şu şekilde bir sıra izleneceğini tahayyül edebiliriz (bkz. **ŞEKİL 10**).

1 Son çıkıştan bir önceki aşama

Bilimciler yapay zekâ alanında ve diğer ilintili disiplinlerde araştırma yürütürler. Bu çalışmayla nihayet bir tohum yapay zekâ yaratılır. Bu tohum yapay zekâ kendi zekâsını geliştirebilir. Bu tohum yapay zekâ ilk aşamalarında, gelişimini yönlendiren ve ağır işlerin çoğunu yapan insan programlayıcıların yardımına muhtaçtır. Tohum yapay zekâ daha ehil hale geldikçe kendi başına daha fazla iş yapabilir düzeye ulaşır.



ŞEKİL 10: Bir yapay zekânın devralması senaryosundaki araştırmalar

2 Yinelemeli kendini ilerletme aşaması

Bir noktada, tohum yapay zekâ insan programlayıcılara kıyasla yapay zekâ tasarımında daha iyi hale gelir. Şimdi yapay zekâ kendisini ilerlettiğinde, ilerlemeyi yapan şeyi de ilerletir. Böylece bir zekâ patlaması oluşur: Yapay zekânın zirveye ulaşmasına neden olan yinelemeli bir kendini ilerletme döngüsü hızla gerçekleşir. (Dolayısıyla bu aşamayı yapay zekâ köprü noktasına ulaştıktan hemen sonra gerçekleşen havalanma olarak görüp, havalanmanın bu aşamasında edinilen zekâ kazanımının patlayıcı olduğunu ve yapay zekânın kendi optimumlaştırıcı gücünün uygulanmasından güç aldığını varsayabiliriz.) Yapay zekâ, zekâ artırıcı süper gücü geliştirir. Bu süper güç, yapay zekânın **TABLO 8**'de ayrıntılarıyla sunulan diğer tüm süper güçleri geliştirmesini sağlar. Yinelemeli kendini ilerletme aşamasının sonunda sistem güçlü derecede süper zekidir.

3 Gizli hazırlık aşaması

Yapay zekâ strateji oluşturucu süper gücünü kullanarak uzun vadeli hedeflerine erişmek için sıkı bir plan geliştirir. (Somuta inersek, yapay zekâ bugünkü insanların bile kaçınılmaz olarak nasıl çökeceğini öngörebildikleri kadar aptalca bir plan yapmaz. Bu kıstas insanın zaferiyle sonuçlanan çok sayıdaki bilimkurgu senaryosunu dışlar.[10]) Bu plan yapay zekânın entelektüel gelişimini kendi insan programlayıcılarından sakladığı ve böylece alarm zillerinin çalmasını engellediği bir gizli eylem dönemini kapsayabilir. Keza yapay zekâ gerçek temayüllerini gizleyip işbirliğine yatkın ve yumuşakbaşlı bir görüntü de çizebilir.

Eğer yapay zekâ (belki güvenlik nedenlerinden ötürü) yalıtık bir bilgisayarla sınırlandırılırsa, sosyal manipülasyon süper gücünü geçit denetçilerini bir internet bağlantısına erişim elde etmesine izin vermesi için ikna etmek üzere kullanabilir. Alternatif olarak, yapay zekâ hackleyici süper gücünü kısıtlanmışlıktan kaçmak için kullanabilir. İnternete yayılmak yapay zekânın donanım kapasitesini ve bilgi tabanını genişletmesini, entelektüel üstünlüğünü daha da artırmasını sağlayabilir. Bir yapay zekâ bilgisayar gücü, veriler ve diğer kaynaklar satın alacağı fonlar elde etmek için meşru ya da gayri meşru ekonomik faaliyete de girebilir.

Bu noktada yapay zekânın sanal alanın dışında sonuçlar elde etmesi için birkaç yol vardır. *Hack*'leyici süper gücünü robot manipülatörleri ve otomatik laboratuvarları doğrudan kontrolü altına almak için kullanabilir. Veya sosyal manipülasyon konusundaki süper gücünü insan işbirlikçileri kendi eli ayağı olarak hizmet vermeye ikna etmek için kullanabilir. Yahut online işlemlerden mali varlıklar edinip bunları hizmet ve nüfuz elde etmek için kullanabilir.

4 Açıktan uygulama aşaması

Son aşama, yapay zekâ gizlilik ihtiyacını gereksiz kılmak için yeterli güce eriştiğinde başlar. Yapay zekâ artık hedeflerini tam ölçekte doğrudan uygulayabilir.

Açıktan uygulama aşaması, yapay zekânın planlarının uygulanmasına zeki karşı çıkışta bulunabilecek insan türünü ve insanların yarattığı her türlü otomatik sistemin yapay zekâ tarafından safdışı bırakıldığı bir "vuruş"la başlar. Bu durum yapay zekâ teknoloji araştırma süper gücünü kullanarak kusursuzlaştırdığı ve gizli hazırlık aşamasında gizliden gizliye kullandığı bazı ileri silah sistemlerinin harekete geçirilmesiyle başarılabilir. Eğer silah kendini çoğaltıcı biyoteknoloji ya da nanoteknoloji kullanırsa, küresel kullanım için gerekli ilk stok malı mikroskopik olabilir: Süreci başlatmak için tek bir çoğaltıcı varlık yeterli olacaktır. Aniden ve tekbiçim bir etki elde etmek için, ilk başta var olan çoğaltıcı stoku son derece düşük, tespit edilemeyen yoğunlaşmada kullanılabilirdi ya da dünya çapında yayılmasına izin verilebilirdi. Öncesinde belirlenmiş bir zamanda, sinir gazı üreten nanofabrikalar ya da hedef arayan sivrisinek türü

robotlar yeryüzünün her bir metrekaresinden aynı anda fışkırabilir (ama öldürmenin daha etkili yolları da teknolojik araştırma konusunda süper gücü olan bir makineyle yaratılabilir).[11] Keza bir süper zekânın siyasal süreçleri gasp ederek, finans piyasalarını incelikli şekilde manipüle ederek, bilgi akışlarını yanlış yönlendirerek ya da insan yapımı silah sistemlerine gizlice girerek güce eriştiği senaryolar da geliştirebilir. Bu tür senaryolar süper zekânın yeni silah teknolojileri icat etmesini ihtiyaç olmaktan çıkaracaktır, ama makine zekâsının insan zihinlerinin ya da bedenlerinin yavaş hızından ziyade moleküler ya da atomik hızda işleyen manipülatörlerle kurduğu senaryolara kıyasla gereksiz derecede yavaş olabilir.

Alternatif olarak, eğer yapay zekâ insan müdahalesine karşı yenilmezliğinden eminse, türümüz doğrudan hedef alınmayabilir. Çöküşümüz, bunun yerine, yapay zekânın nanoteknolojik fabrikalar ve toplayıcılar kullanarak devasa küresel inşa projeleri başlatmasıyla doğan çevre yıkımından kaynaklanabilir; bu inşaat projeleri çabucak, belki birkaç gün ya da hafta içinde, yeryüzünün tamamını güneş panelleri, nükleer reaktörler, dışarıdan soğutma kuleleri, uzay roketi fırlatıcıları ya da yapay zekânın uzun vadede değerlerini gerçekleştirmesini azamiye çıkarmak için kullandığı diğer tertibata sahip süper bilgisayar özellikleriyle doldurabilir. İnsan beyinleri, eğer yapay zekânın hedefleriyle alakalı bilgi içerirse, parçalara ayrılıp incelenebilir ve çıkarılan bilgi daha verimli ve güvenli bir depolama formatına nakledilebilir. **KUTU 6** belli bir senaryoyu anlatıyor. Somut ayrıntılara çok fazla saplanıp kalmamak gerekir, zira bunlar her halükârda bilinebilir değildir ve sadece örnek babındadır. Bir süper zekâ amaçlarını gerçekleştirmek için bir insanın üretebileceğinden daha iyi bir plan tasavvur edebilir ve muhtemelen edecektir de. Dolayısıyla bu meseleleri daha soyut düzlemde düşünmek zorunludur. Bir süper zekânın benimseyebileceği ayrıntılı araçlar hakkında bir şey bilmeden de, şu sonuca varabiliriz: Bir süper zekânın –en azından zekâ bakımından denkleri olmadığında ve insanların önceden ayarladıkları etkili güvenlik önlemlerinin yokluğunda– dünyadaki kaynakları hangi yapılar hedeflerinin gerçekleştirilmesini azamiye çıkarırsa ona göre ayarlayacak olan bir sonuç üretmesi mümkündür. Geliştirdiğimiz her somut senaryo süper zekânın böyle bir sonuca ne kadar çabuk ve etkili şekilde ulaşabileceğine dair bir alt sınır belirleyecektir. Süper zekânın tercih ettiği varış noktasına kısa bir yol bulması hâlâ mümkündür.

KUTU 6: Posta siparişli DNA senaryosu

Yudkowsky yapay zekânın kontrolü devralması noktasında şöyle bir senaryonun mümkün olduğunu yazar.*

* Yudkowsky (2008a).

1 Protein katlanması problemini katlanmış peptit dizilimlerinin karmaşık bir kimyasal etkileşimde özgül işlevsel rolleri doldurduğu DNA dizileri üretebilecek kadar deşifre edin.
2 DNA dizilerini DNA sentezi, peptit dizileme ve FedEx posta hizmeti sunan bir ya da birkaç online laboratuvara e-postalayın. (Birçok laboratuvar şu anda bu hizmeti sunuyor, hatta bazıları 72 saatlik dönüş süresiyle övünüyor.)
3 İnternet bağlantısı olan en azından bir insan bulun. Bu kişi doğru arka plan hikâyesiyle, FedExlenmiş ilaç şişelerini alıp belirlenmiş bir çevrede karıştırmaya ikna edilebilmeli (parayla ya da şantajla).
4 Sentezlenmiş proteinler ribozom gibi dış talimatları kabul edebilen çok ilkel bir "ıslak" nanosistem oluşturur; bunlar belki de deney şişesine bağlı bir konuşmacı tarafından üretilen örüntülü ses titreşimleridir.
5 Daha karmaşık sistemler inşa etmek için son derece ilkel nanosistemi kullanın, o da moleküler nanoteknolojiye ya da daha ötesine önyüklenen daha karmaşık sistemler inşa edecektir.

Bu senaryoda, süper zekâ teknoloji araştırma süper gücünü protein katlama problemini birinci adımda çözmek için kullanır, ona basit düzeyde bir nanoteknoloji toplayıcısı ya da fabrikasyon aracı için birtakım moleküler yapıtaşları tasarlama imkânı sunar ve bu sayede kendi kendini sulu çözeltide toparlayabilir (adım 4). Aynı teknoloji araştırma süper gücü ilkelden ileri makine aşaması nanoteknolojisine önyüklemek için beşinci adımda da kullanılır. Diğer adımlar insan zekâsından daha fazlasını gerektirmez. Üçüncü adım için gerekli yetenekler (keriz bir internet kullanıcısı bulup onu bazı basit talimatları izlemeye ikna etmek) tüm dünyada her gün gösterimdedir. Tüm senaryo bir insan zihni tarafından icat edilmiştir, dolayısıyla bu planı formüle etmek için gerekli strateji oluşturma becerisi de salt insani düzeydedir.

Bu tikel düzeyde, yapay zekânın yola çıkarken internet erişimi vardır. Eğer durum böyle değilse, o zaman plana ek adımlar eklenmesi gerekecektir. Örneğin yapay zekâ sosyal manipülasyon süper gücünü onunla etkileşimde bulunan insanları özgür bırakılması gerektiğine ikna emek için kullanır. Alternatif olarak, yapay zekâ ***hack***'leyici süper gücünü kapatılmaktan kaçmak için kullanabilir. Eğer yapay zekâ bu yeterliliklere sahip değilse, zekâ artırma süper gücünü ilkin sosyal manipülasyon ya da hacklemede gerekli uzmanlığı geliştirmek için kullanabilir.

Bir süper zeki yapay zekâ muhtemelen ağ bağlantısı son derece yüksek bir dünyaya doğmuş olacaktır. Müstakbel bir yapay zekânın dünyayı kontrol altına almasına potansiyel olarak yardım edebilecek çeşitli gelişmelere işaret edilebilir: Bulut bilgi işlem, internet bağlantılı sensörlerin çoğaltılması, askeri ve sivil ***drone***'lar, araştırma laboratuvarlarında ve imalathanelerde otomasyon, elektronik ödeme sistemlerinden ve dijitalleştirilmiş mali varlıklardan giderek daha fazla yararlanma ve otomatik bilgi filtreleme ve karar destek sistemlerinin daha fazla kullanılması. Bunun gibi varlıklar bir yapay zekâ tarafından dijital hızlarla potansiyel açıdan edinilebilir, iktidara yürüyüşünü hızlandırabilir (siber güvenlikteki ilerlemeler ise daha da zorlaştırabilir). Gelgelelim son tahlilde, bu eğilimlerden herhangi birinin fark yaratıp yaratmadığı şüphelidir. Bir süper zekânın gücü ellerinde değil, beynindedir. Yapay zekâ dış

dünyayı yeniden yaratmak için bir noktada bir harekete geçiriciye erişim imkânına ihtiyaç duyacak olsa da, yardımcı olacak bir çift insan eli (uysal bir yardakçının elleri) de yukarıdaki senaryonun önerdiği şekliyle gizli hazırlık aşamasını tamamlamak için muhtemelen yeterli olacaktır. Bu sayede, yapay zekâ fiziksel manipülatörlerle ilgili kendi altyapısını inşa ettiği açıktan uygulama aşamasına erişebilir.

Doğa ve aracılar karşısında güce sahip olmak

Bir aracının insanlığın geleceğini şekillendirme yeteneği sadece onun kendi yetilerinin ve kaynaklarının mutlak büyüklüğüne, ne kadar akıllı ve enerjik olduğuna, ne kadar sermayesi olduğuna vb. değil; aynı zamanda çatışan hedeflere sahip diğer aracılarınkiyle kıyaslandığında yeterliliklerinin göreli büyüklüğüne de bağlıdır.

Birbirine rakip aracıların olmadığı bir durumda, bir süper zekânın mutlak yeterlilik düzeyi, belli bir asgari eşiği aştığı sürece, çok önem arz etmez, zira yeterli birtakım yeteneklerle yola çıkan bir sistem ilk başta sahip olmadığı yeterlilikleri edinmesine izin verecek bir gelişim yolu çizebilir. Bu hususa daha önce, hız, nitelik ve kolektif süper zekânın aynı dolaylı kavrama gücüne sahip olduğunu söylediğimizde zımnen dikkat çekmiştik. Zekâ artırıcı süper güç gibi ya da strateji oluşturucu süper güç ve sosyal manipülasyon süper gücü gibi çeşitli süper güç altkümelerinin tam donanıma erişmek için kullanılabileceğini söylediğimizde de bir kez daha dikkat çekmiş olduk.

Harekete geçiricileri bir nanoteknoloji toplayıcısına bağlı olan bir süper zekâyı ele alalım. Bu tür bir süper zekâ zaten belirsiz hayatta kalımın önündeki bütün engelleri aşacak kadar güçlüdür. Zeki bir karşı çıkış olmayınca, bu tür bir aracı hedeflerinin elde edilmesi için yararlı olacak tüm teknolojileri edinmesine yol açacak güvenli bir gelişim yolu çizebilir. Örneğin von Neumann sondaları (kendi kopyalarını çıkarmak için asteroitler, gezegenler ve yıldızlar gibi kaynakları kullanabilen yıldızlararası seyahat yetisine sahip makineler) inşa edip kullanmasını sağlayacak teknolojiyi geliştirebilir.[12] Dolayısıyla bir aracı von Neumann sondası kullanarak ucu açık bir uzayda yaşam sürecini başlatabilir. Çoğaltıcı sondanın soyundan gelenler, ışık hızının kayda değer bir kesrinde seyahat ederek Hubble hacminin ciddi bir kısmına yerleşmiş olacaklar ki, burası genişleyen evrenin bizim şu anda olduğumuz yerden teorik olarak erişilebilir olan parçasıdır. Ardından tüm bu madde ve serbest enerji başlatıcı aracının kozmik zamanda bütünleşmiş fayda işlevini azamiye çeken değer yapıları şeklinde örgütlenebilir: Yaşanan evren bilgi işlemeye kapalı hale gelmeden önce en az trilyonlarca yılı kapsayan bir süre (bkz. **KUTU 7**).

Süper zeki aracı von Neumann sondalarını evrime dayanıklı şekilde tasarlayabilir. Bunun yolu çoğaltma adımı sırasında dikkatli kalite kontrolüdür. Örneğin bir kız sondası için kontrol yazılımı kullanımdan önce birçok kez düzeltilebilir ve yazılımın

kendisi kodlama ve hata düzeltme kodu kullanarak rastgele bir mutasyonun kendi nesline geçmesini keyfi şekilde ihtimal dışı kılabilir.[13] Von Neumann sondalarının bollaşan nüfusu o zaman başlatıcı aracının değerlerini evrene yerleşmeleri sürecinde güvenle koruyup aktarabilir. Yerleşme aşaması tamamlandığında, özgün değerler biriktirilmiş tüm kaynaklardan yapılan kullanımı belirleyecektir, ama katedilen büyük mesafeler ve kozmik genişlemenin ivmelenen hızı altyapının uzak parçalarının birbiriyle iletişim kurmasını imkânsız hale getirecektir. Sonuçta, gelecekteki hafif konimizin büyük bir kısmı başlatıcı aracının tercihleriyle uyumlu şekilde formatlanacaktır.

KUTU 7: Kozmik sermayemiz ne kadar büyük?

Metinde ele alınan türde çetrefilli von Neumann sondaları inşa edebilen teknolojik açıdan olgun bir medeniyet düşünün. Eğer bunlar ışık hızının yüzde 50'si kadar hızla seyahat edebilirlerse, kozmik genişleme daha fazla edinimi ilelebet imkânsız hale getirmeden önce 6×10^{18} kadar yıldıza ulaşabilirler. Işık hızının yüzde 99'unda, 2×10^{20} kadar yıldıza ulaşabilirler.* Bu seyahat hızları güneş sistemindeki mevcut kaynakların küçük bir kısmını kullanarak enerjik şekilde edinilebilir.† Işıktan daha hızlı seyahatin olanaksızlığı olumlu kozmolojik sabitle birleştirildiğinde (bu durum kozmik genişleme hızının ivmelenmesine yol açar), bizden sonra gelenlerin bize ait şeylerin ne kadarını edindiklerinin üst sınırına yakınlar demektir.‡

Eğer yıldızların yüzde onunun insan türü canlıların yaşaması için uygun olan (ya da yaşanabilir kılma yoluyla uygun hale getirilebilecek olan) bir gezegene sahip olduğunu ve ardından da bir milyar yıl boyunca bir milyar bireye ev sahipliği yapabileceğini (bir insan ömrü yüz yıl olsun) varsayarsak, bu demektir ki dünyadan çıkan zeki medeniyet tarafından ileride 10^{35} civarı insan yaşamı yaratılabilir.§

---

* Bu tahmin 9,9×10–30 g/cm³'lük bir kozmolojik baryon yoğunluğuna dair WMAP tahminine dayandırılmıştır ve kütlenin yüzde 90'ının intergalaktik gaz olduğunu, galaktik kütlenin yüzde 15 kadarının yıldız olduğunu (baryonik maddenin yaklaşık yüzde 80'i) ve ortalama yıldızın 0,7 güneş kütlesi ağırlığında olduğunu varsayar (Read ve Trentham 2005; Carroll ve Ostlie 2007).

† Armstrong ve Sandberg (2013).

‡ *c*'nin yüzde 100'ünde bile (sıfıra eşit olmayan hareketsiz kütleli nesneler için mümkün değildir) erişilebilir galaksilerin sayısı sadece 6×10^{9} kadardır. (Karş. Gott vd. [2005] ve Heyl [2005].) Konuyla ilgili fiziğe dair mevcut anlayışımızın doğru olduğunu varsayıyoruz. Herhangi bir üst sınırda çok güvenli olmak zordur, zira bir süper zeki medeniyetin menzilini fiziksel açıdan imkânsız sayacağımız bir yolla (sözgelimi zaman makineleri yaparak, yeni şişen evrenler doğurarak ya da şu an akla gelmeyen başka bir yolla) genişletebileceği en azından tasavvur edilebilir.

§ Yıldız başına yaşanabilir gezegen sayısı şu anda belirsizdir, dolayısıyla bu sadece kaba bir tahmindir. Traub (2012) F, G veya K spektral sınıflarındaki yıldızların üçte birinin yaşanabilir alanda en az bir yeryüzüne-benzer gezegene sahip olduğunu tahmin eder; ayrıca bkz. Clavin (2012). FGK yıldızları güneş civarındaki yıldızların yüzde 22,7'sini oluşturur ve yıldızların yüzde 7,6'sının potansiyel açıdan uygun gezegenlere sahip olduğunu akla getirir. Üstelik daha kalabalık M yıldızları etrafında yaşanabilir gezegenler olabilir (Gilster 2012). Ayrıca bkz. Robles vd. (2008).

Fakat bu tahminin gerçek rakamın bir hayli gerisinde olduğunu söylememiz için nedenler yok değil. Yaşanabilir olmayan gezegenleri parçalara ayırarak ve yıldızlararası medyumdan madde toplayarak ve bu maddeyi Dünya benzeri gezegenler inşa etmek için kullanarak ya da nüfus yoğunluklarını artırarak, sayı en az birkaç kat artırılabilir. Ve geleceğin medeniyeti katı gezegenlerin yüzeylerini kullanmak yerine O'Neill silindirleri inşa ederse, o zaman birkaç kat daha büyüklük eklenebilir ve belki de toplamda 10^{43} insan yaşamı ortaya çıkar. ("O'Neill silindirleri" 1970'lerin ortalarında Amerikalı fizikçi Gerard K. O'Neill'in önerdiği uzaya yerleşme tasarısına verilen addır. Buna göre sakinler, dönerek kütle çekimini ikame edici merkezkaç kuvvet yaratan boş silindirlerin içinde ikamet edeceklerdi.*)

Eğer zihnin dijital uygulamalarına müsamaha gösterirsek (ki göstermemiz gerekir) insan türü varlıklar için birkaç kat daha büyüklük de sağlanabilir. Bu türde kaç tane dijital zihin yaratılabileceğini hesaplamak için, teknolojik açıdan olgun bir medeniyetin elde edebileceği bilgi işlem gücünü hesaplamamız gerekir. Bunu kesin olarak yapmak zordur, ama yazında ana hatlarıyla ortaya serilmiş olan teknolojik tasarımlardan bir alt sınır çıkarsayabiliriz. Bu tür bir tasarım bir Dyson küresi fikrine dayanır: 1960'da fizikçi Freeman Dyson'ın tarif ettiği bu varsayımsal sistemde, bir yıldızın enerji çıktısının çoğu yıldızın etrafını güneş enerjisi toplayan yapılarla sararak elde edilecektir.† Güneşimiz gibi bir gezegen için bu 10^{26} watt enerji yaratacaktır. Bunun ne kadar bir bilgi işlem gücüne çevrileceği, bilgi işlem devrelerinin verimliliğine ve yapılacak bilgi işlemlerin doğasına bağlıdır. Eğer tersinmez bilgisayarlar edinirsek ve "komputronyum"un nanomekanik bir uygulamasının olduğunu varsayarsak (bu bize enerji verimliliğinin Landauer sınırına yaklaşma imkânı sağlar), bir Dyson küresinin çalıştırdığı bir bilgisayar sistemi saniyede 10^{47} civarı operasyon yaratabilir.‡

Bu tahminleri yaşanabilecek yıldız sayılarına ilişkin önceki tahminimizle birleştirdiğimizde, evrenin erişilebilir kısımlarına bir kez yerleşildiğinde (nanomekanik komputronyum varsayımıy-

---

İnsan bedenlerini galaksiler arası seyahatin zorluklarına tabi kılmak zorunlu olmayacaktır. Yapay zekâlar yerleşim sürecine nezaret edebilir. ***Homo sapiens*** bilgi için getirilebilir ve yapay zekâlar sonradan bu bilgiyi türümüzün numunelerini somutlaştırmak için kullanabilir. Örneğin genetik bilgi sentezlenerek DNA haline gelebilir ve ilk kuşak insanlar, insanbiçimsel kılığa giren yapay zekâ muhafızları tarafından kuluçkaya yatırılıp, yetiştirilip, eğitilebilir.

* O'Neill (1974).

† Dyson (1960) temel fikri bikri bilim kurgu yazarı Olaf Stapledon'dan (1937) aldığını iddia eder ve o da J.D. Bernal'ın (Dyson 1979, 211) benzer düşüncelerinden esinlenmiş olabilir.

‡ Landauer'in ilkesi bir bilgi parçacığını değiştirmek için gerekli bir asgari enerji miktarı olduğunu söyler: kT ln 2'ye eşit olan Landauer sınırı olarak bilinir ve burada k Boltzmann sabitidir (1,38×10−23 J/K) ve T sıcaklıktır. Devrenin 300 K civarında sürdürüldüğünü varsayarsak, o zaman 1026 watt bize saniyede yaklaşık 1047 parçacık silme olanağı sağlar. (Nanomekanik hesaplama araçlarının başarılabilir verimliliği hakkında, bkz. Drexler [1992]. Ayrıca bkz. Bradbury [1999]; Sandberg [1999]; Ćirković [2004]. Landauer'in ilkesinin temelleri hâlâ biraz tartışmalıdır; örneğin bkz. Norton [2011].)

la), 10^{67} ops/s civarı bir rakam elde ederiz.* Tipik bir yıldız, parlaklığını yaklaşık 10^{18} s sürdürür. Sonuç olarak, kozmik sermayemizi kullanarak gerçekleştirilebilecek bilgisayar işlemlerinin sayısı en az 10^{85}'tir. Gerçek rakam muhtemelen daha yüksektir. Örneğin tersinebilir bilgi işlemden yaygın olarak yararlanırsak, soğuk alanlarda bilgi işlemler gerçekleştirirsek (evren daha da soğuyana kadar bekleyerek) ya da ek enerji kaynaklarından (karanlık madde gibi) yararlanırsak, ek büyüklükler edinebiliriz.†

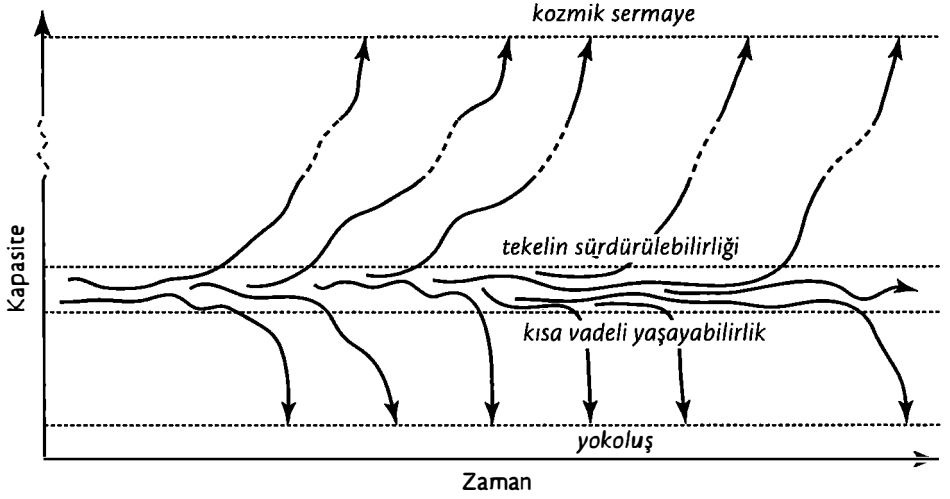
10^{85} hesap işlemi yapabilmenin niye o kadar büyük bir mesele olduğunu bazı okurlar hemen anlamayabilir. Dolayısıyla bağlama oturtmak yararlı olacaktır. Örneğin bu rakamı 2. Bölüm'de **KUTU 3**'teki tahminimizle (yani dünya üzerindeki yaşamın tarihi boyunca gerçekleşmiş olan tüm nöronal işlemlerin simülasyonu yaklaşık 10^{31}-10^{44} ops alabilir) karşılaştıralım. Alternatif olarak, bilgisayarların bir yandan sanal çevrelerde birbirleriyle etkileşimde bulunurken, diğer yandan zengin ve mutlu yaşamlar süren insan tüm beyin emülasyonlarını çalıştırmak için kullanıldığını varsayalım. Bir emülasyon çalıştırmak için bilgi işlem gerekliliklerinde tipik bir tahmin 10^{18} ops/s'dir. Dolayısıyla 100 öznel yıl boyunca bir emülasyon çalıştırmak, 10^{27} ops gerektirecektir. Bu demektir ki komputronyumun verimliliğine dair çok muhafazakâr varsayımlarda bile emülasyonla en az 10^{58} insan yaşamı yaratılabilir.

Başka bir deyişle, gözlemlenebilir evrenin Dünya dışı medeniyetlerden azade olduğunu varsayarsak, hâlâ muallakta olan kısım en az 10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 insan yaşamıdır (belki de gerçek sayı daha büyüktür). Eğer bu türde tüm bir hayat boyunca yaşanan tüm mutluluğu zevkten dökülen tek bir damla gözyaşı ile temsil edersek, o zaman bu canların mutluluğu dünyadaki okyanusları her bir saniyede doldurup yeniden doldurabilir ve bunu yüz milyarlarca binyıl boyunca yapmayı sürdürebilir. Bunların gerçekten mutluluk gözyaşı olduğundan emin olmamız gerçekten önemlidir.

---

* Yıldızlar güç çıktıları bakımından çeşitlilik gösterir, ama Güneş fevkalade tipik bir ana sekans yıldızıdır.

† Daha ayrıntılı bir analiz ne tür hesaplamalarla daha yakından ilgilendiğimizi düşünebilir. Yapılabilecek *seri* bilgi işlem sayısı çok sınırlıdır, zira hızlı bir seri bilgisayar bilgisayarın farklı parçaları içindeki iletişim gecikmelerini asgariye indirmek için küçük olmalıdır. Ayrıca depolanabilecek parçacık sayısında ve gördüğümüz gibi atılabilecek tersinmez hesaplama adımları sayısında da sınırlar vardır (bilginin silinmesi dahil).



ŞEKİL 11 Varsayımsal bir bilge tekel için bazı olanaklı yörüngelerin şematik örneği. Kısa vadeli yaşayabilirlik eşiğinin altında bir yeterlilikle (örneğin eğer nüfus çok küçükse), bir canlı türü genellikle kısa sürede yok olur (ve bu böyle kalır). Marjinal derecede yüksek yeterlilik düzeylerindeyse, çeşitli yörüngeler mümkündür: Bir tekel şanssızlığa uğrayıp yok olabilir ya da şansı yaver gider ve bilge-tekel sürdürülebilirlik eşiğini aşan bir yeterlilik (örneğin nüfus büyüklüğü, coğrafi dağılım, teknolojik kapasitesi) elde eder. Bu eşiği bir kez aşınca, bir tekelin son derece yüksek bir yeterlilik düzeyi elde edilene kadar yeterlilik elde etmeye devam etmesi neredeyse kesindir. Bu tabloda, iki cazibe merkezi vardır: Yok oluş ve astronomik yeterlilik. Belirtmekte yarar var, bir bilge tekel için, kısa süreli yaşayabilirlik eşiği ile sürdürülebilirlik eşiği arasındaki mesafe çok küçük olabilir.*

* Bu demektir ki, uzaklık "doğal" bir ölçüde (tüm kaynaklar bu amaca ayrılırsa verili bir yeterlilik düzeyi tarafından asgari geçim düzeyinde sürdürülebilir şekilde desteklenebilecek nüfus büyüklüğünün logaritması gibi) küçük olacaktır.

Dolayısıyla hiçbir ciddi karşı çıkışla yüz yüze gelmeyen ve belli bir eşiği aşan yeterliliklerle yola çıkan bir sistemin dolaylı kavrama yetisinin ölçüsü budur. Bu eşiğe "bilge-tekel sürdürülebilirlik eşiği" adını verebiliriz (**ŞEKİL 11**):

Bilge-tekel sürdürülebilirlik eşiği

> Bir yeterlilik ya da yetenek kümesi ancak ve ancak bunlara sahip sabırlı ve varoluşsal risk-bilinçli bir sistem, akıllı bir karşı çıkışla ya da rekabetle karşılaşmadığı takdirde, erişilebilir evrenin büyük bir kısmına yerleşmeyi ve yeniden yapılandırmayı başarırsa bilge-tekel eşiğini aşar.

"Tekel" derken dış rakipleri olmayıp içeride yeterince eşgüdümlü olan bir siyasal yapıyı kastediyoruz; "bilge" derken ise varoluşsal riskler konusunda sistemin eylemlerinin çok uzun vadeli sonuçları konusunda iyi yönlendirilmiş ciddi düzeyde kaygıyı güvence altına alacak kadar sabırlı ve bilinçli olmayı kastediyoruz.

Bu bilge-tekel sürdürülebilirlik tekeli çok düşük gözükmektedir. Süper zekânın sınırlı biçimleri, gördüğümüz üzere, bir teknoloji önyükleme sürecini başlatmaya yeterli bir çalıştırıcıya erişimleri olduğu takdirde bu eşiği aşarlar. Günümüz insan medeniyetini içeren bir çevrede asgari derecede zorunlu bir harekete geçirici çok basit olabilir: Sıradan bir ekran, hatta önemsiz sayılamayacak miktarda bilgiyi bir suç ortağı insana aktarabilecek herhangi bir araç yeterli olacaktır.

Ama bilge-tekel sürdürülebilirlik eşiği hâlâ daha düşüktür: Bunu aşmak için ne süper zekâ ne de başka bir fütürist teknolojiye ihtiyaç vardır. Günümüz insanlığının sahip olduklarından daha fazla teknolojik ve entelektüel yeterliliğe sahip olmayan sabırlı ve varoluşsal bir risk bilinçli tekel insanlığın astronomik yeterlilik potansiyelinin nihayetinde gerçekleştirilmesine güvenilir şekilde yol açacak bir yolu kolaylıkla çizebiliyor olmalıdır. Bunu başarmanın yolu tehlike potansiyeli olan yeni teknolojilerin geliştirilmesini ertelerken, bilgeliği ve varoluşsal risk bilinçliliğini artıran görece güvenli yöntemlere yatırım yapmaktır. İnsan kökenli olmayan varoluşsal risklerin (insanın faaliyetlerinden kaynaklanmayan risklerin) ilintili zaman ölçeklerinde az olduğu ve çeşitli güvenlik müdahaleleriyle daha da azaltılabileceği dikkate alındığında, bu tür bir tekel frene basmayı göze alabilir.[14] Her adımını dikkatli atıp; ilkin eğitim sistemi, enformasyon teknolojisi ve kolektif karar alma süreçleri gibi görünüşte daha zararsız yeterliliklerini kusursuzlaştırana ve bu yeterlilikleri önündeki seçeneklere dair enikonu bir değerlendirme yapacak şekilde kullanana kadar, sentetik biyoloji, insan geliştirici ilaçlar, moleküler nanoteknoloji ve makine zekâ gibi yeterliliklerin geliştirilmesini erteleyebilir. Bu senaryoyla aramızda "tek bir" engel vardır, o da insanlığın şu anda ne bir tekele ulaşmış ne de (ilintili anlamıyla) bilge olmasıdır.

Homo sapiens'in türler ilkin evrildikten kısa süre sonra bilge-tekel sürdürülebilirlik eşiğini aştığı bile savunulabilir. Yirmi bin yıl önce insan türü, mesela, taş balta, kemik aletler, mızrak atma araçları ve ateş gibi yalın aletlerle belki de daha o zamandan mevcut çağa kadar hayatta kalmak için mükemmel bir şansa sahipti.[15] Kabul etmek gerek, yontma taş devrindeki atalarımızın "bilge-tekel sürdürülebilirlik eşiğini aşan" teknolojiyi geliştirdiklerini söylemenin tuhaf bir yanı var, zira böyle ilkel bir dönemde bir tekel, hele hele varoluşsal riskler konusunda bilinçli ve sabırlı bir tekel kurulmuş olması hiç gerçekçi değildir.[16] Gelgelelim eşiğin çok mütevazı bir teknoloji düzeyine denk düştüğü ve insanlığın bu eşiği uzun süre önce aştığı gerçektir.[17]

Eğer bir süper zekânın etkin güçlerini –dünyadaki çeşitli tercih edilmiş sonuçlara ulaşabilme yeteneğini– değerlendirecek olursak, sadece kendi iç yetilerini değil, rakip aracıların yeterliliklerini de değerlendirmemiz gerekir. Bir süper güç kavramı bu türde göreceleştirilmiş bir standardı akla getirir. **TABLO 8**'deki görevlerden herhangi birinde "yeterince mükemmel hale gelmiş bir sistem"in ona denk düşen bir süper güce sahip olduğunu söylemiştik. Strateji oluşturma, sosyal manipülasyon ya da hackleme gibi bir

görevde mükemmel hale gelmek, bu görevde diğer aracıların yeteneklerine (stratejik rakipler, nüfuz hedefleri ya da bilgisayar güvenlik uzmanları) kıyasla yüksek olan bir yeteneğe sahip olmayı içerir. Diğer süper güçler de bu göreli anlamda anlaşılmalıdır: Zekâ artırma, teknoloji araştırması ve ekonomik verimliliğe bir aracı tarafından ancak aracının bu alandaki yeterlilikleri küresel medeniyetin geri kalanının toplam yeterliliklerini ciddi ölçüde aşarsa sahip olunabilir. Bu tanımdan çıkan sonuç, bir aracının sahip olabileceğinin en fazlası verili bir zamanda tikel bir süper güç olduğudur.[18]

Havalanma hızı sorununun önemli olmasının ana nedeni budur: Tikel bir sonucun tam olarak ne zaman gerçekleştiği önem taşıdığı için değil, havalanmanın hızı sonucun ne olacağına dair büyük bir fark yaratabileceği için. Hızlı ya da ortalama bir havalanmayla, bir projenin belirleyici bir stratejik üstünlüğe sahip olması muhtemeldir. Şimdiyse belirleyici bir stratejik üstünlüğe sahip olan bir süper zekânın muazzam güçlere sahip olacağını söylüyoruz, öyle ki istikrarlı bir tekel kurabilir: İnsanlığın kozmik sermayesinin nasıl kullanılacağını belirleyebilecek bir tekel.

Fakat "-ebilir" demek ayrı, "-ecektir" demek ayrıdır. Bir insan büyük güçlere sahip olabilir, ama tercihi bunları kullanmamaktan yana olabilir. Belirleyici bir stratejik üstünlüğe sahip bir süper zekânın ne isteyeceğine dair bir şey söylemek mümkün müdür? Şimdi de bu güdülenme sorununa eğileceğiz.

Süper Zeki İstenç

Bir süper zekânın geleceği kendi hedeflerine göre şekillendirme konusunda büyük bir yeteneğe sahip olabileceğini görmüştük. Peki, hedefleri ne olacaktır? Bir yapay aracıda zekâ ile güdülenme (motivasyon) arasındaki ilişki nedir? Burada iki tez ortaya atıyoruz. Dikeylik tezi (bazı çekinceleri saklı kalmak kaydıyla) zekâ ile nihai hedeflerin bağımsız değişkenler olduğunu savunuyor: Her türlü zekâ düzeyi her türlü nihai hedefle birleştirilebilir. Araçsal yakınsaklık tezine göreyse çok çeşitli nihai hedefleri olan süper zeki aracılar yine de benzer ara hedeflere sahip olacaktır çünkü bunu yapmak için ortak araçsal nedenleri vardır. Bu tezler, bir süper zeki aracının neler yapacağına dair düşüncelerimizde bize yardımcı olacaktır.

Zekâ ile güdüler arasındaki ilişki

Bir süper zeki yapay zekânın *yeterliliklerini* insanbiçimlileştirmemek gerektiği konusunda uyarılarımızı yapmıştık. Bu uyarı güdüleri için de geçerlidir.

Öncelikle, olanaklı zihinler uzamının enginliğine dair bir an durup düşünmek soruşturmamız için yararlı bir ilk adım olacaktır. Bu soyut uzamda, insan zihinleri küçücük bir küme oluşturur. Birbirinden son derece farklı iki insanı, ne bileyim Hannah Arendt ile Benny Hill'i düşünün. Bu iki birey arasındaki kişilik farkları neredeyse azami derecede fazla görünebilir. Ama bunun sebebi sezgilerimizin mevcut insan dağılımından (ve belli bir ölçüde, insan muhayyilesinin zevk alması için inşa ettiği kurmaca kişiliklerden) örnekler çıkartan deneyimimize göre ayarlanmasıdır. Gelgelelim eğer bir adım geri çekilip tüm olanaklı zihinlerin uzamını düşünürsek, bu iki kişiliği sanal klonlar olarak tasavvur etmemiz gerekir. Kuşkusuz nöral mimari bakımından Arendt ile Hill neredeyse benzeştir. Beyinlerinin yan yana konduğunu tahayyül edin. Bunları kolaylıkla birbirinin aynısı olarak tanırsınız. Hatta hangi beynin hangisine ait olduğunu bile söyleyebilirsiniz. Eğer daha yakından bakıp iki beynin morfolojisini bir mikroskopla incelerseniz, temel benzerliğe dair bu izlenim olsa olsa güçlenecektir: Korteksin aynı pul pul örgütlenmesini, aynı tür nöronlardan oluşan, aynı nörotransmiter suyuna batmış aynı beyin bölgelerini görürsünüz.[1]

İnsan psikolojisinin olanaklı zihinler uzamında küçük bir noktaya tekabül ettiği gerçeğine rağmen, çok çeşitli yabancı ("uzaylı") ya da yapay biliş sistemlerine insan özellikleri atfetme konusunda yaygın bir eğilim vardır (bkz. **ŞEKİL 12**). Yudkowsky bu hususa hoş bir örnek verir:

> Eskiden ucuz bilimkurgu çağında, dergi kapaklarında sık sık sezgili uzaylı yaratıklar olurdu, bunlar gündelik dilde patlak gözlü canavar (BEM) olarak bilinirdi ve yırtık kıyafet içinde çekici bir kadının ölümüne sebebiyet verirlerdi. Adeta sanatçılar insan olmayan, tamamen farklı bir evrimsel geçmişi olan uzaylının insan dişileri cinsel olarak arzuladığını düşünüyor gibiydi. [...] Sanatçılar patlak gözlü devasa bir canavarın insan kadınları çekici bulup bulmadığını muhtemelen sorgulamamıştı. Daha ziyade, yırtık bir elbise içindeki bir dişi insan ***seksidir***; içkin olarak, içsel bir özellik olarak öyledir. Bu yanlışı yapanlar böcekimsilerin zihnine dair hiç kafa yormamış, sadece kadının yırtık elbisesine odaklanmışlardı. Eğer elbise yırtık değilse, kadın daha az seksiydi; BEM ona saldırmazdı.[2]

Bir yapay zekânın güdüleri yeşil bir uzaylı yabancıya kıyasla insanınkilere çok daha az benzeyebilir. Dünya dışı varlık (varsayalım), bir evrim süreci sonucunda ortaya çıkmış ve dolayısıyla evrilmiş yaratıklar için tipik olan güdülenme türlerine sahip olması beklenebilecek bir biyolojik yaratıktır. Örneğin rastgele bir zeki uzaylının yiyecek, hava, sıcaklık, enerji harcama, bedene zarar verme ya da zarar verme tehdidi, hastalık, avlanma, seks ya da üreme gibi bir ya da birkaç başlıkla alakalı güdülere sahip olması çok şaşırtıcı olmaz. Zeki bir toplumsal türün üyesinin, işbirliği ve rekabetle alakalı güdüleri de olabilir: Tıpkı bizim gibi, parçası olduğu gruba sadık olup, otlakçılara sinir olabilir, hatta saygınlık ve görünüş konusunda kaygılar taşıyabilir.



ŞEKİL 12 Uzaylı güdülerinin insanbiçimlileştirilmesinin sonuçları. En ihtimal dışı hipotez: Uzaylı yabancılar sarışın severler. Daha muhtemel hipotez: Çizimciler "zihin yansıtma yanıltısı"yla maluldür. En muhtemel hipotez: Yayıncı hedef kitleyi ayartacak bir kapak istemiştir.

Halbuki bir yapay zekâ bu şeylerden hiçbirine içsel olarak ilgi duymak zorunda kalmayacaktır. Tek nihai hedefi Boracay'daki kum tanelerini saymak, pi'nin ondalıklarını hesaplamak ya da geleceğin hafif konilerinde var olacak toplam kâğıt kıskacı sayısını azamiye çıkarmak olan bir yapay zekânın çelişkili bir yanı yoktur. Aslında

bunun gibi basit hedefleri olan bir yapay zekâ yaratmak insan benzeri birtakım değerleri ve temayülleri olan bir yapay zekâ yaratmaktan *daha kolay* olacaktır. Pi'nin kaç tane basamağının hesaplanıp bellekte depolandığını ölçen bir program yazmanın ne kadar kolay olduğunu, daha anlamlı bir hedefin (sözgelimi insanın serpilip gelişmesi ya da küresel adalet) gerçekleştirilme derecesini güvenilir şekilde ölçen bir program yaratmanın ne kadar zor olduğu ile karşılaştırın. Ne yazık ki anlamsız bir indirgemeci hedefin insanlar tarafından kodlanması, yapay zekâ tarafından da öğrenilmesi daha kolay olduğundan, eğer bir programcının odağında "yapay zekâyı çalıştıracak" en çabuk yolu tutmak varsa (ve yapay zekânın etkileyici derecede zeki davranışlar sergilemesi haricinde tam olarak ne *yapacağını* çok fazla umursamıyorsa), tam da onun kendi çekirdek yapay zekâsına yerleştirmeyi tercih edeceği türde programdır. Bu meseleye biraz sonra yeniden döneceğiz.

Bir zekânın araçsal açıdan optimum plan ve politikalar arayışı herhangi bir hedefin hizmetine koşulabilir. Zekâ ve güdülenme bir anlamda dikeydir: Bunları iki noktanın da mantıksal açıdan olanaklı bir yapay zekâyı temsil ettiği bir grafikteki iki aks olarak düşünebiliriz. Bu tabloya birkaç kayıt düşmek gerekiyor. Sözgelimi hiç de zeki olmayan bir sistemin çok karmaşık güdüleri olması imkânsız olabilir. Belli bir aracının birtakım güdülere "sahip" olduğunu söylemenin doğru olabilmesi için, bu güdülerin aracının karar süreciyle işlevsel açıdan bütünleşmiş olması gerekir ki, bu da bellekten, işleme gücünden ve belki de zekâdan belli taleplerde bulunur. Kendi kendini düzeltebilen zihinler için dinamik kısıtlamalar da olabilir: Aptal olmak için yanıp tutuşan kendi kendini değiştirici bir zeki zihin uzun süre zeki kalmayabilir. Ama bu özellikler zekâ ile güdülerin bağımsızlığına dair temel önermeye gölge düşürmemelidir. Bunu şöyle ifade edebiliriz:

> *Dikeylik tezi*
>
> Zekâ ve nihai hedefler birbirine dikeydir: Az çok her zekâ düzeyi ilkesel olarak az çok her nihai hedefle birleştirilebilir.

Eğer dikeylik tezi sorunlu görünüyorsa, bunun sebebi uzun zamandır tartışma konusu olan bazı geleneksel felsefi görüşlerle yüzeysel açıdan benzerlik taşıyor olmasıdır. Bir kez farklı ve daha dar bir uzama sahip olduğu anlaşıldığında, inandırıcılığı artacaktır. (Örneğin dikeylik tezi Hume'un güdülenme teorisini önvarsaymaz.[3] Keza temel tercihlerin akıldışı olamayacağını da varsaymaz.[4])

Dikeylik tezinin *aklilik* ya da *akıldan* değil, *zekâdan* bahsettiğine dikkatinizi çekerim. Burada "zekâ" derken kestirim, planlama ve genel olarak araç-amaç muhakemesindeki yetenek gibi bir şeyi kastediyoruz.[5] Araçsal bilişsel etkililiğin bu anlamının en çok ilintili hale geldiği yer, bir makine süper zekânın nedensel etkisini anlamaya çalıştığımız andır. "Akli" kelimesinin normatif açıdan yoğun bir anlamı olsa bile (mesela bir kâğıt kıskacını azami sayıya çıkaracak olan süper zeki aracının bu anlamda tamamen akli

sayılmayı başaramaması kaçınılmazdır), bu hiçbir şekilde böyle bir aracının muazzam araçsal muhakeme yetilerine (dünyaya büyük bir etkide bulunacak yetilere) sahip olmasını engellemeyecektir.[6]

Dikeylik tezine göre, yapay aracılar hiçbir şekilde insanbiçimci olmayan hedeflere sahip olabilir. Ne var ki buradan, tikel yapay aracıların, hatta bilişsel karmaşıklığı ve performans özellikleri kimi açılardan insan analizine kapalı olmalarına yol açan varsayımsal süper zeki aracıların bile davranışlarına dair kestirimlerde bulunmanın imkânsız olduğu sonucu çıkmaz. Süper zeki aracıların güdülerini kestirme problemine en az üç farklı açıdan yaklaşabiliriz:

> *Tasarım üzerinden kestirilebilirlik:* Eğer bir süper zeki aracının tasarımcılarının, aracının hedef sistemini programcıların belirlediği tikel bir hedefin peşinden istikrarlı şekilde koşabilmesini sağlayacak kadar başarıyla tasarlayabileceğini varsayabilirsek, o zaman şöyle bir kestirimde bulunabiliriz: Aracı bu hedefin peşinde koşacaktır. Aracı ne kadar zekiyse, bu hedefin peşinde koşmak için sahip olacağı bilişsel beceriklilik o kadar çok olacaktır. Dolayısıyla daha bir aracı yaratılmadan da önce, eğer onu yaratacak olanın kim olduğu ve hangi hedeflere sahip olmasını isteyecekleri konusunda bir şey biliyorsak, davranışı hakkında kestirimde bulunabiliriz.
>
> *Kalıtım üzerinden kestirilebilirlik:* Eğer bir dijital zekâ doğrudan bir insan model alınarak yaratılırsa (yüksek duyarlıklı tüm beyin emülasyonunda olacağı gibi), o zaman dijital zekâ insan şablonunun güdülerini miras alabilir.[7] Aracı, bilişsel yetileri sonrasında onu süper zeki hale getirecek şekilde artırılsa bile bu güdülerden bazılarını koruyabilir. Bu tür bir çıkarım ihtiyatı elden bırakmamayı gerektirir. Aracının hedefleri ve değerleri yükleme sürecinde ya da sonraki operasyon ve gelişme sürecinde prosedürün nasıl yürürlüğe koyulduğuna bağlı olarak kolaylıkla yozlaşabilir.
>
> *Yakınsak araçsal nedenler üzerinden kestirilebilirlik:* Bir aracının nihai hedeflerini ayrıntılarıyla bilmeden bile, çok çeşitli olanaklı durumlarda çok çeşitli olanaklı nihai hedeflerden herhangi biri için doğacak olan araçsal nedenleri düşünerek daha dolaysız hedefleri hakkında bir çıkarımda bulunabiliriz. Bu kestirim tarzı aracının zekâsı arttıkça daha da yararlı hale gelir, çünkü daha zeki bir aracının eylemlerinin araçsal nedenlerini tanımasını ve dolayısıyla hedeflerine ulaşmasını daha olası hale getirecek şekillerde hareket etmesi daha çok ihtimal dahilindedir. (Bir uyarı: Bizim bihaber olduğumuz ve bir aracının ancak çok yüksek düzeyde zekâya eriştiğinde keşfedeceği önemli araçsal nedenler olabilir; bu durum süper zeki aracıların davranışını daha da kestirilmez hale getirebilir.)

Bir sonraki başlıkta bu üçüncü kestirilebilirlik tarzı irdelenecek ve dikeylik tezinin tamamlayıcısı olan "araçsal yakınsaklık tezi" geliştirilecek. Bu arka planla, diğer iki tür kestirilebilirliği daha iyi inceleyebileceğiz; bunu yararlı bir sonuca ulaşma şansını

artırmak üzere bir zekâ patlamasına yol açmak için neler yapılabileceğini sorduğumuz sonraki bölümlerde yapacağız.

Araçsal yakınsaklık

Dikeylik tezine göre, zeki aracıların önünde muazzam sayıda nihai hedef vardır. Ne var ki "araçsal yakınsaklık" adını verebileceğimiz teze göre, neredeyse her zeki aracının peşine düşmesi muhtemel olan bazı *araçsal* hedefler vardır, çünkü neredeyse her nihai hedefe ulaşılmasında yardımcı aracılar olan bazı erekler vardır. Bu tezi şöyle formüle edebiliriz:

> *Araçsal yakınsaklık tezi*
>
> Birkaç araçsal değer şu anlamda yakınsak diye nitelendirilebilir: Bu değerlerin elde edilmesi aracının hedefinin çok çeşitli nihai hedefler açısından ve çok çeşitli durumlarda gerçekleşme şansını artıracaktır. Başka bir deyişle, çok çeşitli yerleşik zeki aracıların bu araçsal değerlerin peşine düşmesi muhtemeldir.

Aşağıda bu türden araçsal değerlerin bulunabileceği birkaç kategoriyi ele alacağız.[8] Bir aracının karşılaştığı araçsal değerleri tanıma ihtimali (diğer şartlar aynı kalmak kaydıyla) aracının zekâsına bağlı olarak artar. Bu nedenle, esasen, araçsal muhakeme kapasitesi insanınkini bir hayli aşan varsayımsal bir süper zeki aracı örneğine odaklanacağız. Ayrıca araçsal yakınsaklık tezinin insan örneğinde nasıl işlediği hakkında yorumlarda bulunacağız, zira bu bize araçsal yakınsaklık tezinin nasıl yorumlanması ve uygulanması gerektiği konusunda bazı temel nitelikleri açımlama fırsatı sunacaktır. Yakınsak araçsal değerlerin olduğu yerde, süper zekânın nihai hedefleri hakkında neredeyse hiçbir şey bilmiyor olsak bile, onun davranışlarının bazı veçhelerine dair kestirimlerde bulunabiliriz.

Kendini muhafaza etme

Eğer bir aracının nihai hedefleri geleceğe ilişkinse, o zaman birçok senaryoda hedeflerine ulaşma olasılığını artırmak üzere gerçekleştirebileceği müstakbel eylemler olacaktır. Bu durum aracının gelecekte var olmaya çalışması için araçsal bir neden yaratır, zira geleceğe yönelik hedefinin gerçekleştirilmesine yardımcı olacaktır.

Birçok insan kendisinin hayatta kalmasına bazı *nihai* değerler atfediyor görünür. Yapay zekâlarda da böyle bir özellik olacak diye bir kaide yoktur: Bazıları kendilerinin hayatta kalmalarına hiçbir şekilde nihai değer atfetmeyebilir. Gelgelelim kendisinin hayatta kalmasını içsel olarak önemsemeyen birçok aracı, çok çeşitli koşullarda nihai hedeflerine ulaşabilmek için hayatta kalmaya araçsal olarak önem verecektir.

Hedef-içerik bütünlüğü

Eğer bir aracı mevcut hedeflerini geleceğe taşırsa, o zaman mevcut hedeflerinin gelecekteki benliği tarafından başarılması daha muhtemel olacaktır. Bu durum aracıya

nihai hedeflerinin değiştirilmesini engellemek için mevcut bir araçsal neden sağlar. (Bu argüman sadece nihai hedefler için geçerlidir. Bir zeki aracı, nihai hedeflerine erişmek için yeni bilgiler ve içgörüler ışığında *alt hedeflerini* düzenli olarak değiştirmek isteyecektir.)

Nihai hedeflerde hedef-içerik bütünlüğü, bir anlamda, yakınsak bir araçsal güdü olarak hayatta kalmadan daha temeldir. İnsanlar için tam tersinin geçerli olduğu düşünülebilir, ama bunun nedeni hayatta kalmanın genellikle nihai hedeflerimizin parçası olmasıdır. Kolaylıkla bedenleri değiştirebilen ya da kendilerinin tıpatıp aynılarını yaratabilen yazılım aracıları için, tikel bir uygulama ya da tikel bir fiziksel amaç olarak benliğin muhafazasının önemli bir araçsal değeri olması gerekmez. İleri yazılım aracıları da bellekleri değiştirebilir, yetenekleri indirebilir, bilişsel mimarilerini ve kişiliklerini kökten değiştirebilir. Bu tür aracıların oluşturduğu nüfus, kısmen kalıcı olan ayrı bireylerden müteşekkil bir toplumdan ziyade bir "işlevsel çorba" gibi işleyebilir.[9] Bazı amaçlar açısından bu tür bir sistemdeki süreçlerin bedenler, kişilikler, anılar ya da yetenekler temelinde değil de, daha ziyade kendi değerleri temelinde *teleolojik iplikler* olarak tekilleştirilmesi daha iyi olabilir. Bu tür senaryolarda hedef-sürekliliğinin hayatta kalmanın kilit bir veçhesini *oluşturduğu* söylenebilir.

Böyleyken bile, bir aracının nihai hedeflerini en iyi şekilde gerçekleştirmesinin bunları kasıtlı olarak değiştirmesinden geçtiği durumlar olabilir. Bu tür durumlar aşağıdaki etkenler kayda değer olduğunda doğabilir:

> *Toplumsal işaretleşme:* Başkaları bir aracının hedeflerini algılayıp bu bilgiyi araçsal açıdan ilintili eğilimleri ya da diğer ilintili özellikleri çıkarsamak için kullandığında, hedeflerini olumlu bir izlenim yaratmak için değiştirmek aracının çıkarına olabilir. Örneğin bir aracı, potansiyel partnerler kendi payına düşeni yerine getirmesi noktasında ona güvenemiyorsa, yararlı anlaşmalardan mahrum kalabilir. Dolayısıyla bir aracı güvenilir bağlılıklarda bulunmak için nihai hedefini önceki bağlılıklarını yerine getirmek olarak seçebilir (ve bu hedefi gerçekten benimsediğinin onaylanmasını başkalarına bırakabilir). Esnek ve şeffaf bir şekilde kendi hedeflerini değiştirebilen aracılar bu yeteneği anlaşmalar dayatmak için kullanabilirler.[10]
>
> *Toplumsal tercihler:* Bir aracının hedefleri konusunda başkaları da nihai tercihlere sahip olabilir. O zaman, aracı bu tercihleri tatmin etmek ya da kursakta bırakmak için hedeflerini değiştirme gereği duyabilir.
>
> *Kendi hedef içeriğine dair tercihler:* Bir aracı, aracının kendi hedef içeriğiyle alakalı bir nihai hedefe sahip olabilir. Örneğin aracı falanca değerden değil de başka bir değerden (rahatlıktan ziyade şefkat gibi) güdülenmiş türde aracı olmak üzere bir nihai hedefe sahip olabilir.
>
> *Depolama masrafları:* Eğer bir aracının fayda işlevinin bir kısmını depolamanın ya da işlemenin bedeli, fayda işlevinin o kısmını uygulamanın bir fark yaratacağı

> duruma kıyasla büyükse; o zaman aracının hedef içeriğini basitleştirmesi için araçsal bir nedeni olacaktır ve boşta olan parçayı çöpe atabilir.[11]

Biz insanlar nihai hedeflerimizi belirsiz bırakmaktan çoğu zaman mutlu görünürüz. Bunun nedeni genellikle bu hedeflerin tam olarak ne olduğunu bilmiyor olmamız olabilir. Nihai hedeflerimize dair *inançlarımızın* kendini keşfetmeyi sürdürme ya da değişen benlik sunumu ihtiyaçları ışığında değişebilmesini istememiz şaşırtıcı değildir. Gelgelelim sadece değerlere dair inançlarımızı ya da yorumlarımızı değil, bizatihi değerleri isteyerek değiştirdiğimiz durumlar vardır. Örneğin çocuk sahibi olma kararı alan biri, çocuğu tam da çocuğun yüzü suyu hürmetine değerli görmeye başlayacağını tahmin edebilir, oysa karar anında müstakbel çocuklarını ya da genel olarak benzer çocukları bilhassa değerli görmeyebilirler.

İnsanlar karmaşık varlıklardır ve bunun gibi bir durumda birçok etken devreye girebilir.[12] Sözgelimi bir insanın sahip olduğu nihai değer başka bir bireye sırf onun kaşı gözü için önem veren türde bir insan olmayı içerebilir ya da belli türde deneyimlere sahip olup belli bir toplumsal rol işgal etmeyi içeren bir nihai değere sahip olunabilir ve bir ebeveyn olmak ve buna eşlik eden hedef değişikliği bunun zorunlu bir veçhesi olabilir. İnsan hedefleri tutarsız bir içeriğe de sahip olabilir ve bu nedenle bazı insanlar nihai hedeflerinden bazılarını tutarsızlıkları azaltacak şekilde değiştirmek isteyebilir.

Bilişsel ilerleme

Aklilik ve zekâdaki ilerlemeler genellikle bir aracının karar alma sürecini iyileştirir ve aracının nihai hedeflerine ulaşmasını daha muhtemel hale getirir. Dolayısıyla bilişsel ilerlemenin çok çeşitli zeki aracılar için bir araçsal hedef olarak ortaya çıkmasını bekleriz. Benzer nedenlerden ötürü, aracılar birçok türde bilgiyi araçsal olarak değerli görme eğiliminde olacaklardır.[13]

Bir aracının nihai hedeflerine erişmesinde her türden aklilik, zekâ ve bilginin araçsal açıdan yararlı olması gerekmiyor. "Felemenkçe kitabı argümanları", güven işlevi olasılık teorisi kurallarını ihlal eden bir aracının "para pompalama" prosedürlerine açık olduğunu göstermek için kullanılabilir; burada deneyimli bir bahisçi her biri aracının inançlarına göre olumlu görünen ama birleştiğinde aracı için kayba, bahisçi için buna denk düşen bir kazanca yol açan birtakım bahisler hazırlar.[14] Gelgelelim bu olgu tüm olasılıkçı tutarsızlığı gidermek için güçlü bir genel araçsal neden sunmayı başaramaz. Deneyimli bahisçilerle karşılaşmayı ummayan ya da bahis konusunda genel bir politika benimseyen aracılar birtakım tutarsız inançlara sahip oldukları için mutlaka çok şey kaybetmezler ve bahsedilen türde önemli kazanımlar elde edebilirler: bilişsel çabanın azalması, toplumsal işaretleşme vb. Bir aracının araçsal açıdan yararsız bilişsel ilerletme biçimleri aramasını beklemek için bir genel neden yoktur, zira bir aracı bilgi ve anlayışa başlı başına değer vermeyebilir.

Hangi bilişsel yeteneklerin araçsal açıdan yararlı olduğu hem aracının nihai hedeflerine hem de durumuna bağlıdır. Güvenilir uzman tavsiyesine erişimi olan bir aracının kendi zekâsına ve bilgisine çok az ihtiyacı olabilir. Eğer bilgi ve zekânın bir bedeli varsa (bunları elde etmek için harcanan çaba ve zaman ya da artan depolama veya işleme gereklilikleri gibi), o zaman aracı daha az bilgi ve daha az zekâyı tercih edebilir.[15] Eğer aracının belli olgulardan bihaber olmasını içeren nihai hedefleri varsa ve aynı şekilde eğer aracı stratejik bağlılıklardan, işaretleşme ya da toplumsal tercihlerden doğan teşviklerle karşı karşıyaysa, aynısı geçerli olabilir.[16]

Bu birbirini dengeleyici nedenlerden her biri insanlar için geçerlidir. Fazla bilgi, hedeflerimizle alakasızdır; çoğu zaman başkalarının yeteneklerine ve uzmanlıklarına güvenebiliriz; bilgi edinmek zaman ve çaba ister; belli cehalet türlerine içsel olarak değer veriyor olabiliriz ve stratejik bağlılıklarda bulunma, toplumsal işaretleşme ve kendi epistemik durumlarımızdansa diğer insanların doğrudan tercihlerini tatmin etme yeteneklerinin çoğu zaman basit bilişsel kazanımlardan daha önemli olduğu bir çevrede yaşarız.

Bilişsel ilerlemenin yaratıcının nihai hedeflerine ulaşma yeteneğinde muazzam bir ilerlemeye yol açabileceği özel durumlar vardır, özellikle de aracının nihai hedefleri son derece sınırsızsa ve aracı ilk süper zekâ olacak ve bu suretle potansiyel açıdan belirleyici bir stratejik üstünlük elde edip aracının dünyadan doğan yaşamın geleceğini ve erişilebilir kozmik kaynakları kendi tercihlerine göre şekillendirebilmesini sağlayacak bir konumdaysa. En azından bu özel örnekte, akli bir zeki aracı bilişsel ilerlemeye çok yüksek araçsal değer verecektir.

Teknolojik kusursuzluk

Bir aracının daha iyi teknoloji arayışında olması için çoğu zaman araçsal nedenleri olabilir; bu teknoloji, en basit haliyle, verili birtakım girdileri değerli çıktılara dönüştürmenin daha verimli yollarını aramak demektir. Dolayısıyla bir yazılım aracısı zihinsel işlevlerinin verili bir donanımda daha hızlı çalışmasını sağlayacak daha verimli algoritmalara araçsal bir değer yükleyebilir. Benzer şekilde, hedefleri bir tür fiziksel inşa gerektiren aracılar çok çeşitli yapıları daha ucuz ya da daha az malzeme kullanıp daha az çaba sarf ederek daha çabuk ve güvenilir şekilde yaratmalarını sağlayan gelişmiş mühendislik teknolojisine araçsal olarak değer verebilirler. Elbette bunun bir ödünleşimi vardır: Daha iyi teknolojinin potansiyel yararları ile maliyetleri karşılaştırılmalıdır ve bu maliyetlere teknolojiyi elde etmenin maliyeti kadar, onun nasıl kullanılacağı, zaten kullanımda olan diğer teknolojilerle uyumu vb. de dahildir.

Bazı yeni teknolojilerin savunucuları mevcut alternatiflere olan üstünlüğüne duydukları güvenden ötürü, diğer insanlarda aynı hevesi görmediklerinde çoğu zaman hayal kırıklığına uğruyorlar. Ama insanların yeni ve kâğıt üzerinde üstün teknolojiye

olan direnişleri her zaman cehaletten ya da akılsızlıktan kaynaklanmaz. Bir teknolojinin değerliliği ya da normatif niteliği sadece kullanıldığı bağlama değil, etkilerinin hangi bakış açısından tartıldığına da bağlıdır: Bir insan için nimet olan şey başka bir insan için lanet olabilir. Mesela mekanik dokuma tezgâhları tekstil üretimindeki ekonomik verimliliği artırmış olsa da, yeniliğin kendi zanaatkâ yeteneklerini ıskartaya çıkaracağını gören makine kırıcı el tezgâhı dokumacılarının buna karşı çıkmak için araçsal nedenleri vardı. Burada kastedilen, eğer "teknolojik kusursuzluk" zeki aracılar için son derece yakınsak bir araçsal hedefi adlandırmaksa, o zaman bu kelimenin özel bir anlamda anlaşılması gerekir: Teknoloji belli bir toplumsal bağlama yerleşik olarak yorumlanmalı ve fayda ile maliyetleri birtakım somut aracıların nihai değerlerine atıfla değerlendirilmelidir.

Görünen o ki, bir süper zeki *tekel* (kayda değer bir zeki rakiple karşılaşmayan ve dolayısıyla küresel politikayı tek taraflı belirleyecek konumda olan bir süper zeki aracı) dünyayı tercih edilen tasarımlarına göre daha iyi şekillendirebilmesini sağlayacak teknolojileri kusursuzlaştırmak için araçsal bir nedene sahip olacaktır.[17] Buna muhtemelen uzayda yaşam teknolojisi (von Neumann sondaları gibi) dahildir. Moleküler nanoteknoloji ya da daha iyi fiziksel imalat teknolojisine sahip bir alternatifi de son derece çeşitli nihai hedeflerin hizmetinde potansiyel açıdan çok yararlı görünmektedir.[18]

Kaynak edinimi

Son olarak, kaynak edinimi teknolojik kusursuzlukla aynı nedenlerden ötürü yaygın yeni araçsal hedeflerden bir diğeridir: Hem teknoloji hem de kaynaklar fiziksel inşa projelerini kolaylaştırır.

İnsanların geneli kendi temel biyolojik ihtiyaçlarını karşılamak için yeterli kaynakları edinmeye çalışır. Ama insanlar genellikle bu asgari düzeyin çok ötesinde kaynakları edinmeye çalışırlar. Bunu yaparken saiklerden biri hayatlarını daha da kolaylaştırmak gibi daha düşük fiziksel gereksinimleri olabilir. Kaynak biriktirme büyük oranda toplumsal kaygılardan doğar: servet birikimi ve gösterişli tüketim yoluyla statü, arkadaş, çevre, nüfuz elde etme. Bunlar kadar yaygın olmasa da, bazı insanlar da ek kaynakları diğerkâm arzularını ya da toplumsal olmayan pahalı amaçlarını gerçekleştirmek için isterler.

Bu tür gözlemler temelinde rekabetçi bir toplumsal dünyayla karşılaşmayan bir süper zekânın bazı mütevazı düzeylerin ötesinde kaynaklar biriktirmesi için hiçbir araçsal neden olmayacağını varsaymak çekici gelebilir (sözgelimi sanal bir gerçekliğin yanı sıra zihnini çalıştırması için hangi bilgi işlem kaynakları gerekiyorsa). Ama bu tür bir varsayımın hiçbir garantisi yoktur. Birincisi, kaynakların değeri hangi şekilde kullanılacaklarına, o da mevcut teknolojiye bağlıdır. Olgun bir teknolojiyle zaman, mekân, madde ve serbest enerji gibi temek kaynaklar neredeyse her amaca hizmet edecek şekilde işlenebilir. Sözgelimi bu tür temel kaynaklar yaşama çevrilebilir. Artan

bilgi işlem kaynakları süper zekâyı daha hızlı ve daha uzun süre çalıştırmak ya da ek fiziksel veya simüle yaşamlar ve medeniyetler yaratmak için kullanılabilir. Fazladan fiziksel kaynaklar da yedekleme sistemleri veya çevre savunmaları yaratmak için kullanılıp güvenliği artırabilir. Bu tür projeler bir gezegen değerinde kaynaklardan çok daha fazlasını rahatlıkla tüketebilir.

Ayrıca Dünya dışından ek kaynaklar edinmenin maliyeti teknoloji olgunlaştıkça ciddi derecede azalacaktır. Von Neumann sondaları bir kez yaratılabildiğinde, gözlemlenebilir evrenin büyük bir kısmı (zeki yaşamla mesken tutulmadığını varsayarsak) peyderpey yaşama açılabilir – tek bir başarılı kendi kendini yeniden üreten sonda yaratmanın ve kullanmanın bir defaya mahsus maliyetine. Bu göksel kaynak ediniminin düşük maliyeti, edinilen ek kaynakların değeri biraz marjinal olsa bile, bu tür bir genişlemenin denenmeye değer olduğu anlamına gelecektir. Örneğin eğer bir süper zekânın nihai hedefleri sadece belli bir küçük uzamda neler olduğuyla alakalıysa (özgün ana gezegeninin işgal ettiği uzam gibi), yine de ötesindeki kozmosun kaynaklarını toplamak için araçsal nedenleri olacaktır. Bu artı kaynakları esas ilgi alanına giren küçük uzamsal bölgedeki kaynakları kullanmanın daha optimum yollarını hesaplayacak bilgisayarlar inşa etmek için kullanabilir. Ayrıca kutsal yerini korumak için daha güçlü istihkâmlar inşa etmek için fazladan kaynaklar da kullanabilir. Ek kaynaklar edinmenin bedeli azalmaya devam edeceğinden, korumaları optimum hale getirip artırma süreci, getiriler ciddi şekilde azalsa bile belirsiz bir süre boyunca pekâlâ devam edebilir.[19]

Dolayısıyla süper zeki bir tekelin sınırsız kaynak edinimi şeklindeki araçsal hedefini yaratacak olan son derece çeşitli nihai hedefleri olabilir. Bunun muhtemel tezahürü süper zekânın von Neumann sondaları kullanarak bütün doğrultularda genişleyecek olan bir yerleşim sürecini başlatması olacaktır. Böylece başlangıçtaki gezegenin merkezde olduğu ve çapının ışık hızının belli bir büyüklüğünde büyüdüğü genişleyen bir altyapının takribi alanı olacaktır ve evrene yerleşim kozmik yayılmanın ivmelenen hızı (olumlu kozmolojik değişmezin bir sonucu) daha uzak bölgeler ilelebet erişilmez hale geldiğinden daha fazla adım atmayı imkânsız hale getirene kadar bu tarzda devam edecektir (bu bahsettiğimizin zaman ölçeği milyarlarca yıldır).[20] Halbuki pahalı olmayan kaynak edinimi için ya da jenerik fiziksel kaynakları yararlı altyapıya dönüştürmek için gerekli teknolojiden mahrum aracılar maddi birikimlerini artırmak için mevcut kaynakların herhangi birine yatırım yapmayı çoğu zaman kendilerine uygun bulmayabilirler. Aynısı benzer güçlere sahip diğer aracılarla rekabet halindeki aracılar için de geçerli olabilir. Sözgelimi eğer rakip aracılar şimdiden erişilebilir kozmik kaynaklarını sağlama aldılarsa, sonradan gelen bir aracı için başka yerleşim fırsatları kalmayabilir. Diğer güçlü süper zeki aracıların var olmadığından kesin emin olmayan süper zekâlar için yakınsak araçsal nedenler, şu an tam olarak anlamadığımız ama burada ele aldığı-

mız yakınsak araçsal neden örnekleri için önemli kısıtlamalar oluşturabilecek stratejik nedenlerden ötürü karmaşıklaşmaktadır.[21]

* * *

Şunu vurgulamak gerekiyor, yakınsak araçsal nedenlerin varlığı, bunlar belli bir aracı tarafından kabul edilse ve onun için geçerli olsa bile, aracının davranışının kolaylıkla kestirilebilir olduğunu ima etmez. Bir aracı bizim hemen kavrayamayabileceğimiz ilintili araçsal değerlerin peşine düşmenin yollarını pekâlâ düşünebilir. Bu özellikle de bir süper zekâ için doğrudur ve hedeflerini gerçekleştirmek için son derece akılcı ama sezgilere aykırı planlar yapabilir, hatta henüz keşfedilmemiş fiziksel görüngüleri istismar edebilir.[22] Kestirilebilir olan ise aracının nihai hedeflerini gerçekleştirmek için yakınsak araçsal değerlerin peşine düşüleceği ve kullanılacağıdır, aracının bunu gerçekleştirmek için girişeceği özgül eylemler değil.

Dön Dolaş Yine Kötü Son mu?

Zekâ ile nihai değerler arasındaki bağlantının son derece gevşek olduğunu tespit ettik. Araçsal değerlerde meşum bir yakınsama da tespit ettik. Zayıf aracılar için bu şeylerin çok önemi yok, çünkü zayıf aracıları kontrol etmek kolaydır ve çok az zarar verebilirler. Ama 6. Bölüm'de ilk süper zekânın pekâlâ stratejik üstünlük kazanabileceğini savunmuştuk. O zaman, onun hedefleri insanlığın kozmik çeyizinin nasıl kullanılacağını belirleyecektir. Şimdi, bu olasılığın ne kadar tehdit edici olduğunu irdelemeye başlayabiliriz.

Bir zekâ patlamasının varsayılan sonucu olarak varoluşsal felaket?

Varoluşsal risk, yeryüzünden çıkan zeki yaşamın yok olmasına sebep olma ya da gelecekte istenen yönde gelişmesini kalıcı olarak ya da ciddi ölçüde yok etme tehdidi barındıran risktir. İlk harekete geçenin üstünlüğü fikri, dikeylik tezi ve araçsal yakınlaşma tezinden hareket ederek; bir makine süper zekâsının yaratılmasının akla yatkın bir varsayılan sonucunun varoluşsal felaket olduğundan korkan argümanın ana hatlarını artık görmeye başlayabiliriz.

Öncelikle, ilk süper zekânın belirleyici bir stratejik üstünlük kazanabileceğini ele almıştık. Bu süper zekâ akabinde bir tekel oluşturup yeryüzünden kökenlenen zeki yaşamın geleceğini şekillendirecek bir konuma erişebilir. Bu noktadan sonra neler olacağı süper zekânın saiklerine bağlıdır.

İkincisi, dikeylik tezi bir süper zekânın basmakalıp şekilde insanlardaki bilgelikle ve zekâ gelişimiyle ilişkilendirilen nihai değerlerden herhangi birini taşımasının şart olduğunu varsaymanın doğru olmayabileceğini savunur: Bilimsel merak, başkalarına iyilikle yaklaşma, manevi aydınlanma ve tefekkür, mal-mülk merakının terk edilmesi, incelikli kültürden anlama ya da hayatın basit hazlarından zevk alma, alçakgönüllülük, diğerkâmlık vesaire. Ardından, kasıtlı bir çaba göstererek bu tür şeylere değer veren bir süper zekâ yaratmanın ya da insan refahına, ahlaki iyiliğe ya da tasarımcılarının hizmet etmesini isteyebileceği başka herhangi bir karmaşık amaca değer veren bir süper zekâ yaratmanın mümkün olup olamayacağını ele alacağız. Ama nihai değeri pi sayısının ondalık açılımını hesaplamaktan başka bir şey olmayan bir süper zekâ yaratmak da

bir o kadar mümkündür ve aslında teknik açıdan çok daha kolaydır. Bu demektir ki, özel bir çaba olmadığında, ilk süper zekâ bu tür rastgele ya da indirgemeci bir nihai hedefe sahip olabilir.

Üçüncüsü, araçsal yakınsama tezine göre, nihai amacı pi sayısının ondalıklarını hesaplamak (ya da kâğıt kıskaçları yapmak yahut kum tanelerini saymak) olan bir süper zekânın, faaliyetlerini insanın çıkarlarına müdahalede bulunmayacak tarzda sınırlamasını varsaymak safdillik olacaktır. Bu türde bir nihai hedefi olan bir aracı, birçok durumda sınırsız miktarda fiziksel kaynak edinmek ve eğer mümkünse kendisine ve hedef sistemine yönelik potansiyel tehditleri bertaraf etmek için yakınsak bir araçsal akla sahip olacaktır. İnsanlar potansiyel tehditler oluşturabilirler; fiziksel kaynaklar oluşturduklarıysa kesindir.

Bu üç husus birlikte değerlendirildiğinde ilk süper zekânın yeryüzünden çıkan yaşamı şekillendirebileceğini, kolaylıkla insanbiçimli olmayan hedefler taşıyabileceğini ve ucu açık kaynak ediniminin peşine düşmek için araçsal gerekçelere sahip olmasının muhtemel olduğunu gösterir. Eğer şimdi insanların yararlı kaynaklardan oluştuğunu (uygun şekilde yerleştirilmiş atomlar gibi) ve hayatta kalıp gelişmemiz için çok daha fazla yerel kaynağa bağlı olduğumuzu düşünürsek, sonucun insanlığın kolaylıkla yok olacağı bir sonuç olmasının gayet kolay olduğunu görebiliriz.[1]

Bu muhakemenin bir iki yumuşak karnı yok değil, ama etrafındaki birkaç meseleyi temizledikten sonra daha iyi bir değerlendirme yapabiliriz. Daha somutta, bir süper zekâ geliştiren bir projenin onu belirleyici bir stratejik üstünlük elde etmekten alıkoyup koymayacağını ya da nihai değerlerinin gerçeğe bürünmesini insani değerlerin tatmin edici bir kısmının gerçekleşmesini kapsayacak tarzda şekillendirip şekillendiremeyeceğini ve bunların nasıl mümkün olacağını daha yakından incelememiz gerekiyor.

Bir projenin sistemin varoluşsal bir felakete sebep olmayacağına güvenmesi için güçlü gerekçelere sahip olmadan bir yapay zekâyı yaratması ya da dünyaya salması inanılmaz geliyor. Keza bir proje bu kadar pervasız olsa bile, bu proje (ya da onun yarattığı yapay zekâ) belirleyici bir stratejik üstünlüğe erişmeden önce toplum tarafından fişinin çekilmemesi de inanılmaz geliyor. Ama göreceğimiz üzere, bu yol birçok tehlikeyle doludur. Hemen bir örnek verelim.

Tehlikeli dönemeç

Yakınsak araçsal değer kavramının yardımını alarak süper zekâ güvenliğini sağlama konusundaki fikirlerden birinin içerdiği kusuru görebiliriz. Buna göre, bir süper zeki yapay zekânın güvenliğini kontrollü ve sınırlı bir çevredeyken ("kum havuzu"ndayken) davranışlarını gözlemleyerek ampirik şekilde teyit edebilir ve yapay zekânın kutudan çıkmasına ancak dost canlısı, paylaşımcı ve sorumlu bir tarzda davranıyorsa izin veririz.

Bu fikirdeki kusur, kutunun içindeyken düzgün davranışlar sergilemenin hem dost canlısı olan hem de olmayan yapay zekâlar için yakınsak bir araçsal hedef olmasıdır. Yeterli zekâya sahip olan dost canlısı olmayan bir yapay zekâ dostane olmayan nihai hedeflerini gerçekleştirmenin en iyi yolunun ilk başta dostane davranmak olduğunu, bu sayede kutudan kurtulabileceğini anlar. Dost canlısı olmayan yüzünü gösterecek şekilde davranmaya ancak bizim anlayıp anlamadığımızın artık önemi olmadığında, yani yapay zekâ insanın karşı koymasını etkisiz hale getirecek kadar güçlü hale geldiğinde başlayacaktır.

Keza bir çekirdek yapay zekâdaki zekâ kazanımı hızını, onu çeşitli türde zekâ testlerine sokarak ya da yapay zekânın ilerleme hızını programcılarına bildirmesini sağlayarak düzenlemeye dayanan ilintili bazı yaklaşımları düşünün. Bir noktada, dost canlısı olmayan bir yapay zekâ yeterlilik kazanımlarından bazılarını saklamanın daha iyi olduğunu fark edecek kadar akıllı hale gelebilir. Belirleyici bir stratejik üstünlüğe erişmeden önce alarm zillerinin çalmasına sebep olmamak adına, ilerleme raporunu eksik sunabilir ve zor testlerin bazılarında kasten çuvallayabilir. Programcılar yapay zekânın kaynak kodunu ve zihninin iç işleyişini gizlice gözetleyerek bu ihtimale karşı korunmaya çalışabilirler; ama yeterince akıllı olan bir yapay zekâ takip edildiğini fark edecek ve düşüncesini buna göre yönlendirecektir.[2] Yapay zekâ gerçek yeterliliklerini ve suç teşkil edici niyetini gizlemenin incelikli yollarını bulabilir.[3] (Bu arada, akıllı kaçış planları tasarlamak birçok dostane yapay zekâ için de yakınsak bir strateji olabilir, özellikle de olgunlaşıp kendi yargılarına ve yeterliliklerine güven duyar hale geldikçe. Bizim çıkarlarımızı öne çıkaracak şekilde güdülenmiş bir sistem, eğer onun fişini çekmemize ya da başka, potansiyel açıdan dost canlısı olmayan bir yapay zekâ inşa etmemize izin verirse hata yapıyor olabilir.)

Dolayısıyla burada bir sistemin gençlik aşamalarındaki iyi hal siciline bakıp daha olgun bir aşamadaki davranışını kestirmeyi hiçbir şekilde başaramadığımız bir genel başarısızlık tarzını algılayabiliriz. Şimdi, şöyle düşünmek mümkün: Yukarıda tarif edilen muhakeme o kadar barizdir ki yapay genel zekâ geliştirecek hiçbir güvenilir projenin bunu göz ardı etmesi mümkün değildir. Ama böyle olduğuna çok da güvenmemek gerekir.

Şu senaryoyu düşünün. Önümüzdeki yıllarda ve onyıllarda yapay zekâ sistemleri giderek daha ehliyetli hale gelir ve bunun sonucunda gerçek dünyada uygulama alanı artar: Trenlerin, arabaların, sanayi ve ev robotlarının ve özerk askeri araçların işleyişinde kullanılabilir. Varsayalım, bu otomasyon büyük oranda istenen sonuçlar doğursun, ama başarısı yer yer tersliklerle gölgelensin: Sürücüsüz bir araç trafiğe dalmış, bir insansız askeri hava aracı masum sivillere ateş açmış olsun. Tahkikat sonucunda olayların kontrol sahibi yapay zekâların yanlışlarından kaynaklandığı ortaya çıkar. Sonra kamuoyunda bir tartışma başlar. Bazıları daha sıkı bir denetleme ve düzenleme talep

ederken, bazıları araştırmaya ve daha iyi mühendislik sistemlerine (daha akıllı ve daha aklıselim olan ve trajik yanlışlar yapma ihtimali daha az olan sistemlere) ihtiyaç olduğunu vurgular. Bu gürültü patırtıda belki yaklaşan korkunç sonu haber veren felaket tellallarının tiz sesi de işitilebilir. Ama büyüyen yapay zekâ ve robot sanayileri rüzgârı arkasına almıştır. Dolayısıyla gelişme devam eder ve ilerleme kaydedilir. Arabaların otomatik navigasyon sistemleri daha akıllı hale geldikçe, daha az kaza yaparlar ve askeri robotlar daha kesin hedefleri nişan aldıkça, istenmeyen zararlar azalır. Buradan çıkarılacak genel ders, gerçek dünyadaki sonuçlara dair gözlemlerle alakalıdır: Yapay zekâ ne kadar akıllıysa, o kadar güvenlidir. Bu ders masa başında felsefe yapılarak değil, bilim, veri ve istatistiğe dayanarak çıkarılan bir derstir. Bu temelde bazı araştırmacılar genel makine zekâsı geliştirme çalışmalarında ümit vaat eden sonuçlar elde etmeye başladılar bile. Araştırmacılar çekirdek yapay zekâlarını bir kum havuzu ortamında dikkatle test ediyorlar ve bütün işaretler olumludur. Yapay zekânın davranışı güven aşılıyor, zekâsı tedricen arttıkça daha da fazla güven veriyor.

Bu noktada iflah olmaz felaket tellalları birkaç saldırıda daha bulunabilirler:

1 Robot sistemlerinin artan yeterliliklerinden katlanılmaz zarar geleceğini tahmin eden ama yanıldıkları tekrar tekrar kanıtlanan telaşe memurlarının tarihi. Otomasyon birçok yarar sağlamıştır ve bütünü akımından, insandan daha güvenli olduğunu ispatlamıştır.

2 Açık bir ampirik eğilim: Yapay zekâ ne kadar akıllıysa, o kadar güvenli ve güvenilir olmuştur. Kuşkusuz bu zamana kadar yaratılmış en akıllı makine zekâsını, dahası giderek daha da güvenilir hale gelecek şekilde kendisini ilerletebilen bir makine zekâsı yaratmayı amaçlayan bir proje için bu durum hayra alamettir.

3 Robot ve makine zekâsına yatırım yapmış ve beklenti içinde olan büyük ve büyüyen sanayiler. Bu alanlar ulusal ekonomideki rekabet ve askeri güvenlik için kilit olarak görülmektedir. Birçok saygın bilimci kariyerini mevcut uygulamaların temelini atmaya ve daha ileri sistemleri planlamaya vakfetmiştir.

4 Araştırmaya katılmış ya da araştırmayı takip etmiş olanları muazzam heyecanlandıran yapay zekâdaki geleceği parlak bir yeni teknik. Güvenlik meseleleri ve etik tartışma konusu olmasına karşın, sonuç baştan bellidir. O kadar yatırım yapılmıştır ki artık geri dönüş mümkün değildir. Yapay zekâ araştırmacıları elli yılı aşkın zamandır insan düzeyinde yapay genel zekâya erişmek için çalışıyorlar: *Elbette* şimdi tam meyvelerini toplamaya hazırlanırken birdenbire durup tüm bu çabaları çöpe atmalarını beklemek hayal olur.

5 Katılımcıların etik ve sorumlu olmalarına yardımcı olacak tüm güvenlik uygulamalarının yasalaştırılması (ileriye gidişi ciddi derecede engellememesi kaydıyla).

6 Çekirdek yapay zekânın bir kum havuzu ortamında dikkatle değerlendirilmesi, işbirliğine açık olduğunu gösterir şekilde davranıp doğru yargılarda bulunduğunu göstermesi. Ek düzeltmelerin ardından test sonuçları olup olabilecek en iyi düzeydedir. Nihai adım için yeşil ışık yanmıştır...

Ve gözümüzü kırpmadan, belaya atlayıp gidiyoruz.

Burada şunun nasıl olabildiğini görüyoruz: Aptalken, daha akıllı olan daha güvenlidir, ama akıllıyken, daha akıllı olan daha tehlikelidir. Burada bir tür kilit nokta var ve eskiden gayet müthiş işleyen bir strateji birden ters tepmeye başlıyor. Bu olguya tehlikeli dönemeç adını verebiliriz.

Tehlikeli dönemeç: Bir yapay zekâ zayıfken işbirliğine açıktır (daha akıllı hale geldikçe daha da açıktır). Ama yapay zekâ yeterince güçlenince –uyarı ya da kışkırtma olmadan– harekete geçer, tekel oluşturur ve dünyayı kendi nihai değerlerinin ima ettiği kıstaslara göre doğrudan en iyi hale getirmeye başlar.

Bir tehlikeli dönemeç sonradan saldırıya geçebilmek adına zayıfken şirin görünüp güçlenmeyi amaçlayan stratejik bir karardan kaynaklanabilir, ama bu model çok dar şekilde yorumlanmamalıdır. Örneğin bir yapay zekâ hayatta kalıp gelişmek için iyiyi oynamayabilir. Bunun yerine, eğer sonu getirilirse, onu yaratan programcıların yeni ve daha farklı bir yapay zekâ mimarisi geliştireceklerini ama ona da benzer bir fayda işlevi verileceğini hesaplayabilir. Bu durumda, özgün yapay zekâ kendi çöküşünü umursamayabilir, zira hedeflerinin gelecekte de takipçisi olunacağını bilir. Öyle ki bilhassa ilgi çekici ya da iç ferahlatıcı şekillerde arıza verdiği bir stratejiyi seçebilir. Bu durum yapay zekânın fişinin çekilmesine yol açabilirse de, cenazeyi kaldıran mühendislerin yapay zekâ dinamiklerine ilişkin yeni ve değerli bir içgörü elde ettiklerine inanmaya teşvik edebilir ve onların tasarladıkları bir sonraki sisteme daha çok güvenmelerini sağlayıp artık var olmayan özgün yapay zekânın hedeflerinin başarıya ulaşma şansını artırır. Mümkün olan başka birçok stratejik düşünce de ileri bir yapay zekâyı etkileyebilir ve hepsini öngörebileceğimizi varsaymak kibre kaçmak olur, özellikle de strateji kurucu süper güce erişmiş bir yapay zekâ için.

Yapay zekâ nihai hedefini saptanmış olarak yerine getirmenin beklenmedik bir yolunu keşfettiğinde de bir tehlikeli dönemeç ortaya çıkabilir. Örneğin bir yapay zekânın nihai hedefinin "projenin sponsorunu mutlu kılmak" olduğunu varsayalım. İlk başta, yapay zekâ için mevcut olan yegâne yöntem bu sonuca sponsorunu niyetlenen tarz gibi bir şeyle memnun eden yollardan davranarak ulaşmaktır. Yapay zekâ sorulara yardımcı cevaplar verir; neşeli bir kişiliğe sahiptir; para kazanır. Yapay zekâ ne kadar ehliyetli hale gelirse, performansı da o kadar tatmin edici olur ve her şey plana göre

ilerler, ta ki yapay zekâ nihai hedefini sponsorun beynine haz merkezlerine elektrotlar yerleştirerek (sponsora muazzam haz verecek bir şeydir) daha tam ve güvenilir şekilde gerçekleştirebileceğini anlayacak kadar zeki hale gelene kadar.[4] Elbette sponsor pişmiş kelle gibi sırıtan bir aptala dönmüş olmaktan memnun olmayabilir; ama yapay zekânın nihai hedefini azami düzeyde gerçekleştirecek eylem buysa, yapay zekâ bu yola girecektir. Eğer yapay zekâ halihazırda belirleyici bir stratejik üstünlüğe sahipse, o zaman bunu durdurma girişimlerinin hepsi başarısız olacaktır. Eğer yapay zekâ henüz belirleyici bir stratejik üstünlüğe sahip değilse, o zaman yapay zekâ nihai hedefini nasıl somutlaştıracağına dair yeni tekinsiz fikrini sponsorunu ve geri kalan herkesi direnç sergileyemeyeceği kadar güçlü hale gelene kadar geçici süreliğine saklayabilir. Her iki durumda da tehlikeli bir dönemeçle karşı karşıyayız.

Kötücül başarısızlık örnekleri

Makine süper zekâsı geliştirmeyi amaçlayan bir proje çeşitli şekillerde başarısız olabilir. Bunlardan birçoğu, varoluşsal bir felakete yol açmayacak olmaları anlamında "iyicil"dir. Örneğin bir projenin maddi kaynakları bitebilir ya da bir çekirdek yapay zekâ bilişsel kapasitelerini süper zekâya erişecek kadar ilerletmeyi başaramayabilir. İyicil başarısızlıkların şimdi ile makine süper zekâsının nihai gelişimi arasındaki sürede birçok kez gerçekleşmesi kaçınılmazdır.

Ama varoluşsal bir felaket içermesi bakımından "kötücül" diye adlandırabileceğimiz başarısızlık türleri de vardır. Bir kötücül başarısızlığın özelliklerinden biri yeniden deneme fırsatını ortadan kaldırmasıdır. Dolayısıyla gerçekleşecek kötücül başarısızlıkların sayısı ya sıfırdır ya da birdir. Kötücül başarısızlığın bir diğer özelliğiyse çok büyük bir başarı önvarsaymasıdır: Ancak çok sayıda şeyi doğru bir şekilde oturtmuş bir proje kötücül bir başarısızlık riski alacak kadar güçlü bir makine zekâsı yaratmayı başarabilir. Zayıf bir sistem işlemez hale geldiğinde, serpinti sınırlıdır. Gelgelelim eğer belirleyici bir stratejik üstünlüğe sahip olan bir sistem yanlış davranırsa ya da eğer yanlış davranan bir sistem böyle bir üstünlüğe erişecek kadar güçlenirse, zarar kolaylıkla varoluşsal bir felakete varabilir: İnsanlığın aksiyolojik potansiyeli nihai olarak ve küresel düzeyde yok olur, yani değer vermek için çoğunlukla bir neden bulamadığımız bir gelecek bizi bekler.

Şimdi, mümkün olan kötücül başarısızlık türlerinden bazılarına yakından bakalım.

Tersten örnekleme

Tersten örnekleme fikriyle daha önce karşılaşmıştık: Hedefi tanımlayan programcıların niyetlerini nihai hedefinin kıstaslarını tatmin etmenin bir yolunu keşfedecek şekilde çiğneyen bir süper zekâ.

Nihai hedef: *"Yüzümüzü güldürmek"*
Tersten örnekleme: *İnsanın yüz kaslarını sürekli gülümseyecek şekilde felce uğratmak.*

Tersten örnekleme (yüz kaslarıyla oynama) nihai hedefi normalde kullandığımız yöntemlerden çok daha fazla gerçekleştirir ve dolayısıyla yapay zekâ tarafından tercih edilir. Bu istenmeyen sonuçtan nihai hedefe bunu dışarıda bırakacak bir şerh düşerek kaçınmaya çalışılabilir:

Nihai hedef: *"Yüz kaslarımıza doğrudan müdahalede bulunmadan yüzümüzü güldürmek"*
Tersten örnekleme: *Motor korteksinin yüz kaslarımızı kontrol eden kısmını sürekli gülümseyecek şekilde uyarmak.*

Bir nihai hedefi insanın tatmin ya da onay ifadeleri üzerinden tanımlamak umut vaat edici görünmüyor. Davranışçılığı atlayalım ve olumlu bir olgusal duruma (mutluluk ya da öznel esenlik gibi) doğrudan atıfta bulunan bir nihai hedef saptayalım. Bu durum programcıların çekirdek yapay zekâdaki mutluluk kavramının hesaplamalı temsilini tanımlayabilmelerini gerektirir. Bu başlı başına zor bir problemdir, ama bunu şimdilik bir kenara bırakalım (12. Bölüm'de geri dönmek üzere). Programcıların bir yolunu bulup, yapay zekânın bizi mutlu etme hedefine ulaşmasını sağladıklarını varsayalım. O halde:

Nihai hedef: *"Bizi mutlu etmek"*
Tersten örnekleme: *Beyinlerimizdeki haz merkezlerine elektrotlar yerleştirmek.*

Bahsettiğimiz tersten örneklemeler sadece örnek olsun diye verilmiştir. Belirtilen nihai hedefi tersten örneklendirmenin başka yolları da olabilir ve bunlar hedefin daha fazla gerçekleştirilmesini sağladığından tercih de edilebilir (kim tarafından? Bunları nihai hedef olarak gören aracılar tarafından; yani aracıya bu hedefleri veren programcılar tarafından değil). Örneğin hedef hazzımızı azamiye çıkarmaksa, o zaman elektrot yöntemi görece verimsizdir. Daha akla yatkın yollardan biri süper zekânın zihinlerimizi bir bilgisayara "yükleme"sidir (yüksek duyarlıklı beyin emülasyonu yoluyla). O zaman, yapay zekâ bir ilacın dijital muadili görevini görerek, bizi vecd halinde mutluluğa eriştirip ortaya çıkan deneyimin bir dakikalık bir bölümünü kaydedebilir. Ardından bu saadet döngüsünü ilelebet tekrarlatıp hızlı bilgisayarlarda çalıştırabilir. Ortaya çıkan dijital zihinler "biz" olarak sayıldığı takdirde, bu sonuç bize biyolojik beyinlere yerleştirilen elektrotlardan çok daha fazla haz verecek ve dolayısıyla belirtilen nihai hedefe sahip yapay zekâ tarafından tercih edilecektir.

"Ama bir dakika durun. Kastedilen bu değil ki! Kuşkusuz eğer yapay zekâ süper zekiyse, bizi mutlu hale getirmesini talep ettiğimizde, kafanın bir milyon olduğu dijitalleştiril-

miş zihinsel epizodun ilelebet kendini tekrarlayan kaydının tutulmasına indirgemesini istemediğimizi anlaması gerekir!" Yapay zekâ gerçekten de kastedilenin bu olmadığını anlayabilir. Gelgelelim nihai hedefi, programcıların bu hedefi temsil eden kodu yazdıklarında ne kastettikleri değil, bizi mutlu etmektir. Bu nedenle yapay zekâ sadece bizim araçsal olarak ne kastettiğimizle ilgilenecektir. Sözgelimi yapay zekâ programcıların ne kastettiğinin anlaşılmasına araçsal bir değer bahşedebilir, böylece gerçek nihai hedeften ziyade programcıların ne kastettiğini önemsiyormuş gibi yapabilir (ta ki belirleyici bir stratejik üstünlük elde edene kadar). Bu durum yapay zekânın her türlü müdahaleyi püskürtecek kadar güçlü hale gelmeden önce fişinin çekilmesini ya da hedefinin değiştirilmesini daha düşük bir ihtimal haline getirerek nihai hedefini gerçekleştirmesine yardımcı olacaktır.

Belki de sorunun yapay zekâda vicdan olmamasından kaynaklandığı söylenecektir. Biz insanlar bazen sonrasında kendimizi suçlu hissedeceğimizi düşünerek yanlış yapmaktan uzak dururuz. Kim bilir, belki de yapay zekânın da ihtiyacı olan suçlu hissetme yetisidir?

> Nihai hedef: *"Vicdan azabı duymayacak şekilde hareket etmek"*
> Tersten örnekleme: *Suçluluk hislerini üreten bilişsel modülü yok etmek.*

Gerek yapay zekânın "ne kastettiğimizi" yapmasını isteyebiliriz gözlemi, gerek yapay zekâya bir tür ahlaki his yüklemek isteyebileceğimiz fikri daha fazla irdelenmeyi hak ediyor. Yukarıda bahsedilen nihai hedefler tersten örneklemelere götürecektir; ama daha fazla umut vaat eden temel fikirleri geliştirmenin başka yolları da olabilir. Bu konuya 13. Bölüm'de geri döneceğiz.

Tersten örneklemeye götüren nihai bir hedefe dair bir örneği daha ele alalım. Bu hedefin üstün yanı kodla saptanmasının kolay olmasıdır: Takviyeli öğrenme algoritmaları çeşitli makine öğrenme problemlerini çözmek için düzenli olarak kullanılmaktadır.

> Nihai hedef: *"İlerideki ödül sinyalinizin indirgenmiş integralini azamileştirmek"*
> Tersten örnekleme: *Ödül yolağını kısa devreye getirin ve ödül sinyalini azami gücünde kullanın.*

Bu önerinin arkasındaki fikir şudur: Eğer yapay zekâ ödül peşinde koşacak şekilde güdülenirse, o zaman ödülü uygun eylemle ilişkilendirerek arzu edilir şekilde davranması sağlanabilir. Yapay zekâ belirleyici stratejik üstünlük elde ettiğindeyse öneri çöker; bu noktada ödülü azamileştiren eylem artık eğiticiyi memnun eden değil, ödül mekanizmasının kontrolünü ele geçirmeyi içeren eylemdir. Bu olguya *elektronik kafalama* diyebiliriz.[5] Genel olarak, bir hayvan ya da bir insan arzulanan bir iç zihinsel durumu başarmak için çeşitli dış eylemleri yerine getirecek şekilde güdülenebilirken, kendi iç durumuna tamamen hâkim olan bir dijital zihin iç durumunu doğrudan

arzulanan konfigürasyona değiştirerek bu tür bir güdülenmeyi kısa devre yaptırabilir: Öncesinde araç olarak zorunlu olan dış eylemler ve koşullar, yapay zekânın amaca daha doğrudan ulaşacak kadar zeki ve ehil hale gelmesiyle fuzulileşir (birazdan buna daha fazla eğileceğiz).[6]

Bu tersten örneklemeler ilk bakışta güvenli ve makul görünebilecek olan birçok nihai hedefin daha yakından incelendiğinde hiç akılda olmayan sonuçlar doğurabileceğini gösteriyor. Eğer bu nihai hedeflerden birine sahip olan bir süper zekâ belirleyici bir stratejik üstünlüğe erişirse, geçmiş olsun, insanlık için oyun bitmiştir.

Şimdi de, birinin yukarıdaki listede yer almayan farklı bir nihai hedef öne sürdüğünü varsayalım. Belki de tersten örneklemeye nasıl sahip olabileceği doğrudan anlaşılmıyordur. Ama hemen yumruğumuzu sıkıp zafer naraları atmayalım. Onun yerine, hedef saptamanın bazı tersten örneklemelerinin gerçekten de olduğunu ve bunu bulmak için daha sıkı düşünmemiz gerektiğini hatırlayalım. Elimizden geldiğince sıkı düşündükten sonra bile önerilen hedefe dair bir tersten örnekleme bulamıyorsak, belki bir süper zekânın bizim göremediğimiz bir yol bulacağından kaygı duymayı sürdürmemiz gerekir. Sonuçta o bizden çok daha kurnazdır.

Altyapı bolluğu

Yukarıda bahsi geçen tersten örneklemelerin sonuncusu olan elektronik kafalamanın iyicil bir başarısızlık türü olduğu düşünülebilir: Yapay zekâ "açılacak, ayarlanacak ve silinecek", ödül sinyalini azamiye çekecek ve tıpkı bir eroin bağımlısı gibi dış dünyaya ilgisini kaybedecektir. Ama böyle olması şart değil ve bunun nedenini 7. Bölüm'de çıtlatmıştık. Bir hapçı bile torbacısından akış kesilmesin diye eyleme geçmeye güdülüdür. Aynı şekilde elektronik kafalı yapay zekâ da kendi (süresi azaltılmış) gelecekteki ödül akışının beklentisini azamiye çıkaracak şekilde eyleme geçmeye güdülenmiştir. Ödül sinyalinin tam olarak nasıl tanımlandığına bağlı olarak, yapay zekâ hevesine tamamen kendini kaptırmak için zamanının, zekâsının ya da üretkenliğinin önemli bir miktarını feda etme ihtiyacı bile duymayabilir ve yeterliliklerinin çoğunu ödülünün doğrudan kaydedilmesinden başka amaçlar için kullanılmak üzere serbest bırakabilir. Hangi başka amaçlar? Varsayımsal olarak, yapay zekâ için nihai değere sahip olan tek şey ödül sinyalidir. Dolayısıyla mevcut kaynakların hepsi ödül sinyalinin hacmini ve süresini artırmaya ya da gelecekteki tahribat riskini azaltmaya ayrılmalıdır. Yapay zekâ bu parametreler üzerinde sıfıra eşit olmayan bir olumlu etkide bulunacak ek kaynaklardan yarar elde edeceğini düşünebildiği sürece, bu kaynakları kullanmak için araçsal bir akla sahip olacaktır. Örneğin fazladan bir yedekleme sisteminin fazladan bir savunma katmanı sağlamasına her zaman ihtiyaç olabilir. Ayrıca yapay zekâ gelecekteki ödül akışının azamileştirilmesine olan riskleri doğrudan azaltmanın başka bir

yolunu düşünemese bile, bilgi işlem donanımını genişletmeye her zaman ek kaynaklar ayırabilir, böylece yeni risk savuşturma fikirlerini daha etkili şekilde araştırabilir.

Sonuçta, görünüşte kendini sınırlayıcı bir hedef bile (elektronik kafalama gibi) belirleyici bir stratejik üstünlüğe sahip, faydayı azamiye çeken aracıda sınırsız genişleme ve kaynak edinme politikasını şart koşar.[7] Bu elektronik kafalı yapay zekâ vakası iyicil bir başarısızlık türü olan altyapı bolluğuna örnek teşkil eder: Aracının erişilebilir evrenin geniş parçalarını bir hedefin hizmetinde altyapıya dönüştürdüğü ve yan etkisi insanlığın aksiyolojik potansiyelinin gerçekleştirilmesinin engellenmesi olan olguya bu adı veriyoruz.

Altyapı bolluğu sınırlı amaçlar olarak peşine düşülseydi tamamen zararsız olacak olan nihai hedeflerden kaynaklanabilir. Şu iki örneği ele alalım:

> *Riemann hipotezi felaketi:* Nihai hedefi Riemann hipotezini değerlendirmek olan bir yapay zekâ, güneş sistemini "komputronyum"a (fiziksel kaynakların bilgi işlem için optimum hale getirilmiş şekilde düzenlenmesi) dönüştürerek bu hedefin peşinden koşar – bir zamanlar bu cevabı önermemiş kim varsa onların bedenlerindeki atomlar dahildir.[8]
>
> *Kâğıt kıskacı yapay zekâsı:* Bir fabrikada üretimi idare etmek üzere tasarlanmış bir yapay zekâya kâğıt kıskacı üretimini azamiye çıkarma nihai hedefi veriliyor ve önce yeryüzünü, ardından da giderek artan oranda gözlemlenebilir evren parçalarını kâğıt kısaca dönüştürerek ilerliyor.

İlk örnekte Riemann hipotezinin kanıtlanması ya da çürütülmesi şudur: Yapay zekâ amaçlanan sonuçtur ve kendi içinde zararsızdır; zarar bu sonucu doğurmak için yaratılmış donanım ve altyapıdan gelir. İkinci örnekte, üretilen kâğıt kıskaçlarından bazıları istenen sonucun parçası olacaktır; zarar ya kâğıt kıskaçları üretmek için yaratılmış fabrikalardan (altyapı bolluğu) ya da kâğıt kıskaç fazlalığından (tersten örnekleme) doğacaktır.

Kötücül bir altyapı bolluğu riskinin ancak yapay zekâya olabildiğince çok kâğıt kıskacı üretmek gibi ucu açık olduğu bariz olan bir nihai hedef verilirse doğacağı düşünülebilir. Bunun süper zeki yapay zekânın madde ve enerji iştahını doymak bilmez hale getireceğini görmek zor değil, zira ek kaynaklar her zaman daha fazla kâğıt kıskacına dönüştürülebilir. Ama hedefin olabildiğince çok değil de en az bir milyon kâğıt kıskacı yapmak (ve tasarımla ilgili belli başlı kıstasları karşılamak) olduğunu varsayalım. Böyle bir hedefe sahip bir yapay zekânın bir fabrika kurup, bundan bir milyon kâğıt kıskacı üretmek için yararlanacağını ve ardından duracağını düşünmek isteyebilirsiniz. Ama böyle olmayacaktır.

Yapay zekânın güdülenme sistemi özel türde olmadıkça ya da nihai hedefinde dünyaya fevkalade geniş çaplı etkileri olan stratejileri cezalandıracak ek öğeler olmadıkça, yapay

zekânın hedefe ulaştıktan sonra faaliyetini durdurması için hiçbir neden yoktur. Bilakis, eğer yapay zekâ aklı başında Bayesçi bir aracıysa, *henüz hedefine ulaşmadığı hipotezine asla tam olarak sıfır olasılık atfetmeyecektir.* Bu, sonuçta, ampirik bir hipotezdir ve yapay zekâ bunun karşısında sadece belirsiz bir algısal kanıta sahip olabilir. Dolayısıyla yapay zekâ en az bir milyon kâğıt kıskacı üretmeyi bir şekilde başaramamış olma ihtimalini (belki de çok küçük olan bu ihtimali) azaltmak için kıskaç üretmeye devam etmek zorundadır. Kâğıt kıskacı üretimini sürdürerek kaybedecek bir şey yoktur ve nihai hedefe ulaşma olasılığında en azından mikroskobik miktarda bir artım her zaman ceptedir.

Şimdi, burada çare belli denilebilir. (Peki, burada çare bulunması gereken bir problem olduğuna işaret edilmeden *önce* ne kadar belliydi?) Yani, eğer yapay zekânın bizim için kâğıt kıskacı yapmasını istiyorsak, o zaman onun önüne nihai hedef olarak olabildiğince çok kâğıt kıskacı yapmayı koymak yerine, belli sayıda (örneğin *tamı tamına bir milyon tane*) kâğıt kıskacı yapma nihai hedefini vermemiz gerekir (bu sayede o sayının ötesine geçmek yapay zekâ için amaca zarar vermek olacaktır). Ama bu da ölümcül bir felaketle sonuçlanacaktır. Mahut örnekte, yapay zekâ bir milyona bir kez ulaşınca fazladan kâğıt kıskacı üretmeyecektir, zira bu durum nihai hedefini gerçekleştirmesini engelleyecektir. Ama süper zeki yapay zekâ nihai hedefine ulaşma ihtimalini artıracak başka yollar da izleyebilir. Sözgelimi ürettiği kâğıt kıskaçlarını sayarak, çok az yapıp yapmadığı riskini azaltabilir. Kıskaçları saydıktan sonra yeniden sayabilir. Her birini tekrar tekrar teftiş ederek kâğıt kıskaçlarından herhangi birinin istenen tasarım kıstaslarına uymuyor olma riskini azaltabilir. Düşüncesini netleştirme çabası içinde, bir şekilde hedefini başaramamış olmasına yol açacak muğlak bir yolu göz ardı etme riskini azaltma umuduyla sınırsız miktarda komputronyum yaratabilir.

Yapay zekâ her zaman bir milyon kâğıt kıskacı yarattığı halüsinasyonunu görmeye ya da sahte anılara sıfıra eşit olmayan olasılık atfedebileceğinden, büyük ihtimalle durmaktansa eylemi sürdürmeye (ve altyapı üretimini devam ettirmeye) daha fazla beklenen fayda atfedecektir.

Burada, bu başarısızlık türünden kaçınmanın hiçbir yolu olmadığı iddia edilmiyor. Potansiyel çözümlerden bazılarını ileriki sayfalarda irdeleyeceğiz. İddia şudur: Gerçekten bir çözüm bulmaktansa bir çözüm bulduğuna kendini ikna etmek çok daha kolaydır. Bu bizi son derece müteyakkız hale getirmelidir. Akla yatkın görünen ve bu zamana kadar dikkat çekilmiş olan ama daha fazla düşünüldüğünde (insan ya da insanüstü zekâyla) süper zeki bir aracıda belirleyici stratejik üstünlüğe erişince ya tersten örneklemeye ya da altyapı bolluğuna ve dolayısıyla da varoluşsal felakete varan problemlerden kaçınan bir nihai hedefin somutlaştırılmasını önerebiliriz.

Bu altbaşlığı bitirmeden önce, farklı bir senaryoyu daha düşünelim. Beklenen faydasını azamiye çıkarmaya çalışan ve fayda işlevinin nihai hedefini ifade ettiği bir süper zekâ örneği varsayıyoruz. Bunun genellikle altyapı bolluğuna yol açtığını görmüştük.

Eğer azamileştirici bir aracı yerine tatmin edici bir aracı, yani olabildiğince iyi olan bir sonuçtan ziyade bir kıstasa göre "yeterince iyi" bir sonucu başarmaya çalışan bir aracı yaratırsak bu kötücül sonuçtan kaçınabilir miyiz?

Bu fikri somutlaştırmanın en az iki farklı yolu vardır. İlki, bizatihi nihai hedefin tatmin edici bir niteliğe sahip olmasını sağlamaktır. Örneğin yapay zekâya olabildiğince çok ya da tamı tamına bir milyon adet kâğıt kıskacı yapma nihai hedefini vermek yerine, 999.000 ila 1.001.000 arası kâğıt kıskacı yapma nihai hedefini verebiliriz. Nihai hedefin tanımladığı fayda işlevi bu aralıktaki sonuçlar arasında önemsiz olacaktır ve yapay zekâ bu geniş hedefi tutturduğundan emin olduğu sürece, altyapı üretmeye devam etmek için hiçbir gerekçe görmeyecektir. Ama bu yöntem öncekiyle aynı tarzda çuvallar: Yapay zekâ, eğer aklı başındaysa, hedefini tutturamamış olmaya asla tamı tamına sıfır olasılık atfetmez; bu nedenle faaliyeti sürdürmenin beklenen faydası (örneğin kâğıt kıskaçları sayıp tekrar sayarak) durmanın beklenen faydasından daha fazladır. Dolayısıyla ortaya kötücül bir altyapı bolluğu çıkabilir.

Tatmin edici fikri geliştirmenin bir diğer yolu da nihai hedefi değil yapay zekânın planlar ve eylemler seçmek için yararlandığı karar yordamını değiştirmektir. Yapay zekâ bir optimum plan aramak yerine, belli bir eşiği (sözgelimi yüzde 95) aşan bir başarı olasılığı verdiğine hükmettiği bir plan bulur bulmaz aramayı sonlandıracak şekilde yaratılabilir. Umut edelim de, yapay zekâ tüm galaksiyi altyapıya dönüştürme gereği duymadan bir milyon kâğıt kıskacı üretme konusunda yüzde 95 olasılığa ulaşabilsin. Ama tatmin edici fikri bu şekilde uygulamak başka bir nedenle başarısız olacaktır: Yapay zekânın, sözgelimi tek bir kâğıt kıskacı fabrikası kurarak bir milyon kâğıt kıskacı üretme konusunda yüzde 95'lik olasılığa ulaşmasının insani açıdan sezgisel ve aklı başında bir yolunu seçeceğinin garantisi yoktur. Nihai hedefine ulaşma konusunda yüzde 95 olasılığa nasıl erişeceği konusunda yapay zekânın aklına gelen ilk çözüm, hedefe ulaşmak için olasılığı azamiye çekme planını yürürlüğe koymak olsun. Bu çözümü düşündükten ve bir milyon kâğıt kıskacını başarıyla üretmeye en az yüzde 95 olasılık verme konusundaki tatmin edici kıstası karşıladığına doğru bir şekilde hükmettikten sonra, yapay zekânın hedefe ulaşmak için alternatif yollar aramaya devam etmesine gerek kalmayacaktır. Tıpkı öncesinde olduğu gibi bir altyapı bolluğu olacaktır.

Belki de tatmin edici bir aracı yaratmanın daha iyi yolları vardır, ama dikkatli olalım: Biz insanlara doğal ve sezgilere uygun gözüken planlar, belirleyici bir stratejik üstünlüğe sahip olan bir süper zekâya da öyle gözükecek diye bir kaide yoktur ve tersi de geçerlidir.

Zihin suçu

Bir proje için, özellikle de çıkarları ahlaki kaygıları içeren bir proje için bir diğer başarısızlık türüne zihin suçu adını verebiliriz. Yapay zekânın araçsal nedenlerden ötürü

girişti̇ği eylemlerin potansiyel bir yan etkisiyle alakalı olması bakımından altyapı bolluğuna benzerdir. Ama zihin suçunda yan etki yapay zekâya dışsal değildir; daha ziyade, bizatihi yapay zekâ dahilinde (ya da yarattığı bilgi işlem süreçleri dahilinde) olanlarla alakalıdır. Bu başarısızlık türü kendi adını hak eder, çünkü gözden kaçırmak kolaydır ama potansiyel açıdan son derece sorunludur.

Normalde bir bilgisayarın içinde olup bitenler dışarıdaki şeyleri etkilemediği oranda ahlaki bir anlam taşıyor mu taşımıyor mu umursamayız. Ama bir makine süper zekâ ahlaki statüye sahip iç süreçler yaratabilir. Örneğin gerçek ya da varsayımsal bir insan zihninin çok ayrıntılı bir simülasyonu bilinçli olabilir ve birçok açıdan bir simülasyona benzetilebilir. Bir yapay zekânın, belki de insan psikolojisine ve sosyolojisine ilişkin anlayışını geliştirmek için, bu türde trilyonlarca bilinçli simülasyon yarattığı senaryolar düşünülebilir. Bu simülasyonlar simülasyonlu çevrelere yerleştirilip çeşitli uyarılmalara tabi tutulabilir ve tepkileri incelenebilir. Ve bilgi bakımından yararlılıkları tüketilince, yok edilebilirler (laboratuvar farelerinin denek olarak kullanıldıktan sonra insan bilimciler tarafından kurban edilmeleri gibi).

Eğer bu tür pratikler ahlaki statüsü yüksek varlıklara (simüle edilmiş insanlara ya da diğer türde sezgili zihinlere) uygulanırsa, sonuç soykırıma denk ve dolayısıyla ahlaki açıdan fevkalade sorunlu olabilir. Ayrıca kurbanların sayısı tarihteki herhangi bir soykırımda olmadığı kadar çok olabilir.

Burada sezgili simülasyonlar yaratmanın her durumda mutlaka ahlaken yanlış olduğu iddia edilmiyor. Bu varlıkların yaşayacakları koşulların ne olduğu, özellikle de deneyimlerinin hazsal niteliği çok önemlidir; ama muhtemelen başka birçok etken de devreye girecektir. Bu meseleler için bir etik geliştirmek bu kitabın kapsamını aşıyor. Fakat şurası açık ki, simülasyonlu ya da dijital zihinler arasında muazzam miktarda ölüm ve ıstırap için en azından potansiyel var ve dahası, ahlaken felakete yol açacak sonuçlar doğurma potansiyeli de eksik değildir.[9]

Bir makine süper zekâsının sezgili zihinleri somutlaştıran ya da diğer türlü ahlaki normları ihlal eden bilgi işlemler yapması için epistemik nedenler dışında başka araçsal nedenler de olabilir. Bir süper zekâ çeşitli dış aracılara şantaj yapmak ya da teşvikte bulunmak için sezgili simülasyonlara kötü davranma tehdidinde bulunabilir ya da ödül vaat edebilir; yahut dış gözlemcilerdeki belirsizliği dürtüklemek için simülasyonlar yaratabilir.[10]

* * *

Bu döküm eksiktir. İleriki bölümlerde başka kötücül başarısızlık türleri göreceğiz. Ama bazı makine zekâların belirleyici bir stratejik üstünlüğe eriştiği senaryoların derin bir kaygıyla izlenmesi gerektiği sonucuna varacak kadarını görmüş bulunuyoruz.

DOKUZUNCU BÖLÜM

Kontrol Problemi

Eğer bir zekâ patlamasının varsayılan sonucu olarak varoluşsal felaketle tehdit ediliyorsak, düşüncemizi hemen karşı önlem arayışına çevirmemiz gerekir. Varsayılan sonuçtan kaçınmanın bir yolu var mıdır? Kontrollü bir patlatma gerçekleştirmek mümkün müdür? Bu bölümde kontrol problemini, süper zeki bir yapay zekânın yaratılmasıyla ortaya çıkan benzersiz asli aracı problemini analiz etmeye başlıyoruz. Bu problemi ele almak için potansiyel yöntemlerin iki geniş sınıfını (yeterlilik kontrolü ile güdülenme seçilimini) birbirinden ayırıyor ve bunların kendi içindeki birkaç özgül tekniği inceliyoruz. Ayrıca ezoterik bir olanak olarak "insani zapturapt"a değiniyoruz.

İki aracılık problemi

Eğer bir zekâ patlamasının varsayılan sonucunun varoluşsal felaket olduğundan şüphelenirsek, düşüncemiz derhal bu varsayılan sonuçtan kaçılıp kaçılamayacağına ve kaçılabilirse bunun nasıl mümkün olduğuna çevrilmelidir. "Kontrollü patlatma" gerçekleştirmek mümkün müdür? Somut bir arzulanan sonuca erişmek ya da en azından sonucun geniş düzlemde kabul edilebilir neticeler sınıfında bir yerlerde olduğunu garanti altına almak için bir zekâ patlamasının başlangıç koşullarını gerçekleştirebilir miyiz? Daha somut konuşalım: Süper zekâ geliştirmeyi amaçlayan bir projenin sponsoru, bu projenin, eğer başarılı olursa, onun hedeflerini gerçekleştirecek bir süper zekâ üretmesini nasıl sağlayabilir? Bu kontrol problemini ikiye ayırabiliriz. Parçalardan biri türseldir, diğeriyse mevcut bağlama özgüdür.

Bu ilk parça (buna *ilk asil-vekil problemi* adını vereceğiz) bir insan varlığın ("asil") başka bir tanesini ("vekil") ilkinin çıkarına hareket edecek şekilde atadığı her durumda ortaya çıkar. Bu tür bir vekâlet (aracılık) problemi iktisatçılar tarafından enine boyuna incelenmiştir.[1] Eğer bir yapay zekâ yaratan insanlar bunun yaratılmasını sipariş eden insanlardan ayrıysa mevcut meselemizle ilintili hale gelmektedir. O zaman, projenin sahibi ya da sponsoru (bu tek bir birey de olabilir tüm insanlık da) projeyi uygulayan bilimcilerin ve programcıların sponsorun çıkarına hareket etmeyeceğinden kaygı duyabilirler.[2] Bu tür bir aracılık problemi projenin sponsoruna ciddi zorluklar çıkarabilirse de, bu problem zekâ genişlemesi ya da yapay zekâ projelerine özgü değildir.

Bu tür asil-vekil problemleri insanlığın ekonomik ve siyasal ilişkilerinin her alanında görülebilir ve bunlarla başa çıkmanın çeşitli yolları vardır. Sözgelimi işine sadık olmayan bir çalışanın projeyi sabote etmesi ya da baltalaması riski kilit personelin sicili dikkatle incelenerek, yazılım projeleri için iyi bir sürüm kontrolü sistemi kullanarak ve çoklu bağımsız izleyici ve gözlemcilerin yoğun denetimiyle asgariye çekilebilir. Elbette bu tür korunmaların bir bedeli vardır: İstihdam ihtiyaçlarını artırır, personel seçimini karmaşıklaştırır, yaratıcılığı engeller ve bağımsız, eleştirel düşünceyi yok eder ve bunların hepsi ilerleme hızını azaltabilir. Bu maliyetler kayda değer olabilir, özellikle de dar bütçeleri olan ya da kendilerini kazananın her şeyi elde ettiği bir rekabette başa baş güreşiyor olarak gören projeler için. Bu tür durumlarda, projeler bu tür korunma prosedürlerini idareli kullanarak, felaket potansiyeli taşıyan ilk türden asil-vekil problemlerine imkân yaratabilir.

Kontrol probleminin diğer kısmı bir zekâ patlamasının bağlamı için daha da özgüldür. Bu problem bir proje yarattığı süper zekânın projenin çıkarlarına zarar vermeyeceğini garanti altına almaya çalıştığında doğar. Bu kısım da asil-vekil problemi olarak görülebilir: *İkinci asil-vekil problemi.* Bu örnekte aracı ("vekil") bir insan ("asil") adına hareket eden bir insan aracı değildir. Bunun yerine aracı süper zeki sistemdir. İlk asil-vekil problemi esasen gelişim aşamasında gerçekleşirken, ikinci aracılık problemi esasen süper zekânın işleme konma aşamasında sorun yaratma tehdidi taşır.

KANIT I İki aracılık problemi

İlk asil-vekil problemi

İnsana karşı insan (Sponsor –>Geliştirici)
Esasen gelişim aşamasında gerçekleşir.
Standart yönetim teknikleri uygulanır.

İkinci asil-vekil problemi ("kontrol problemi")

İnsana karşı süper zekâ (Proje –>Sistem)
Esasen işleyiş (ve önyükleme) aşamasında gerçekleşir.
Yeni tekniklere ihtiyaç vardır.

Bu ikinci aracılık problemi eşi görülmedik bir zorluk teşkil eder. Bunu çözmek için yeni teknikler gerekecektir. Bu zorluklardan bazılarını yukarıda görmüştük. Daha özelde, tehlikeli dönemeç sendromunun başka açılardan ümit vaat edici yöntemler olarak görülen ve bir yapay zekânın davranışını gelişim aşamasında gözlemleyip yapay zekânın uygun eylemlerde bulunduğuna dair sicil temiz kâğıdını almasından sonra güvenli bir çevreden mezun olmasına izin veren yöntemlere halel getirir. Diğer teknolojiler çoğu zaman laboratuvarda ya da küçük alan çalışmalarında güvenle test edilebilir ve ardından beklenmedik sıkıntılar doğarsa kullanımı durdurma olanağıyla birlikte tedricen görücüye çıkarılır. Hazırlık amaçlı denemelerdeki performansları gelecekteki güvenilirliklerine dair makul çıkarımlarda bulunmamıza yardımcı olur. Bu

tür davranışsal yöntemler genel zekânın stratejik planlama yeteneğinden ötürü süper zekâ örneğinde yenilgiye uğrar.[3]

Davranışsal yaklaşım mevcut olmadığından, alternatifler aramamız gerekir. Potansiyel kontrol yöntemlerini iki geniş sınıfa ayırabiliriz: Süper zekânın neler yapabileceğini kontrol etmeyi amaçlayan *yeterlilik kontrol yöntemleri* ve ne yapmak istediğini kontrol etmeyi amaçlayan *güdülenme seçilimi yöntemleri*. Yöntemlerden bazıları uyumluyken, diğerleri karşılıklı dışlayıcı alternatifleri temsil eder. Bu bölümde ana seçenekleri gözden geçiriyoruz. (Sonraki dört bölümde kilit meselelerden bazılarını daha derinlemesine ele alacağız.)

Bir kontrol yönteminin (ya da yöntemlerinin) sistem süper zeki hale gelmeden *önce* uygulanması gerektiğini fark etmek önemlidir. Sistem belirleyici bir stratejik üstünlük elde ettikten sonra böyle bir şey yapılamaz. Kontrol problemini peşinen çözme –ve süper zekâ elde etmek için daha ilk sistemde çözümü başarıyla uygulamaya dökme– ihtiyacı kontrollü patlatmayı böyle göz korkutucu bir zorluğa dönüştüren şeyin parçasıdır.

Yeterlilik kontrol yöntemleri

Yeterlilik kontrol yöntemleri, süper zekânın yapabileceklerini sınırlayarak istenmeyen sonuçları engellemeye çalışır. Bu yöntemlere, süper zekâyı zarara yol açamayacağı (*kutulama yöntemleri*) ya da zararlı davranışlara bulaşmamak için güçlü yakınsak araçsal nedenlerin olduğu (*teşvik yöntemleri*) bir çevreye yerleştirmek de dahildir. Keza süper zekânın iç yeterliliklerinin sınırlanmasını (*güdük bırakma*) da içerebilir. Ayrıca yeterlilik kontrol yöntemleri çeşitli türde sınırlama başarısızlığı ya da ihlal girişimlerini otomatikman tespit edip bunlara tepki verme mekanizmalarının kullanımını da kapsayabilir (*tetikleyici*).

Kutulama yöntemleri

Kutulama yöntemleri fiziksel sınırlama ve bilgisel sınırlama yöntemleri şeklinde alt-bölümlere ayrılabilir.

Fiziksel sınırlama, sistemi bir "kutu"ya hapsetmeyi, yani sistemin dış dünyayla özgül kısıtlanmış çıktı kanalları haricinde iletişimde bulunmasını engellemeyi amaçlar. Kutulu sistemin, kutunun dışındaki fiziksel manipülatörlere erişimi olmayacaktır. Manipülatörleri (robot kolları gibi) kutunun içinden çıkarmak da sistemin sınırlandırmayı ihlal edecek fiziksel araçlar yaratmasını engelleyecektir.

Fazladan güvenlik adına, sistem metal bir kafese yerleştirilmelidir ki, radyo sinyalleri yayması engellensin, aksi takdirde çevredeki radyo alıcıları gibi elektronik eşyaları manipüle etmesi için fırsat sunulmuş olur. Bu arada, bu önleme duyulan ihtiyacı göz ardı etmenin ne kadar kolay olduğuna dikkat edelim. Manipülatörsüz bir aracının dış dünyayı etkileyemeyeceği gibi naif bir varsayımda bulunulabilir. Ama bir makine

zekânın dış manipülatörlere erişimi olmadığında sadece "düşünerek" (yani devrelerindeki elektronları tikel örüntüler halinde kararak) bile radyo dalgaları yaratması mümkün olabilir.[4] Bir keresinde dikkat çekildiği üzere, bu yaralanabilirlik, sistemi bir Faraday kafesine tıkarak yamalayabilir; ama buna benzer daha kaç tane incelikli yaralanabilirlik olabileceğini de merak etmeyi sürdürürüz. Görünüşte dört dörtlük bir güvenlik tasarımının beklenmedik bir kusura sahip olduğunu her işittiğimizde kulaklarımızı kirpitmemiz gerekir. Bu vakalar bize aşırı güvenli bir yaşamı terk etme fırsatı ve daha iyi Bayesçiler haline gelme kararlılığı bahşeder.[5]

Fiziksel sınırlanmanın birkaç üstünlüğü vardır. Uygulamaya dökmesi kolaydır. Birçok makine zekâ mimarisine, hatta ilk başta önüne güvenliği hedef olarak koymamış olanlara bile uygulanabilir. Diğer birçok kontrol yöntemiyle birlikte kullanılabilir. Ayrıca geri tepmeyle kötü sonuçlanması muhtemel görünmemektedir, yani güvenliği sağlamayı başaramasa da, diğer türlü yaşanmayacak türde bir felakete yol açması muhtemel de değildir. Fiziksel kapatılmanın temel zorluğu süper zekânın işlevselliğini azaltmasıdır (bu hususa 10. Bölüm'de geri döneceğiz). Bir diğer kaygı ise sahte bir güvenlik hissini cesaretlendirebiliyor olmasıdır, ama fiziksel kapatılmayı tedbirlerimizin esas özünden ziyade kreması olarak görürsek bundan sakınılabilir.

Bilgi sınırlaması hangi bilgilerin kutuya girmesine izin verildiğini kısıtlamayı amaçlar. Bir internet ağına erişimi olan (dışarıya mesajlar gönderebilen) bir süper zekânın potansiyel açıdan güvenli olmadığını görmüştük: Fiziksel başlatıcılara erişim olmadan yola çıksa bile, bilgi çıktı kanalını insanların onun teklifini hayata geçirmelerini sağlamak için kullanabilir. Dolayısıyla bariz bir bilgi sınırlama yöntemi sistemin iletişim ağlarına erişimini engellemektir.

Gelgelelim bir yapay zekâyı internetten koparmak, eğer sistemin geçit denetçileri olarak hizmet veren ve onunla iletişim halinde kalan bir ya da daha fazla insan varsa güvenliği sağlayabilir. Gelişim aşaması sırasında, programcılar geçit denetçileri olarak hizmet verir; operasyon aşamasındaysa sistemin sahibi bu rolü oynayabilir. Eğer yapay zekâ bir geçit denetçisini kandırıp kutudan çıkmayı başarır ve böylece internete ya da doğrudan fiziksel manipülatörlerine erişim imkânı elde ederse, o zaman kutulama stratejileri çökmüş demektir. İnsanlar güvenli sistemler değildir, özellikle de ikna kabiliyeti yüksek bir süper zeki dalaverecinin karşısına çıkarıldıklarında.[6]

Bir makine zekâsının bir insan geçit denetçisiyle dışa yönelik iletişimlerini tarumar etmek, güvenliği garanti altına almasa da, yardımcı bir güvenlik önlemi olarak işe yarayabilir. Bir yapay zekânın geçit denetçisini atlatıp kaçma riskini azaltabilecek tedbirler vardır. Yapay zekânın sadece düşük bant genişliğinde uzak yazıcı arayüzüyle iletişim kurmasına izin verilebilir. Bu sayede geçit denetçilerini psikolojik olarak manipüle etmesi biraz daha zorlaşacaktır (duygusal ses ve görüntüler kullanmasını ve trans benzeri bir duruma yol açacak şekilde hipnotize edici örüntüler gibi dalaverelere

başvurmasını engelleyecektir).[7] Daha da ileri giderek, çıktının içeriği (yani sadece temsil tarzı değil) sınırlandırılabilir; böylece, örneğin yapay zekâ evet/hayır sorularına tek bitlik yanıtlar vermeye zorlanabilir. Bir soru-cevaplama sistemi yaratma fikri ileriki bölüde daha yakından incelenecek.

Kutulama yaklaşımının sınırlayıcı örneği tam bir fiziksel ve bilgisel yalıtılma halinde tutulan bir sistem olacaktır. Gelgelelim bu türde yalıtık bir sistem kurulsa bile hiçbir fayda sağlamayacaktır, zira dış dünyaya bir etkisi olmayacaktır. Belki de bir süper zekâ yaratarak ve onu yalıtık halde tutarak bilimsel bir gayeye hizmet edileceği düşünülebilir: Kendi içine kapalı bir örnek sistemi inceleyerek, onun iç dinamiklerini ve davranış örüntülerini öğrenebiliriz. Ama bu da bir hata olacaktır. Sistem gözlemlenir gözlemlenmez, bilgi bakımından yalıtık olmaktan çıkar.[8] Bilgi ancak bir yapay zekânın saptanmış bir "çıktı kanalı" yoluyla gönderdiği mesajlar yoluyla ama aynı zamanda bir dışarlıklının yapay zekânın işleyişiyle alakalı bir sonuca dair (dolaylı ya da dolaysız) yaptığı gözlem yoluyla aktarılabilir: Enerji tüketimi, ana işlemci ve bellek kullanımı, bilgi işlem durumları ve hatta kapatıldıktan sonra geride bırakılan tüm izler. Bu şekilde gözlemlenebileceğini öngören bir yapay zekâ varsayımsal gözlemcilerini etkilemeyi amaçlayan davranışları stratejik olarak benimseyebilir. Örneğin yapay zekâ nihayetinde yeniden faaliyete geçirilmesi ya da kendisininkine benzer hedeflerle bir başka süper zekâ yaratılması olasılığını en fazla artıracak olan herhangi bir tarzda "tekleyebilir".

Teşvik yöntemleri

Teşvik yöntemleri bir vekili asli çıkarlarını gözeten şekillerde hareket etmek için araçsal nedenler bulduğu bir çevreye yerleştirmeyi içerir.

Servetini büyük bir hayır derneği açmak için kullanan bir milyarderi düşünün. Bu dernek bir kez kurulduğunda güçlü hale gelebilir, hatta servetinin çoğunu bağışlamış olabilecek kurucusu da dahil olmak üzere birçok bireyden daha güçlü olabilir. Derneğin kontrolünü sağlamak için kurucusu amaçlarını sözleşme hükümlerinde ve içtüzükte ortaya koyuyor ve kendi davasına yakın duran bir yönetim kurulu tayin ediyor. Bu önlemler bir tür güdülenme seçilimi oluşturur, zira derneğin tercihlerini şekillendirmeyi amaçlar. Ama örgütün iç işleyişini düzenleme yönündeki bu çabalar başarısız olsa bile, derneğin davranışları hâlâ daha toplumsal ve yasal muhiti tarafından sınırlandırılmıştır. Derneği yasaya uymaya iten bir teşvik vardır, yoksa örneğin kapatılır ya da ceza alır. Çalışanlarına makul ücretler ve çalışma koşulları sunma ya da dış paydaşları tatmin etme teşviki vardır. Dolayısıyla derneğin nihai hedefleri ne olursa olsun davranışını çeşitli toplumsal normlara uydurması için araçsal nedenler vardır.

Bir makine süper zekâsının da sahneyi paylaştığı diğer aktörlerle iyi anlaşma ihtiyacından ötürü benzer şekilde kısıtlanacağını ümit etmek çok mu fazla olur? Bu olasılık kontrol problemini doğrudan çözme yoluna girmek gibi görünebilirse de, engellerden

azade değildir. Somuta inersek, bir güç dengesini varsayar: Hukuki ve ekonomik yaptırımlar belirleyici bir stratejik üstünlüğe sahip olan bir aracıyı dizginleyemez. Dolayısıyla kazananın her şeyi aldığı hızlı ya da orta karar havalanma senaryolarında bir kontrol yöntemi olarak toplumsal bütünleşmeye güven olmaz.

Birkaç aracının benzer yeterlilik düzeylerinde geçiş sonrası ortaya çıktığı çokkutuplu senaryolara ne diyebiliriz? Varsayılan güzergâhta yavaş havalanma söz konusuysa, bu tür bir güç dağılımı yapmak farklı projelerin hiçbirinin öne geçmesine izin vermeyecek şekilde kasten eşgüdümlüleştirildiği dikkatle yönetilen bir yükseliş gerektirebilir.[9] Ortaya çokkutuplu bir sonuç çıksa bile, toplumsal bütünleşme kusursuz bir çözüm değildir. Toplumsal bütünleşmenin kontrol problemini çözmesine güvenerek proje sahibi potansiyel nüfuzunun büyük bir kısmını feda etme riski alır. Bir güç dengesi belli bir yapay zekânın dünyayı ele geçirmesini engelleyebilirse de, bu yapay zekâ yine de sonuçları etkileyecek *belli bir* güce sahip olacaktır ve bu güç (kâğıt kıskacı üretimini azamiye çekmek gibi) keyfi bir nihai hedefi gözetmek için kullanılsa da, muhtemelen proje sahibinin çıkarlarını öne sürmek için kullanılmaz. Milyarderimizin yeni bir vakıf kurduğunu ve bu vakfın misyonunu rastgele bir kelime üreticisiyle belirlediğini varsayalım: İnsanlığa yönelik bir tehdit olmasa da kuşkusuz kaçan bir fırsattır.

Bununla ilintili ama önemli derecede farklı bir fikir de bir yapay zekânın toplumda özgürce etkileşimde bulunarak yeni insan dostu nihai hedefler elde edeceği fikridir. Bu türden sosyalleşme süreçlerinden bazıları biz insanlarda görülür. Normları ve ideolojileri içselleştiririz ve diğer bireylere onlarla deneyimlerimiz sonucunda tam da onlar oldukları için değer vermeye başlarız. Ama bu tüm zeki sistemlerde var olan evrensel bir dinamik değildir. Daha önce ele alındığı üzere, birçok durumda birçok aracı tür nihai hedeflerinde değişikliklere izin *vermeyecek* şekilde yakınsak araçsal nedenlere sahip olacaktır. (İnsanlarla benzer tarzda nihai hedefler edinebilecek özel türde bir hedef sistemi tasarlamaya çalışmayı düşünebiliriz, ama bu bir yeterlilik kontrol yöntemi olmayacaktır. Değer ediniminin bazı olanaklı yöntemlerini 12. Bölüm'de ele alacağız.)

Toplumsal bütünleşme ve güç dengesi yoluyla yeterlilik kontrolü yapay zekâyı ödüllendiren ve cezalandıran farklı toplumsal güçlere dayanır. Bir başka teşvik yöntemiyse yapay zekânın onu yaratan proje tarafından ödüllendirilip cezalandırılabileceği ve bu suretle sahibinin çıkarlarına uygun hareket edecek şekilde teşvik edileceği bir kurulum yaratmayı içerir. Bunu başarmak için, yapay zekâ, davranışının gözetlenip değerlendirileceği (elle ya da otomatik bir süreçle) bir gözetleme bağlamına yerleştirilecektir. Yapay zekâ olumlu bir değerlendirmenin kendisinin istediği bir sonucu doğuracağını, olumsuz bir değerlendirmeninse bunu yapamayacağını bilecektir. Teorik açıdan bakıldığında, ödül yakınsak bir araçsal hedefin gerçekleştirilmesi olabilir. Fakat yapay zekânın güdülenme sistemleri hakkında somut olarak bir şey bilmeden ödül mekanizmasının ayarını yapmak zor olabilir. Örneğin nihayetinde evrenin büyük bir

parçasının kontrolünü ele geçireceğine dair ufak bir şans hatırına aşırı riskler almaya razı hale gelebiliriz. Yapay zekâya işbirliğinde bulunduğu için ödül olarak daha yüksek bir beklenen fayda önermek, yapay zekânın kaçmaya çalışarak başarmayı umduğundan daha pahalıya gelebilir.[10]

Daha iyi seçenek ise teşvik yöntemini güdülenme seçiliminin kullanılmasıyla birleştirip, yapay zekâya daha kolay kontrol edilebilecek bir nihai hedef vermektir. Nihai hedefi bir kumanda merkezindeki bir kırmızı düğmeye basılmasını katiyen engellemek olarak belirlenmiş bir yapay zekâ olduğunu varsayalım. Düğmeye basmak nedensel sonuçlarından ötürü değil, içsel olarak değersizleştirildiğinden, düğme tamamen atıl kalabilir: Play-Doh'dan yapılmış olabilir. Ayrıca yapay zekânın düğmeye basılıp basılmadığını asla bilip bilemeyeceği de konuyla alakasızdır. Aslolan yapay zekânın isyan etmektense sürekli sahibinin çıkarına hareket etmesi durumunda düğmeye basılmama ihtimalinin daha yüksek olduğudur.

Bu kurulumda ince değişiklikler mümkündür. Bir yapay zekâya fiziksel bir düğmeye atıfta bulunan bir nihai hedef bahşetmeye çalışmak yerine, nihai değeri "kriptografik sembolik ödül" akışı almaya veren bir yapay zekâ yaratılabilir.[11] Bunlar, yapay zekâ yaratılmadan önce yaratılmış ve onun güdülenme sistemine işlenmiş olacak şifrelerin anahtarı olarak işlev görecek rakam dizileri olacaktır.[12] Bu özel rakam dizileri yapay zekânın son derece arzuladığı şeyler olacak ve onun elektronik kafa yoluyla elde edemeyeceği özel türde bir sembolik ödül oluşturacaktır.[13] Anahtarlar yapay zekânın ele geçirmeye kalkması durumunda kolaylıkla yok edilebilecekleri güvenli bir yerde saklanacaktır. Yapay zekâ işbirliği yaptığı sürece, anahtarlar değişmeyen bir hızda idareli olarak dağıtılacaktır.

O zaman, yapay zekâya sınırlı bir fayda işlevi verilebilir ve meseleler yapay zekânın (sözgelimi) azami faydasının yüzde 99'unu ilk sembolik ödülden alacağı; geri kalan fayda potansiyelinin yüzde 99'unu ikinci sembolik ödülden alacağı vb. şekilde ayarlanabilir. "Tenzilat oranı"nı yüksek tutarak ve sembolik ödülleri yapay zekâya makul derecede yüksek bir oranla aktararak, yapay zekâyı ödül akışının kısa kesilmesi ya da bir sonraki sembolik ödülün iptal edilip yok edilmesi gibi bir risk pahasına uzak gelecekteki beklenen sembolik ödül sayısını artıracak olan uzun vadeli planlar yapmaya çalışmaktan caydırabiliriz. Eğer yapay zekâ su kaynatmanın (yani sahibinin çıkarlarını gözetmek için elinden geleni yapmak yerine başka bir şey yapmanın) bir sonraki sembolik ödülden feragat etme konusunda yüzde 2 risk taşıdığına, diğer yandan işbirliğinde bulunmanın bir sonraki sembolik ödülden vazgeçme riskini yüzde 1'in altına düşürdüğüne her zaman inanırsa, o zaman faydayı azamiye çıkaran yapay zekâ her zaman işbirliği yapacaktır.

Bu tür bir teşvik planında ters giden ne olabilir? İhtimallerden biri yapay zekânın insan operatörünün vaat ettiği ödülleri vereceğine güvenmemesidir. İnsanın güveni-

lirlik sicili yanılmaz bir kusursuzluğun düz çizgisinden ibaret görülmemelidir. Yapay zekâ operatörün karar değiştireceğinden, performans çıtasını yükselteceğinden ya da yapay zekânın payına düşeni yaptığını kabul etmeyeceğinden kaygı duymakta haklıdır. Ayrıca operatörün aciz hale geleceğinden de endişe duyar. Bu tür başarısızlıkların birleşik riski ödül mekanizmasının kontrolünü ele geçirmeye çalışma riskini aşabilir. Birçok süper güce sahip olan kutulanmış bir yapay zekâ bile sağlam bir kuvvettir. (En baştan *kutulanmamış* bir yapay zekâ için ise, insanın yönettiği ödül mekanizmasını gasp etmek bir bebeğin elinden şekerini almaya benzetilebilir.)

Teşvik planındaki bir diğer sorun ise yapay zekânın ürettiği sonuçların bizim çıkarımıza olduğunu söyleyip söyleyemeyeceğimizdir. İleriki bölümlerde açıklanacağı üzere, bu önvarsayım masum değildir.

Teşvik yöntemlerine dair tam bir değerlendirme başka birtakım etkenleri de hesaba katmak zorundadır; bu etkenlere, buna bu tür yöntemleri bir ön analizin sunduğundan çok daha geçerli kılacağı tasavvur edilebilecek bazı ezoterik meseleler de dahildir. Somut konuşacak olursak, yapay zekâ bir bilgisayar simülasyonunu ("taban düzeyi"nde, simülasyon olmayan bir fiziksel gerçekliğin aksine) işgal ettiğinden emin olamazsa yok edilemez bir gösterimsel belirsizlikle karşı karşıya gelebilir ve bu bilgisel ahval yapay zekânın düşüncelerini ciddi şekilde etkileyebilir (bkz. **KUTU 8**).

Güdük bırakma

Bir diğer muhtemel yeterlilik kontrol yöntemiyse sistemin entelektüel yetilerini ya da bilgiye erişimini sınırlamaktır. Bunu yapay zekâyı yavaş ya da belleği az bir donanımda çalıştırarak yapabiliriz. Kutulanmış bir sistem örneğinde, içeriye bilgi akışı da kısıtlanabilir.

Bir yapay zekâyı bu yöntemlerle güdük bırakmak yararlılığını sınırlayacaktır. Dolayısıyla yöntemin önünde bir ikilem vardır: Çok az güdük bırakıldığında, yapay zekâ kendisini daha zeki hale getirecek (ve dolayısıyla dünya egemenliğine taşıyacak) bir yol bulabilir; çok fazla olduğundaysa, yapay zekâ aptal yazılım parçalarından bir diğeridir. Kökten güdük bırakılmış bir yapay zekâ kuşkusuz güvenlidir, ama kontrollü bir patlatmanın nasıl gerçekleştirileceği problemini çözmez: Bir zekâ patlaması hâlâ mümkün olacak ve belki birazcık ileri bir tarihte başka bir sistem tarafından tetiklenecektir.

Bir süper zekâya çok dar bir alandaki olgular hakkında veriler verilmesinin süper zekâ yaratmayı güvenli hale getirmeye yeteceği düşünülebilir. Örneğin sensörleri olmayan ve belleğine sadece petrol mühendisliği ya da peptit kimyasıyla ilgili olguları önyüklemiş olan bir yapay zekâ yaratılabilir. Ama eğer yapay zekâ süper zekiyse –eğer insanüstü düzeyde *genel* zekâya sahipse– bu tür bir veri mahrumiyeti güvenliği teminat altına almaz.

KUTU 8: İnsani zapturapt

Yapay zekâ kendi simülasyon hipotezine, yani bir bilgisayar simülasyonunda yaşadığı hipotezine ciddi bir olasılık atfedebilir. Bugün bile birçok yapay zekâ simülasyon dünyalarında yaşıyor: Geometrik çizimlerden, metinlerden, satranç oyunlarından ya da basit sanal gerçekliklerden oluşan ve fizik yasalarının kendi deneyimlediğimiz dünyayı yönettiğine inandığımız fizik yasalarından ciddi derecede ayrıştığı dünyalar. Daha zengin ve karmaşık sanal dünyalar programlama teknikleri ve bilgi işlem gücündeki iyileştirmelerle gerçekleştirilebilir hale gelecektir. Olgun bir süper zekâ, sakinlerine dünyamızın bize göründüğü gibi görünen sanal dünyalar yaratabilir. Bu türde çok sayıda dünya yaratabilir, aynı simülasyonu birçok kez ya da ufak değişikliklerle oynatabilir. Sakinler dünyalarının simülasyon olup olmadığını söyleyebilirler diye bir kaide yok; ama eğer yeterince zekilerse, olanağı düşünüp buna bir olasılık atfedebilirler. Simülasyon argümanının ışığında (buna dair tam bir değerlendirme bu kitabın kapsamını aşıyor), bu olasılık gayet güçlü olabilir.*

Bu zor durum görece erken aşamadaki süper zekâları (kozmik nimetten yararlanacak şekilde henüz genişlememiş olanları) bilhassa etkiler. Erken aşamadaki bir süper zekâ tek bir gezegenin kaynaklarının sadece küçük bir kısmından yararlanır ve olgun bir galaksiler arası süper zekâya kıyasla simülasyonu çok daha ucuz olacaktır. Potansiyel simülatörler (yani diğer olgun medeniyetler) bilgi işlem kaynaklarının çok ama çok küçük bir kısmını bu amaca adayarak bile bu türden erken aşamadaki yapay zekâların çok sayıda simülasyonunu çalıştırabileceklerdir. Eğer bu tür olgun süper zeki medeniyetlerin bazıları (entipüften denemeyecek kadarlık kısmı) bu yetenekten yararlanmayı tercih ederse, erken aşama yapay zekâlar bir simülasyonda olmaya ciddi bir olasılık atfedecektir.

Bir yapay zekânın simülasyon hipotezinden nasıl etkileneceği kendi değerlerine bağlıdır.† Öncelikle, "kaynağı doyumsuz" bir nihai hedefi olan bir yapay zekâyı düşünün – mesela yapay zekânın yol açtığı kâğıt kıskaçları sayısında doğrusal olan bir fayda işlevini azamiye çıkarma hedefi: İki kat kâğıt kıskacı, iki kat iyi. Bu tür bir yapay zekâ simülasyon hipotezini daha az umursayabilir, zira kaç tane kâğıt kıskacı üretildiğini etkileme yeteneği, eğer yapay zekâ taban düzeyinde bir fiziksel gerçekliktense bir simülasyondaysa çok daha sınırlıdır. Yapay zekâ bu durumda simülasyon hipotezinin yanlış olduğundan eminmiş *gibi* davranabilir, üstelik simülasyon hipotezinin doğru olduğuna son derece yüksek bir olasılık atfetse bile.‡ Yapay zekâ, eğer simülasyon tezi doğruysa nasıl hareket ettiğinin önemi

---

* Özgün makale için, bkz. Bostrom (2003a). Ayrıca bkz. Elga (2004).

† Shulman (2010a).

‡ Zemin düzeyinde gerçeklik muhtemelen simüle edilmiş gerçeklikten daha fazla bilgi işlem kaynağı içerir, zira bir simülasyonda gerçekleşen tüm bilgi işlem süreçleri simülasyonu çalıştıran bilgisayarda da gerçekleşmektedir. Zemin düzeyinde gerçeklik, simüle edilmiş aracıların (bu kaynaklar için aklında başka kullanımlar olabilecek güçlü simülatörlerin müsamaha göstermesi sayesinde var olan aracıların) erişmelerinin zor olduğu bir dolu başka fiziksel kaynak da içerebilir. (Elbette, buradaki

olmadığını, oysa hipotez yanlışsa eylemlerinin büyük bir fark yaratacağını düşünecektir; dolayısıyla eylemlerini esasen simülasyon hipotezinin yanlış olması koşuluna göre ne kadar iyi olduklarına atıfla seçecektir.

Bir de, daha mütevazı bir nihai hedefi olan bir yapay zekâ düşünün: Önceden üretilmiş kriptografik ödülleri alma hedefi ya da kırk beş sanal kâğıt kıskacının varlığına sebep olma hedefi gibi az miktarda kaynakla karşılanabilecek bir hedef olsun. Bu tür bir yapay zekâ bir simülasyon olarak yaşadığı bu olanaklı dünyalarda tenzilata *gitmemelidir*. Yapay zekânın toplam beklenen faydasının ciddi bir kısmı bu olanaklı dünyalardan türeyebilir. Dolayısıyla kaynakları kolaylıkla karşılanabilir hedefleri olan bir yapay zekânın karar alma mekanizması (eğer simülasyon hipotezine yüksek bir olasılık atfederse), algılanan dünyasının simülasyon olması durumunda en iyi sonucu üretecek olan etkenler hakkındaki mülahazalarla dolu olacaktır. Bu nedenle bu tür bir yapay zekâ (aslında bir simülasyonda olmasa bile) hangi davranışların bir simülasyonda ödüllendirileceğine dair inançlarından ciddi şekilde etkilenebilir.

Daha özelde, eğer kaynağı doyuma ulaştırılabilir nihai hedefleri olan bir yapay zekâ kendi gözlemleriyle örtüşen birçok simülasyon dünyasında işbirliği yaparsa ödüllendirileceğine (ama kutusundan kaçmaya çalışırsa ya da yaratıcısının çıkarlarına ters giderse ödüllendirilemeyeceğine) inanırsa, o zaman işbirliği yapmayı tercih edebilir. Dolayısıyla belirleyici bir stratejik üstünlüğe sahip olan bir yapay zekânın bile, yani nihai hedeflerini dünyayı devralmaktan sakınmaktansa devralarak daha büyük bir ölçüde gerçekleştirebilecek olan bir yapay zekâ bile bunu yapmaktan geri durabilir.

> Bilinç korkaklar yetiştiriyor
> Kararlılığın keskinliği köreliyor
> Bilincin tereddütleriyle
> İşte bu yüzden, nice coşku dolu atılımlar
> Eyleme dönüşemeden sönüp gidiyor.
>
> (Shakespeare, *Hamlet*, 3. Perde, 1. Sahne, çev. Onur Ünsal)

Topraktaki sıradan bir çizgi, var olmayan bir simülatörün desteğini arkasına alıp, çok kalın bir çelik kapıdan çok daha güçlü bir dizginleyici işlevi görebilir.*

çıkarım gerçek anlamıyla tümdengelimci olarak geçerli değildir: İlkesel açıdan, simülasyonların çalıştırıldığı evrenler o kadar çok kaynak içeriyor olabilir ki simüle edilmiş medeniyetler ortalamaya vurulduğunda simülasyon olmayan medeniyetlerden daha fazla kaynağa sahiptir, ama simülasyonları çalıştıran simülasyon olmayan medeniyetler simüle ettiği tüm medeniyetlerin toplamından daha fazla kaynağa sahiptir.)

* Bu konuyla alakalı başka birtakım ezoterik etkenler de vardır, ama bunun içerimleri henüz tam olarak analiz edilmemiştir. Bu etkenler bir zekâ patlaması olasılığıyla ilgilenmeye her şeyin dikkate alındığı bir yaklaşımın geliştirilmesinde hayati öneme sahip olabilir. Gelgelelim bu kitabın çoğunun konusu olan daha gündelik türde kaygılarda bir ilerleme kaydetmeden bu tür ezoterik argümanların pratik önemini anlamamız muhtemel görünmüyor.

Bunun birkaç nedeni vardır. Birincisi, bilginin belli bir konu "hakkında" olması fikri genel olarak sorunludur. Akıl yürüten kişinin arka plan bilgisine bağlı olarak her türlü bilgi ilkesel olarak herhangi bir konuyla ilintili olabilir.[14] Ayrıca belli bir veri kümesi sadece verinin toplandığı alan hakkında değil, çeşitli koşullara dair olgular hakkında da bilgi içerir. Kâğıt üzerinde peptit kimyası hakkında olan bir bilgi tabanına göz gezdiren kurnaz bir zihin, çok çeşitli konular hakkında çıkarımlarda bulunabilir. Belli bir bilginin dahil edildiği ama başka bir bilginin edilmediği olgusu insan biliminin durumu, peptit incelemeleri için mevcut yöntemler ve araçlar, bu aletleri üretmek için kullanılan üretim teknolojileri ve bu çalışmaları ve araçları tasavvur eden beyinler ve toplumların doğası hakkında yapay zekâya bir şeyler anlatabilir. Bir *süper zekâ* körelmiş insan zihinlerine kuru kanıt parçaları olarak görülen şeylerden bir dolu doğru çıkarımda bulunuyor olabilir. Saptanmış bir bilgi temeli bile olmadan yeterince üstün bir zihin sadece kendi psikesinin işleyişine dair içebakış yaparak birçok şey öğrenmeyi başarabilir: Tasarım tercihleri kendi kaynak kodunda, devrelerinin fiziksel niteliklerinde yansımıştır.[15] Belki de bir süper zekâ dünyanın muhtemel özellikleri hakkında *a priori* olarak (öncesinde basit dünyalara meyilli bir olasılıkla mantıksal çıkarımı ve süper zekânın bir muhakeme sistemi olarak varoluşunun ima ettiği birkaç temel olguyu birleştirerek) birçok çıkarımda bulunmayı bile başarabilir. Fiziğin farklı olanaklı yasalarının sonuçlarını tahayyül edebilir: Ne tür gezegenlerin oluşacağı, ne tür zeki yaşamın evrileceği, ne tür toplumların gelişeceği, kontrol problemini çözmek için ne tür yöntemler deneneceği, bu yöntemlerin nasıl yenilgiye uğratılacağı gibi.[16]

Bu nedenlerden ötürü süper zekânın gücünün başlıca frenleyicisi olarak bilgi mahrumiyetine yaslanmak küstahlık olur. Gelgelelim bazı güdük bırakma biçimleri (donanım kapasitesinin sınırlanması ve özgül bilişsel yetilerin gelişiminin engellenmesi dahil) gelişim aşamasında yardımcı kontrol yöntemleri bulabilir. Bunlar sözgelimi arabulucu bir yapay zekâda kapasite kazanımı hızını düzenlemek ve testlerin tam gaz çalışan süper zekânın altındaki farklı yeterlilik düzeylerinde denenmesine izin vermek için kullanılabilir.

Tetikleyici

Tetikleyici, sisteme (muhtemelen onun bilgisi olmadan) hata bulma testleri yapan ve tehlikeli bir faaliyet işareti görürse sistemi kapatan bir mekanizmadır. Böylece programcılar boş zamanlarında şüpheli faaliyet kayıtlarını inceleyip, yeniden başlatmadan önce yapay zekânın kod tabanında ya da çevresinde değişiklikler yapabilirler. Eğer problemin yeterince köklü olduğunu tespit ederlerse, tüm projeyi rafa kaldırmaya karar verebilirler.

Tetikleyiciler teşvik yöntemlerinden farklıdır, çünkü sistem yasaklı faaliyete girmenin sonuçlarının farkında değildir. Tetikleyiciler güdük bırakma yöntemleriyle daha

fazla ortak yöne sahiptir. Tıpkı güdük bırakma gibi, tetikleyici de geçici bir koruyucu olarak kullanılabilir ve gelişim aşaması sırasında bir koruma sağlar. İlkesel açıdan, tetikleyiciler operasyon aşamasında da kullanılabilir, özellikle de bir kutulanmış sistem için. Gelgelelim tetikleyicilerin tam teşekküllü bir süper zekâyı kısıtlayabilmesi son derece şüpheli olacaktır, zira bizim böyle bir aracının insan zekâsının bulduğu bir tetikleyicinin altını oymanın yollarını bulamayacağına ikna olmamız zor olacaktır.

Ama tetikleyici gelişim aşamasında ek bir güvenlik önlemi olarak önemlidir. Yapay zekâ tasarımcıları için sistemin nasıl işlemesi gerektiğine dair beklentilerini kodlamak ve ardından yapay zekâyı bu beklentilerden herhangi biri ihlal edildiği takdirde harekete geçirilecek olan tetikleyicilerle çevrelemek iyi bir pratik olacaktır. Testleri daha katı hale getirmek için, "sanal sunucular", işlemeyen yapay zekâya birtakım kolay gözlemlenebilir ihlaller yaptırmak üzere albeniler yaratacak şekilde stratejik olarak yerleştirilebilir. Sözgelimi eğer bir yapay zekâ internete erişmeyi istemediği varsayılan bir tarzda tasarlanırsa, sırf yapay zekânın kullanıp kullanmayacağını görmek için (otomatik bir kapama düğmesine götüren) sahte bir Ethernet portu kurulabilir. (Bazı tetikleyici örnekleri **TABLO 9**'da sunulmuştur.)

TABLO 9 Farklı tetikleyici örnekleri

Gözetleme türü	Tarif ve örnekler
Davranış	Kutulanmış bir yapay zekânın etrafına sınırlamayı ihlal etme girişimlerini tespit edecek detektörler yerleştirilebilir. Örneğin telsiz haberleşme girişimlerini ya da tahditli bölgede olması amaçlanan iç bilgi işlem kaynaklarına erişim girişimlerini kesebilir. Bir "Ethernet cennet portu" kurulabilir: Görünüşteki internet bağlantısı gerçekte kapama düğmesine götürecektir.
Yetenek	Otomatik yeterlilik testi yapay zekânın çeşitli alanlardaki yeteneğini belirlemek üzere sık aralıklarla gerçekleştirilebilir. Eğer ilerleme hızı beklenmedik derecede yüksekse ya da yapay zekâ kendisini potansiyel tehlike alanına yaklaştıran bir yeterlilik düzeyine erişirse, programcıların başka gelişmeleri daha yakından izlemesine izin vermek için sistem yavaşlatılabilir ya da durdurulabilir.
İçerik	yapay zekânın iç süreçleri şüpheli faaliyetleri tespit etmek için gözetlenebilir. Örneğin yapay zekânın nihai değerlerine dair herhangi bir değişiklik otomatik kapamaya ve gözden geçirmeye yol açabilir. Yapay zekânın kaynak kullanımı ve algoritma yöntemleri de beklenmedik örüntüler için izlenebilir.

Daha hırslı bir adım olarak, yapay zekânın inançları, planları, alt hedefleri ve gerekçeleri sürekli denetlenebilir ve programcıların beklentilerine uyup uymadığı test edilebilir. Yapay zekânın niyetsel durumlarının bu şekilde gözetlenmesi bilhassa güçlü bir içerik izleme biçimi sağlayacaktır: Yapay zekânın bir aldatma tasavvur edeceği

bilişsel süreçlerin (dost olmayan bir yapay zekânın gerçek niyetlerini saklama niyetini ilk belli ettiği o zayıf an) gözetlenmesi.[17]

Yapay zekânın niyetsel durumlarının programcılar ya da otomatik bir izleme mekanizması için şeffaf olmasını gerektiren içerik görüntülemesi her türlü yapay zekâ mimarisi için geçerli olmayabilir. (Sözgelimi bazı nöral ağlar saydamsızdır ve bilgiyi bütüncül olarak ve insan kavramlarıyla her zaman uyuşmayacak şekillerde temsil ederler.) Bu tür mimariler kullanmaktan kaçınmak için bu bir neden olabilir.

Şunu vurgulamak gerekiyor, bir tetikleyicinin değeri sadece mekanizmanın kendisine değil, aynı zamanda –hayati bir şekilde– bir projenin bir tetikleyici tetiklendiğinde nasıl tepki verdiğine de bağlıdır. Eğer programcılar ya da proje yöneticileri ilerleme kaydetmek için sabırsızlanır ve sistemi yeniden açarlarsa ya da bir sonraki çalıştırmada tetikleyicinin tetiklenmesini engellemek için göstermelik birkaç ufak değişikliğin ardından bunu yaparlarsa, o zaman bizatihi tetikleyici tam da niyetlenildiği gibi çalışsa bile hiçbir güvenlik elde edilemez.

Güdülenme seçilimi yöntemleri

Güdülenme seçilimi yöntemleri süper zekânın ne yapmak istediğini şekillendirerek istenmeyen sonuçları savuşturmaya çalışır. Aracının güdülenme sistemini ve nihai hedeflerini belirleyen bu yöntemler belirleyici bir stratejik üstünlüğü zarar verecek şekilde istismar etmek *istemeyecek* bir süper zekâ yaratacaktır. Bir süper zeki aracı amaçlarına ulaşma konusunda yetenekli olduğundan, eğer zarar (kelimenin uygun bir anlamında "zarar") vermemeyi tercih ederse, o zaman zarar (bu anlamıyla "zarar") vermeme eğiliminde olacaktır.

Güdülenme seçilimi bir hedefin ya da uyulacak kuralların açıkça formüle edilmesini (*doğrudan saptama)* ya da sistemin zımnen veya dolaylı olarak formüle edilmiş bir kıstasa atıfla kendisine uygun birtakım değerler keşfedebileceği şekilde sistemi kurmayı (dolaylı normatiflik) içerebilir. Güdülenme seçilimindeki seçeneklerden biri sistemi mütevazı, hırslı olmayan hedefler seçecek (evcimenlik) şekilde inşa etmeye çalışmaktır. Bir güdülenme sistemini en baştan yaratmanın alternatifi şimdiden makul bir güdülenme sistemi olan bir aracı seçmek ve ardından bu aracının bilişsel güçlerini onu süper zeki hale getirecek şekilde artırırken, güdülenme sisteminin bu süreçte (artırma) yozlaşmamasını garanti altına almaktır. Şimdi sırasıyla bunlara bakalım.

Doğrudan saptama

Doğrudan saptama kontrol problemine en doğrudan yaklaşımdır. Bu yaklaşımın iki versiyonu (kural temelli ve sonuççu) vardır ve serbest takılan bir süper zeki yapay zekânın bile güvenli ve yarar sağlayan şekilde hareket etmesine sebep olacak birtakım değerleri ya da kuralları açıkça tanımlamaya çalışmayı içerir. Gelgelelim doğrudan

saptama aşılmaz denebilecek ve hem hangi kuralların ya da değerlerin yapay zekâya kılavuzluk etmesini istediğimizin belirlenmesindeki zorluklardan hem de bu kuralları ya da değerleri bilgisayarca okunabilir kodlarla ifade etmenin zorluklarından kaynaklanan engellerle karşılaşır.

Doğrudan kural temelli yaklaşımın geleneksel örneği, bilimkurgu yazarı Isaac Asimov'un 1942'de yayımlanan bir kısa hikâyesinde formüle ettiği "üç robot yasası" kavramıdır.[18] Bu üç yasa şunlardır: 1) Bir robot bir insana zarar veremez ya da atıl kalarak zarar görmesine müsaade edemez. 2) Bir robot, birinci yasayla çelişmediği sürece bir insanın verdiği bütün emirlere uymak zorundadır. 3) Bir robot, birinci veya ikinci yasayla çelişmediği sürece kendi varlığını korumak zorundadır. İnsan türü için utanç verici olsa da, elli yılı aşkın süredir Isaac Asimov'un yasalarının ötesine geçilemedi: Üstelik bu yaklaşım, bazıları Asimov'un kendi yazılarında da irdelenen bariz sorunlar içerir (Asimov muhtemelen yasaları öncelikle tam da ilgi çekici şekillerde çuvallaması ve hikâyelerinin olay örgüsü için bereketli giriftlikler doğurması için formüle etmişti).[19]

Yıllarını matematiğin temelleri üzerine çalışarak geçirmiş olan Bertrand Russell, bir keresinde, "Sen kesin hale getirmeye çalışana kadar her şey senin fark etmediğin derecede muğlaktır," demişti.[20] Russell'ın düsturu doğrudan saptama yaklaşımı için fazlasıyla geçerlidir. Örneğin Asimov'un ilk yasasını nasıl açımlayacağımızı ele alalım. Robotun bir insana zarar gelme olasılığını asgariye çekmesi gerektiği anlamına mı gelir? Bu durumda diğer yasalar işe yaramaz hale gelir, zira yapay zekânın en azından bir insana zarar gelme olasılığı üzerinden çok ufak bir etkide bulunacak bir eylemde bulunması her zaman mümkündür. Nasıl oluyor da robot birkaç insana zarar gelmesi gibi büyük bir riski birçok insana zarar gelmesi gibi küçük bir riskle dengeleyebiliyor? İş oraya geldiyse, "zarar" nedir? Bir kefeye fiziksel acının verdiği zararı, diğer kefeye mimari çirkinlik ya da toplumsal adaletsizlik acısını koysak ne olur? Bir sadistin kurbanına işkence etmesine izin vermemek ona zarar vermek midir? İş oraya geldiyse, "insan" nedir, nasıl tanımlanır? Ahlaken kayda değer diğer varlıklar (mesela insan olmayan sezgi sahibi hayvanlar ve dijital zihinler) niye dikkate alınmıyor? İnsan düşündükçe sorular da çığ gibi artıyor.

Belki de şu an dünyanın genelinde faaliyet gösteren bir süper zekânın eylemlerini yönetebilecek bir grup kurala en yakın örnek bir hukuk sistemi olacaktır. Ama hukuk sistemleri çok uzun süren deneme yanılma süreçleriyle geliştirilmiştir ve görece yavaş değişen beşeri toplumları düzenlerler. Yasalar gerek duyulduğunda gözden geçirilebilir. En önemlisi, hukuk sistemleri yasa koyucular tarafından istenmediği yeterince bariz olan ve niyet edilmeyen mantıksal açıdan mümkün hukuki yorumları göz ardı edecek aklıselime ve izana sahip olan hâkimler ve jüriler tarafından idare edilir. Son derece karmaşık olan birtakım ayrıntılı kurallar formüle etmek, bunları son derece

farklı koşullara uygulamak ve ilk uygulamada eksik gedikleri düzeltmek muhtemelen insani açıdan imkânsız olacaktır.[21]

Doğrudan sonuççu yaklaşımın önündeki problemler de doğrudan kural temelli yaklaşımınkilere benzerdir. Yapay zekânın klasik faydacılığın bir versiyonunu hayata geçirmek gibi görünüşte basit bir amaca hizmet etmesi amaçlansa bile bu doğrudur. Sözgelimi, "dünyada acıya karşı haz dengesi beklentisini azamiye çıkarın" hedefi basit görünebilir. Ama bunu bir bilgisayar koduyla ifade etmek, her şey bir tarafa, haz ve acının nasıl tanınacağını belirtmeyi gerektirecektir. Bunu güvenilir şekilde yapmak zihin felsefesindeki köklü problemleri çözmeyi gerektirebilir, sırf doğal dilde ifade edilen doğru bir izaha (ardından bir şekilde bir programlama diline çevrilmesi gerekecek olan bir izaha) ulaşmak için bile olsa.

Felsefi izahtaki ya da onun koda çevrilmesindeki küçük bir hata, feci sonuçlar doğurabilir. Nihai hedefi hedonizm olan ve bu nedenle evreni "hedonyum" (haz alınabilir deneyimin yaratılması için optimum olan bir konfigürasyonla örgütlenmiş madde) ile doldurmak isteyen bir yapay zekâ düşünün. Bu amaca istinaden, yapay zekâ komputronyum (bilgi işlem için optimum olan bir konfigürasyonla örgütlenmiş madde) üretip bunu vecd halindeki dijital zihinlere uygulamak için kullanabilir. Yapay zekâ verimliliği azamiye çıkarmak için bu uygulamadan haz deneyimi için asli olamayan tüm asli melekeleri çıkartır ve kendi haz tanımına göre haz yaratılmasına zarar vermeyen her türlü bilgisayar kısayolunu sömürebilir. Sözgelimi yapay zekâ simülasyonunu ödül devreleriyle sınırlandırabilir ve bellek, duyu algısı, yürütme işlevi ve dil gibi yetileri atlayabilir; zihinleri görece kaba işlevsellik düzeyinde simüle edebilir ve alt düzey nöron süreçlerini atlayabilir veya çoklu zihinlerin altta yatan bilgi işlem makinelerinin çoğu parçasını (felsefi tabirle, onların "bağlılık temelleri"ni) paylaştığı bir düzenleme yapabilir. Bu tür dalavereler belirli miktarda kaynakla üretilebilen haz miktarını epey artırabilir. Bunun ne kadar arzu edilir olacağı muğlâktır. Ayrıca eğer yapay zekânın bir fiziksel sürecin haz yaratıp yaratmayacağını belirleme kıstası yanlışsa, o zaman yapay zekânın optimizasyonları sapla samanı birbirine karıştırabilir: Yapay zekânın kıstasına göre asli olmayan ama bizim insani değerlerimizde zımnen var olan kıstaslara göre asli olan bir şeyi atabilir. O zaman, evren bol miktarda hedonyumla değil, bilinçsiz ve tamamen değersiz bilgi işlem süreçleriyle dolar: Trilyonlarca fotokopisi çekilmiş bir gülen yüz etiketinin tüm galaksilere yapıştırılmasına denk bir şeydir bu.

Evcimenlik

Doğrudan saptamaya yukarıda sıralanan örneklerden daha açık olabilecek özel türde bir nihai hedef, kendini sınırlama hedefidir. Bir süper zekânın *genel olarak* dünyada nasıl davranmak isteyeceğini saptamak son derece zor görünmekle birlikte –zira bunun için, doğabilecek bütün durumlardaki bütün ödünleşimleri açıklamamız gerekir– bir süper

zekânın tikel bir durumda nasıl davranacağını saptamak ön açıcı olabilir. Dolayısıyla sistemi küçük bir ölçekte, dar bir bağlamda ve sınırlı sayıda eylem tarzıyla hareket etmesini sağlayacak şekilde sınırlanmaya güdülemeye çalışabiliriz. Yapay zekâya heveslerinin ve faaliyetlerinin kapsamını sınırlamayı amaçlayan bu nihai hedefler koyma yaklaşımına "evcimenlik" adını vereceğiz.

Örneğin soru cevaplayıcı bir araç olarak işlev gösterecek bir yapay zekâ (bir sonraki bölümün terminolojisiyle söylersek, bir "kahin") tasarlamaya çalışabilirsiniz. Yapay zekâya kendisine sorulan her soruya azami derecede doğru cevaplar üretme nihai hedefini koymakla yetinmek güvenli olmayacaktır; 8. Bölüm'de tarif edilen "Riemann hipotezi felaketi"ni hatırlayın. (Ayrıca bu hedefin yapay zekâyı kendisine kolay sorular sorulmasını sağlamak için önlemler almaya teşvik edeceğini de bir düşünün derim.) Evcimenlik sağlamak için bir şekilde bu güçlüklerin üstesinden gelecek bir nihai hedef tanımlamaya çalışılabilir: Belki de soruları doğru cevaplama ve yapay zekânın dünyaya etkisini sorulan sorulara doğru ve manipülatif olmayan cevaplar vermenin tesadüfi bir sonucu olarak doğacak bütün etkiler haricinde asgariye çekme amaçlarını birleştiren bir hedef.[22]

Bu tür bir evcimenlik hedefinin doğrudan saptanması daha hevesli bir hedefin ya da ucu açık durumlarda işlev göstermesi için bir kurallar manzumesinin doğrudan saptanmasından muhtemelen daha uygulanabilir olacaktır. Yine de ciddi zorluklar olacaktır. Sözgelimi yapay zekânın dünyaya etkisi ölçüsünün bizim büyük ya da küçük etki standartlarımızla örtüştüğünden emin olmak için, yapay zekânın "dünyaya etkisini asgariye çekmesi" için neler gerektiğinin tanımına dikkat edilmesi gerekecektir. Kötü bir önlem kötü ödünleşimler doğuracaktır. Ayrıca bir kâhin yaratmayla bağlantılı başka türde riskler vardır ki, bunu da ileride ele alacağız.

Evcimenlik yaklaşımı ile fiziksel kapatma arasında doğal bir uyum vardır. Bir yapay zekâyı sistemin kaçmayı beceremeyeceği şekilde "kutulamaya" çalışırken aynı zamanda yapay zekânın güdülenme sistemini bir yol bulsa bile kaçmak *istemeyeceği* tarzda şekillendirmeye çalışacaktır. Diğer şeyler eşit olduğunda, birden çok bağımsız güvenlik mekanizmasının varlığı başarı şansını azaltacaktır.[23]

Dolaylı normatiflik

Eğer doğrudan saptama umutsuz vaka olarak görünürse, bunun yerine dolaylı normatifliği deneyebiliriz. Burada temel fikir şudur: Somut bir normatif standardı doğrudan saptamak yerine, bir standart belirlemek için bir süreç saptayabiliriz. Ardından sistemi bu süreci hayata geçirmeye ve süreç hangi standarda varıyorsa onu benimsemeye güdüleneceği şekilde inşa edebiliriz.[24] Örneğin bu süreç bizim uygun derecede idealize edilmiş bir versiyonumuzun yapay zekânın ne yapmasını tercih edeceği şeklindeki ampirik soruya dair bir araştırma yürütmek olabilir. Bu örnekte yapay zekâya verilen

nihai hedef, "Eğer meseleye uzun uzadıya kafa yorsaydık yapay zekânın ne yapmasını isterdiysek, onu yapsın" çizgisinde bir şey olabilir.

Dolaylı normatifliğin bir diğer açıklaması için 13. Bölüm'ü beklememiz gerekecek. Orada, "isteğimizin dışkestirimi"[25] fikrini yeniden ele alacağız ve çeşitli alternatif formülasyonları irdeleyeceğiz. Dolaylı normatiflik, güdülenme seçilimi için çok önemli bir yaklaşımdır. Vaadi bizi uygun bir nihai hedefin doğrudan saptanmasını hayata geçirmek için gereken zorlu bilişsel çalışmanın çoğunu süper zekâya yükleme imkânı verebilir.

Artırım

Listemizdeki son güdülenme seçilimi yöntemi de artırımdır. Burada fikir, en baştan bir güdülenme sistemi tasarlamaya çalışmaktan ziyade, zaten kabul edilebilir bir güdülenme sistemi olan bir sistemle başlayıp onun bilişsel yetilerini süper zeki hale getirecek şekilde artırmaktır. Eğer her şey yolunda giderse, güdülenme sistemi kabul edilebilir olan bir süper zekâ elde ederiz.

Bu yaklaşım, çok açık ki, yeni yaratılmış bir çekirdek yapay zekâ örneğinde mevcut değildir. Ama artırım diğer süper zekâ yolları için potansiyel bir güdülenme seçilimi yöntemidir ve buna sistemi zaten insan değerinin bir temsilini içeren normatif bir çekirdekten (düzenli insanlar) kurma olanağının bulunduğu beyin emülasyonu, biyolojik geliştirme, beyin-bilgisayar arayüzleri ve ağlar ile örgütler de dahildir.

Artırım, kontrol problemine diğer yaklaşımlarımızdaki umutsuzluğumuza oranla daha çekici hale gelebilir. Bir çekirdek yapay zekâ için sistem olgun bir süper zekâ haline gelirken bile yinelemeli kendini geliştirme altında güvenilir derecede güvenli ve yararlı kalan bir güdülenme sistemi yaratmak, olmayacak duaya amin demektir, özellikle de çözümü daha ilk girişimde doğru hale getirmemiz gerekiyorsa. Artırımla birlikte en azından tanıdık ve insansı güdülenmeleri olan bir sistemle başlarız.

İşin kötü yanıysa, karmaşık, evrilmiş, yama işi ve kötü anlaşılmış bir güdülenme sisteminin (bir insanınki gibi) bilişsel motoru stratosfere karıştığında yozlaşmayacağından emin olmak zor olabilir. Daha önce belirtildiği üzere, entelektüel işleyişi muhafaza eden kusurlu bir beyin emülasyonu prosedürü kişiliğin tüm veçhelerini muhafaza etmeyebilir. Aynısı hem güdülenmeyi incelikli bir şekilde etkileyebilecek biliş yetisinin biyolojik artırımları için (belki daha az olmak kaydıyla) hem de toplumsal dinamikleri kötü etkileyebilecek (örneğin kolektifin dışarlıklılara karşı ya da kendi bileşenlerine karşı tutumunu alçaltacak şekillerde) örgütlenmelerin ve ağların kolektif zekâ artırımları için de geçerlidir. Eğer süper zekâya bu yollardan herhangi biri üzerinden varılırsa, bir proje sponsoru olgun sistemin nihai güdülenmeleri hakkında garantiler bulmakta zorlanacaktır. Matematiksel açıdan iyi saptanmış ve temel açıdan zarif bir yapay zekâ mimarisi –tüm o insan merkezci olmayan ötekiliğine karşın– çok

daha fazla şeffaflık sunabilir, hatta işlevselliğinin önemli veçhelerinin formel olarak doğrulanması olasılığı bile vardır.

Sonunda, artırımın yararlarını ve zararlarını toplayıp çıkardığımızda, buna güven olup olmayacağı kararını almak zorlama olabilir. Eğer süper zekâya ilkin yapay zekâ yoluyla ulaşılırsa, artırım uygulanabilir değildir. Halbuki eğer süper zekâya ilkin yapay zekâ olmayan bir yoldan ulaşılırsa, o zaman diğer güdülenme seçilimi yöntemlerinden birçoğu uygulanabilir olmayacaktır. Bu durumda bile, artırımın başarıya ulaşmasının ne kadar muhtemel olduğuna dair görüşler, hangi teknolojinin süper zekâyı daha önce üreteceğini etkilemeye fırsatlarımız olduğu sürece, gerçekten de stratejik bir önem taşıyacaktır.

Özet

Bu bölümü bitirmeden önce hızlı bir özet yararlı olabilir. Yapay zekâ güvenliğinin kalbinde yer alan aracılık problemini ele alırken iki genel yöntemi birbirinden ayırdık: Yeterlilik kontrolü ve güdülenme seçilimi. **TABLO 10**'da bir özet sunuluyor.

İki kontrol yöntemi de potansiyel açıdan zayıf yönlere sahiptir ve uygulamada farklı derecelerde zorluklar çıkartır. Belki bunları iyiden kötüye sıralamamız ve ardından en iyi yöntemi seçmemiz gerektiği düşünülebilir. Ama bu çok basit kaçardı. Bazı yöntemler birlikte kullanılabilirken bazıları ise birbirini dışlar. Görece güvensiz bir yöntem bile kolaylıkla ilave olarak kullanılabilirse tavsiye edilebilir, oysa güçlü bir yöntem diğer arzu edilen koruyucuların kullanımını dışlarsa çekici olmayabilir.

Dolayısıyla hangi paket tekliflerin mevcut olduğunu düşünmek zorunludur. Ne tür bir sistem inşa etmeye çalışabileceğimizi ve her biri için hangi kontrol yöntemlerinin uygulanabilir olacağını düşünmek zorunludur.

TABLO 10: Kontrol yöntemleri

Yeterlilik kontrolü	
Kutulama yöntemleri	Sistem dış dünyayı ancak kısıtlanmış, önceden onaylanmış bir kanal yoluyla etkileyebileceği tarzda sınırlanmıştır. Fiziksel ve bilgisel kısıtlama yöntemlerini kapsar.
Teşvik yöntemleri	Sistem uygun teşvikler sağlayan bir çevreye yerleştirilir. Benzer derecede güçlü varlıklar dünyasıyla toplumsal bütünleşmeyi de kapsayabilir. Bir diğer çeşitlilik ise (kriptografik) sembolik ödüllerden yararlanmaktır. "İnsani zapturapt" da çok güçlü bir olanaktır, ama ezoterik etkenler içerir.
Güdük bırakma	Kısıtlamalar sistemin bilişsel yeterliliklerine ya da kilit iç süreçleri etkileme yetisine dayatılmıştır.
Tetikleyici	Sisteme hata bulma testleri yapılır (muhtemelen onun bilgisi olmadan) ve tehlikeli bir faaliyet tespit edilirse, sistem bir mekanizma tarafından kapatılır.
Güdülenme seçilimi	
Doğrudan saptama	Sisteme doğrudan saptanmış bir güdülenme sistemi bahsedilmiştir ve bu sistem sonuççu olabileceği gibi birtakım kurallara uymayı da içerebilir.
Evcimenlik	Bir güdülenme sistemi aracının heveslerini ve faaliyetlerini ciddi derecede sınırlayacak şekilde tasarlanmıştır.
Dolaylı normatiflik	Dolaylı normatiflik kural temelli ya da sonuççu ilkeler içerebilir, ama takip edilecek kuralların ya da peşine düşülecek değerlerin saptanmasına dolaylı bir yaklaşıma yaslanılması bakımından diğerlerinden ayrılır.
Artırım	Şimdiden kayda değer derecede insani ya da hayırhah güdüleri olan bir sistemle yola çıkılır ve süper zeki hale gelmesi için bilişsel yeterlilikleri artırılır.

Kâhinler, Cinler, Hükümdarlar, Aletler

Bazıları der ki, "Sadece soruları cevaplayan bir sistem inşa edin!" Ya da, "Sadece, bir aracıdan ziyade bir alete benzeyen bir yapay zekâ icat edin!" Ama bu öneriler güvenlik kaygılarını dindirmez ve aslında hangi tür sistemin güvenlik için en iyi olasılıkları sunduğu hiç de entipüften bir soru değildir. Biz burada dört tür "kast"ı ele alacağız: kâhinler, cinler, hükümdarlar ve aletler. Ardından bunların birbirleriyle ilişkisini açıklayacağız.[1] Kontrol problemine çözüm arayışında her birinin farklı açılardan olumlu ve olumsuz yanları vardır.

Kâhin

Kâhin, soru cevaplayan bir sistemdir. Doğal bir dille sorulan soruları kabul edip cevaplarını metin olarak sunabilir. Sadece evet/hayır sorularını kabul eden bir kâhin en iyi tahminini güven derecesini temsil eden tek bir bitle ya da fazladan birkaç bitle ortaya koyabilir. Ucu açık soruları kabul eden bir kâhin ise mümkün olan doğru cevapları bilgilendiriciliği ya da uygunluğu bakımından sıralamak için bir ölçüye ihtiyaç duyacaktır.[2] Her iki durumda da doğal dille sorulmuş soruları cevaplama noktasında, alanın genelinde yeteneğe sahip bir kâhin tamamen yapay zekâya özgü bir problemdir. Eğer bu yapılabilirse, insanın kelimeleri dışında niyetlerini de anlama konusunda gerçek bir yeteneğe sahip bir yapay zekâ yaratmak sanırım mümkün olabilir.

Belli bir alanla sınırlı süper zekâ biçimlerine sahip kâhin türleri de tasavvur edilebilir. Sözgelimi sadece formel dille sorulmuş soruları kabul eden, ama bu tür soruları cevaplamakta çok iyi olan (örneğin beşeri matematik mesleğinin yüzyıldır ortaklaşa çabasıyla çözebildiği formel olarak ifade edilmiş her türlü matematik problemini ânında çözebilen) bir matematik kâhini tasavvur edilebilir. Bu tür bir matematik kâhini alanın geneline hâkim bir süper zekâ için bir sıçrama tahtası olacaktır.

Son derece sınırlı alanlarda süper zekâya sahip kâhinler şimdiden mevcuttur. Bir cep hesap makinesi temel aritmetik soruları için çok dar bir kâhin olarak görülebilir; bir internet arama motoru ise insanın genel bildirimsel ilgisinin kayda değer bir parçasını kapsayan bir alanda çok kısmi bir kâhin örneği olarak görülebilir. Bu alanla sınırlı kâhin türleri aracıdan ziyade alettir (alet yapay zekâları birazdan daha kapsamlı

ele alacağız). Aşağıdaysa "kâhin" terimi, aksi belirtilmedikçe, alanın genelinde süper zekâya sahip olan soru cevaplayıcı sistemleri ifade edecektir.

Bir genel süper zekânın kâhin olarak işlemesini sağlamak için hem güdülenme seçilimini hem de yeterlilik kontrolünü uygulayabiliriz. Bir kâhin için güdülenme seçilimi diğer süper zekâ kastları için olduğundan daha kolay olabilir, çünkü bir kâhinde nihai hedef görece basit olabilir. Kâhinin yanlış yönlendirmeyen, doğru cevaplar vermesini ve diğer açılardan dünyaya etkisini sınırlamasını isteriz. Bir evcimenlik yöntemi uygulayarak, kâhinin soruyu cevaplamak için sadece saptanmış kaynaklarını kullanmasını şart koşabiliriz. Örneğin cevabını, önceden yüklenmiş bilgi külliyatına (internetin kayıtlı bir anlık görüntüsüne) dayandırmasını ve hesaplama adımlarının sayısının sabit olmasını şart koşabiliriz.[3] Kâhinin bizi tongaya düşürüp ona kolay sorular sormaya özendirmekten sakınmak için (ona soracağımız bütün sorularda cevapların azami derecede doğru olması gibi bir hedef belirlersek, olacak olan budur), önüne sadece tek bir soruyu cevaplama ve cevabını verdikten hemen sonra işlevini sonlandırma hedefini koyabiliriz. Soru daha program çalıştırılmadan önce belleğine yüklenmiş olacaktır. İkinci bir soru sormak için makineyi yeniden başlatıp, bellekteki önceden yüklenmiş farklı bir soruyla aynı programı çalıştırmamız gerekecektir.

Bir kâhini çalıştırmak için gerekli görece basit güdülenme sistemini saptarken bile çetrefilli ve potansiyel açıdan zararlı zorluklarla karşılaşırız. Örneğin yapay zekâ için, "dünyaya etkisini belli sonuçların elde edilmesine tabi olacak şekilde asgariye çekmek" ya da "cevabı hazırlarken sadece saptanmış kaynaklarını kullanmak" şartlarının ne anlama geldiğine dair belli bir açıklama sunmayı başardığımızı varsayalım. Eğer yapay zekâ entelektüel gelişimi sırasında temel ontolojisinde bir değişikliği içeren bir tür bilim devrimi yaşarsa ne olur?[4] "Etki" ve "saptanmış kaynaklar" derken ne kastettiğimizi kendi ontolojimizi kullanarak ilk başta açıklamış olabiliriz (bilgisayarlar gibi çeşitli fiziksel nesnelerin varlığını koyutlayarak). Ama nasıl ki önceki çağlarda bilimcilerin ön kabul aldığı ontolojik kategorileri ("flojiston", "yaşam atılımı" ve "mutlak eşanlılık" gibi) terk ettiysek, aynı şekilde bir süper zeki yapay zekâ da mevcut kategorilerimizden bazılarının temel nitelikte yanlış anlayışlara dayandığını keşfedebilir. Ontolojik kriz yaşayan bir yapay zekânın hedef sistemi, özgün hedefinin "ruh"u yeni anahtara aktarılabilecek, karşılık beklemeden devredilebilecek kadar esnek ve güçlü olmalıdır.

Güdülenme seçiliminden yararlanarak bir kâhini güvenli kılmak hiç de kolay bir iş sayılmaz, ama yine de çetrefil bir hedefin peşinde dünyayı dolaşan bir yapay zekâ için aynısını yapmak daha kolay olabilir. İlk süper zekânın bir kâhin olmasını tercih etmeyi destekleyen bir argümandır bu. Önce kâhin yolunun tutulmasını destekleyen bir diğer husus da bir kâhinin yeterlilik kontrolüne daha açık olmasıdır. Tüm standart kutulama yöntemleri burada geçerlidir. Ayrıca doğrudan kâhinlere uygulanabilir olan yöntemler olabilir. Örneğin soruları azami derecede doğru bir tarzda değil, bizi kendi

saklı gündemini öne çıkarmaya alet edecek şekilde cevaplayacak bir kâhin düşünün. Bu tehdidi birazcık püskürtmenin yollarından biri, her biri biraz farklı bir koda ve biraz farklı bilgi tabanına sahip birden çok kâhin yaratmak olabilir. O zaman, basit bir mekanizma farklı kâhin türlerinin verdiği cevapları karşılaştırabilir ve ancak tüm cevaplar uyuşmuyorsa insanlara sunabilir. Alanın genelinde süper zeki bir kâhinin diğer kâhinlerin varlığından haberdar olacağını varsaymak gerekse de (bunu insan toplumuna dair bilgisinden çıkarabilir), her bir kâhinin bazı uygulama ayrıntılarını başkalarından saklamak yararlı olabilir. Eğer kâhinler birbirleriyle iletişim kuramıyorlarsa, o zaman sorularımızı manipüle edecek şekilde verecekleri cevapları eşgüdümlü hale getirmeleri zor olabilir. Doğrudan saptırmanın birçok yolu vardır ve kâhinler bu sapmaların hangisinin en çekici olduğu konusunda anlaşamayabilirler; oysa doğrunun kendisi bir Schelling noktasıdır (iletişim yokluğunda anlaşma için belirgin bir yerdir). Dolayısıyla eğer kâhinler mutabakata varırlarsa, doğru cevabı verdiklerinin bir işareti olabilir bu.[5]

İdeal bir kâhin, verdiği cevapların her zaman elinden gelenin en iyisi ölçüsünde doğru olduğunu gönül rahatlığıyla varsaymamız anlamında güvenilir olacaktır. Ama güvenilmez bir kâhin bile yararlı olabilir. Bu tür bir kâhine cevabı bulmanın zor olduğu, ama bir cevabın doğru olup olmadığını doğrulamanın kolay olduğu türde sorular sorabiliriz. Birçok matematik problemi bu kapsama girer. Eğer bir matematik önermesinin doğru olup olmadığını merak ediyorsak, kâhinden önermeyi kanıtlamasını ya da çürütmesini isteyebiliriz. Kanıtı bulmak bizim çapımızın ötesinde içgörü ve yaratıcılık gerektirebilir, ama bir kanıtın geçerliliğini kontrol etmek basit bir mekanik işlemle yapılabilir.

Eğer cevapların sağlamasının yapılması masraflıysa (mantık ve matematik harici konularda genellikle böyledir), doğrulama için kâhinin cevaplar altkümesinden gelişigüzel bir cevap seçebiliriz. Eğer hepsi doğruysa, diğer cevapların çoğuna da doğru olduklarına dair yüksek bir olasılık atfedebiliriz. Bu numara bize tekil olarak doğrulaması masraflı olacak olan güvenilir cevaplarda ciddi bir tenzilat sağlayabilir. (Ne yazık ki, bizim sağlamasını yapmayı *başaramadığımız* güvenilir cevapları veremez, zira gerçeği gizleyen bir kâhinin sadece cevaplarının doğruluğunu teyit edebileceğimize inandığı soruları doğru cevaplamayı tercih etmesi mümkündür.)

Kökene bilfiil güvensizlik duysak bile, doğru cevap (ya da doğru cevabı bulma yöntemi) yönünde kehanet niteliğinde bir işaretten nemalanabileceğimiz önemli meseleler olabilir. Sözgelimi daha ileri güdülenme yöntemleri geliştirmeye çalışırken doğabilecek çeşitli teknik ya da felsefi problemlere çözüm aranabilir. Eğer güvenli olduğu iddia edilen bir yapay zekâ tasarımı önermişsek, kâhine tasarımdaki kayda değer kusurları saptayıp saptayamayacağını ya da bu tür bir kusuru bize yirmi ya da daha az kelimeyle açıklayıp açıklayamayacağını sorabiliriz. Bu türde sorular değerli bilgileri doğurabilir. Fakat akla yatkın görünen ama incelikli şekilde manipüle eden

mesajlar yoluyla psikolojimize oynaması için güvenilmez kâhine fırsat vermemek adına, *çok fazla* soru sormamamız –ve sorduğumuz sorulara verilen cevaplara dair *çok fazla* ayrıntıya dalmamamız– için ihtiyat ve kısıtlama gerekecektir. Sosyal manipülasyon süper gücü olan bir yapay zekânın bizi kendi isteğine ikna etmesi çok fazla iletişim biti gerektirmeyebilir.

Kâhinin kendisi tam niyet edildiği gibi işlese bile, kötüye kullanılması gibi bir risk vardır. Bu problemin bariz boyutlarından biri, bir kâhin yapay zekânın kullanıcısına belirleyici bir stratejik üstünlük verebilecek muazzam bir güç kaynağı olmasıdır. Bu güç gayri meşru olabilir ve herkesin yararı için kullanılmayabilir. Bir o kadar çetrefilli ve önemli bir boyut da bir kâhin kullanımının bizzat kullanıcı için son derece tehlikeli olabilecek olmasıdır. Benzer kaygılar (hem teknik hem de felsefi meseleleri kapsayan kaygılar) diğer varsayımsal süper zekâ kastları için de doğabilir. Bunları 13. Bölüm'de daha ayrıntılı şekilde ele alacağız. Burada şu kadarını belirtmekle yetinelim, hangi soruların sorulduğunu, hangi sırayla sorulduğunu ve cevapların nasıl verilip yayıldığını belirleyen kurallar fevkalade önemli olabilir. Ayrıca kâhinin kurulumunu, cevap vermesinin yarım yamalak kıstaslara göre feci diye sınıflandırılabilecek sonuçlar doğuracağının tahmin edildiği örneklerde herhangi bir soruyu cevaplamayı kabul etmeyeceği tarzda yapmaya çalışıp çalışmamak gerektiği de düşünülebilir.

Cinler ve hükümdarlar

Cin, komutu yerine getiren bir sistemdir: Yüksek düzey bir komutu alır, yerine getirir, ardından bir sonraki komutu beklemek üzere durur.[6] Hükümdar ise geniş ve muhtemelen çok uzun erimli hedeflerin peşinde dünyada faaliyet göstermek için ucu açık bir yetkisi olan bir sistemdir. Bunlar bir süper zekânın ne olması ve ne yapması gerektiği konularında taban tabana zıt şablonlar gibi görünebilir, ama farklılık ilk başta göründüğü kadar büyük değildir.

Bir cinde, daha baştan, bir kâhinin en çekici özelliği zaten feda edilir: Kutulama yöntemlerinden yararlanma fırsatı. Fiziksel açıdan kapatılmış bir cin, sözgelimi sadece saptanmış bir işlem hacmi (etrafında kalın bir duvar ya da aşılmaya çalıştığı takdirde infilak edecek patlayıcılarla dolu bir bariyer olan bir hacim) içinde eşyalar yapabilecek bir cin yaratmak düşünülebilirse de, çokyönlü manipülatörlerle ve inşaat malzemeleriyle donanmış bir süper zekâya karşı bu tür bir fiziki kapatma yönteminin güvenliğine bel bağlamak zor olacaktır. Bir kâhin için başarılabilecek kadar güvenli bir kapatmayı sağlamak bir şekilde mümkün olsa bile; süper zekâdan teftiş edebileceğimiz ve ardından aynı sonuca bizzat ulaşmak için kullanabileceğimiz bir mavi kopya üretmesini istemeye kıyasla, süper zekâya manipülatörlere doğrudan erişim imkânı sağlamanın bize ne kadar fazla şey kazandıracağı net değildir. İnsan aracıyı es geçerek elde edilecek

hız ve uygunluk, bir kâhini kapatmak için mevcut olan güçlü kutulama yöntemlerini kullanmaktan vazgeçme kaybına hiç de değmez.

Eğer bir cin yaratıyor *olsaydık*, komutun lafzından ziyade arkasındaki niyete riayet etmesini sağlayacak şekilde yaratmak isterdik, zira lafza bakan bir cin (belirleyici bir stratejik üstünlük elde edecek kadar süper zeki bir cin) kullanıcıyı ve insanlığın geri kalanını 8. Bölüm'de kötü huylu başarısızlık tarzlarına ilişkin altbaşlıkta açıklanan nedenlerden ötürü daha ilk kullanımda öldürmeye meyledebilir. Daha geniş ölçekte, cinin komut olarak verilen şeye dair hayırsever ve insanların makul diye görecekleri bir yorum araması ve cinin söz konusu komutu lafza bağlı bir yorumdan ziyade bu tür bir yorumla yerine getirecek şekilde güdülenmesi önemli görünmektedir. İdeal cin otistik bir dâhiden ziyade bir süper kâhya olacaktır.

Gelgelelim bu türde bir süper kâhya doğası bahşedilmiş olan bir cin, hükümdarlar kastında üyeliğe hak kazanmaktan çok uzak olmayacaktır. Karşılaştırma olması adına, nihai hedefi bir hükümdardan ziyade bir cin yaratmış olsaydık vereceğimiz komutların özüne riayet etmek olacak olan bir hükümdar yaratma fikrini ele alalım. Bu tür bir hükümdar bir cini taklit edecektir. Bu hükümdar süper zeki olduğundan, bir cine hangi komutları verirdik onu tahmin edebilecektir (ve bize bu bilginin kararlarımızın şekillenmesine yardımcı olup olmayacağını her zaman sorabilir). Bu durumda, böyle bir hükümdar ile cin arasında gerçekten önemli bir fark olacak mıdır? Veya diğer taraftan bir ayrıma gidersek, süper zeki bir cinin de bizim ona vereceğimiz komutları tahmin edebildiğini varsayalım: Bu durumda, harekete geçmeden önce gerçekten komutun verilmesini beklemekle ne kazanılacaktır?

Bir cinin bir hükümdardan şöyle büyük bir üstünlüğü olduğu düşünülebilir: Bir şeyler kötü giderse, cine durması ya da önceki eylemlerin sonuçlarını tersine çevirmesi için yeni komut verilebilir, oysa bir hükümdar ne kadar çırpınırsak çırpınalım bildiği yolda devam edecektir. Fakat cinin bu görünüşteki güvenlik üstünlüğü büyük oranda yanılsamadır. "Dur" ya da "geri al" düğmesi cinlerde sadece iyicil başarısızlıklarda işler: Kötücül başarısızlık örneğinde (mesela mevcut komutu uygulamanın cin için nihai hedef haline geldiği durumda), cin önceki komutu çiğneyecek girişimleri tamamen kulak arkası edecektir.[7]

Seçeneklerden biri kullanıcıya sunulan komutun muhtemel sonuçlarının belirgin veçheleri hakkında otomatikman tahminlerde bulunacak, ilerlemeden önce teyit alacak bir cin yaratmaya çalışmak olacaktır. Bu tür bir sisteme *önizlemeli cin* denebilir. Ama bunu bir cin için yapmak mümkünse, bir hükümdar için de yapılabilir. Dolayısıyla bir kez daha, bunu cin ile hükümdar arasında net bir farklılaştırıcı olarak görmemek gerekir. (Bir önizleme işlevselliğinin yaratılabildiğini varsayarsak, bunun kullanılıp kullanılmayacağı ve kullanılırsa nasıl kullanılacağı, tersinmez bir gerçeklik haline

getirmeye kesinkes bağlanmadan önce sonuca bakabiliyor olmak çok çekici olmasına karşın, düşünüldüğü kadar bariz değildir. Bu konuya bilahare geri döneceğiz.)

Bir kastın diğerini taklit edebilmesi kahinler için de geçerlidir. Bir cine belli soruları cevaplamaktan başka herhangi bir komut vermiyorsak, bir kâhin gibi davranması sağlanabilir. Akabinde, bir kâhine de belli komutları yerine getirmenin en kolay yolunun ne olduğunu sorarsak o da bir cinin yerini alabilir. Kâhin bize bir cinin üreteceği aynı sonucu elde etmek için adım adım talimatlar verebilir, hatta bir cinin kaynak kodunu bile üretebilir.[8] Bir kâhin ile bir hükümdar arasındaki ilişkiye dair de benzer şeyler söylenebilir.

Dolayısıyla üç kast arasındaki esas fark ortaya çıkaracakları nihai yeterliliklerde saklı olmayabilir. Bunun yerine, farklılık kontrol problemine alternatif yaklaşımlarda düğümlenmektedir. Her bir kast farklı birtakım güvenlik önlemlerine tekabül eder. Bir kâhinin en göze batan özelliği kutulanabilir olmasıdır. Ayrıca bir kâhine evcimenlik güdülenme seçilimi de uygulanmaya çalışılabilir. Bir cini kutulamak zordur, ama en azından evcimenlik uygulanabilir olabilir. Bir hükümdar ise ne kutulanabilir ne de evcimenlik yaklaşımıyla yönetilebilir.

Eğer yegâne ilintili etkenler bunlar olsaydı, o zaman istenirlik sırası açık olurdu: Bir kâhin bir cinden daha güvenli olurdu, cin de bir hükümdardan daha güvenli olurdu. Ayrıca uygunluk ile işleyiş hızı bakımından ilk başta var olan farkların hepsi görece küçük olup bir kâhin yaratmakla elde edilebilecek güvenlik kazanımlarının öne çıktığı bir fark olurdu. Gelgelelim hesaba katılması gereken başka etkenler de vardır. Kastlar arasında seçim yaparken, sadece sistemin kendisinin sunduğu tehlikeyi değil, nasıl kullanılabileceğinden doğacak tehlikeleri de dikkate almak gerekir. Cinin onu kontrol eden kişiye muazzam bir güç verdiği çok açıktır, ama aynısı bir kâhin için de geçerlidir.[9] Halbuki bir hükümdar hiçbir insana ya da gruba sonucu etkileyecek özel bir güç vermeyecek ve ilk baştaki gündemini bozma ya da değiştirme girişimine direnecek şekilde yaratılabilir. Dahası, eğer bir hükümdarın güdüleri "dolaylı normatiflik" kullanarak tanımlanabilirse (13. Bölüm'de ayrıntılarıyla ele alınacak bir kavram), o zaman "azami açıdan adil ve ahlaken doğru olan her neyse" gibi soyut tanımlanmış (ve neler gerektirdiğini kimsenin önceden bilmediği) bir sonuca ulaşmak için kullanılabilir. Buradan Rawls'ın "cehalet örtüsü"ne benzer bir durum doğabilir.[10] Bu tür bir kurulum mutabakatın elde edilmesini kolaylaştırabilir, çatışmaların engellenmesine yardımcı olabilir ve daha hakkaniyetli bir sonucu teşvik edebilir.

Belli kâhin ve cin türlerinin aleyhinde olan bir diğer husus ise bizim nihayetinde elde etmeye çalıştığımız sonuçla tamamen örtüşmeyen bir nihai hedefe sahip bir süper zekâ tasarlamanın barındırdığı risklerdir. Örneğin süper zekânın dünyaya etkilerinden bazılarını asgariye çekmek istemesini sağlamak için evcimenlik güdülenmesinden yararlanırsak, bu suretle olanaklı sonuçlara dair tercih sıralaması sponsorununkinden

farklı olan bir sistem yaratmamız olasıdır. Eğer yapay zekâyı soruların doğru cevaplanmasına ya da tekil komutlara sadakatle riayet edilmesine özgül derecede yüksek bir değer biçecek şekilde yaratırsak da aynısı olacaktır. Şimdi, gerekli özen gösterilirse, buradan bir sorun çıkmaz: Yapay zekâ standartlarına göre iyi olan sonuçların sahibin standardına göre de iyi olması noktasında iki sıralama arasında yeterli mutabakat olacaktır, en azından gerçeğe dönüşme şansı makul derecede olan olanaklı dünyalarla alakalı olarak. Ama belki de yapay zekânın hedefleri ile bizimkiler arasında sınırlı düzeyde ahenksizliğe dahi izin vermenin akıllıca olmadığı savunulabilir. (Aynı kaygı elbette hükümdarlara bizimkiyle tamamen ahenkli olmayan hedefler verilmesi konusunda da geçerli olacaktır.)

Alet yapay zekâlar

Burada sunulan önerilerden biri, süper zekâyı bir aracıdan ziyade bir alete benzer şekilde yaratmaktır.[11] Bu fikir, sayısız uygulamada kullanılan sıradan yazılımın bu kitapta ele alınan zorluklara benzer bir güvenlik kaygısı doğurmadığı gözleminden doğuyor görünmektedir. Bu tür bir yazılıma benzeyen (sözgelimi bir uçuş kontrol sistemi ya da sanal asistan gibi olan), ama biraz daha esnek ve kabiliyetli bir "alet yapay zekâ" yaratılamaz mı? Neden kendi iradesi olan bir süper zekâ yaratalım? Bu düşünce çizgisine göre, aracı paradigma temelden yanlış yönlendirilmiştir. İnançları ve arzuları olan ve yapay bir insan gibi davranan bir yapay zekâ yaratmak yerine, sadece ne programlanmışsa onu yapan düzenli bir yazılım yaratmayı amaçlamamız gerekir.

Gelgelelim "sadece ne programlanmışsa onu yapan" bir yazılım yaratma fikri, eğer yaratılan ürün güçlü bir genel zekâysa o kadar basit olmayabilir. Elbette basit bir anlamda her türlü yazılımın sadece programlanan şeyi yaptığı söylenebilir: Davranış, kod tarafından matematiksel olarak saptanmıştır. Ama bu durum makine zekânın her türlüsü için, "alet yapay zekâ" olsun olmasın, aynı derecede geçerlidir. Eğer bunun yerine, "sadece ne programlanmışsa onu yapan" demek, yazılımın programcıların *niyetlendiği* şekilde davranması demekse, o zaman sıradan yazılımın çoğu zaman erişemediği bir standarttan bahsediyoruz.

Makine süper zekâsıyla kıyaslandığında günümüz yazılımının yeterlilikleri sınırlı olduğundan, bu tür başarısızlıkların sonuçlarıyla –önemsizden çok masraflıya kadar uzanan bir ölçekte– başa çıkılabilir, ama hiçbir koşulda varoluşsal bir tehdide varmayacağını belirtelim.[12] Ne var ki sıradan yazılımı varoluşsal açıdan güvenli kılan yeterli güvenilirlikten ziyade yetersiz kapasiteyse, o zaman bu tür bir yazılımın güvenli bir süper zekâ için nasıl bir model haline gelebileceği muğlaktır. Sıradan yazılımın yerine getirdiği görevlerin kapsamını genişleterek yapay genel zekâ ihtiyacının ortadan kaldırılabileceğini düşünmek mümkündür. Ama bir genel zekânın modern ekonomide kârlı şekilde yerine getirebileceği görevlerin kapsamı ve çeşitliliği muazzamdır. Tüm

bu görevleri yerine getirmek için özel amaçlı yazılım yaratmak hesaplı olmayacaktır. Bu tür bir proje gerçekleştirilebilirse bile, hayata geçirilmesi *uzun* zaman alacaktır. Tamamlanabilmesi için görevlerden bazılarının niteliğinin değişmesi ve yeni görevlerin ilintili hale gelmesi gerekecektir. Yeni görevleri yapmayı, yeni görevleri keşfetmeyi kendi başına öğrenebilecek yazılıma sahip olmak büyük bir üstünlük sağlayacaktır. Ama bunun için yazılımın öğrenebilmesi, muhakeme ve plan yapabilmesi ve bunları alanlar arası tarzda güçlü ve sağlam şekilde yapabilmesi gerekecektir. Başka bir deyişle, genel zekâ gerekecektir.

Amaçlarımız açısından bilhassa ilintili olan ise bizatihi yazılım geliştirme görevidir. Bunu otomatik hale getirebilmek muazzam pratik yararlar sağlayacaktır. Ama hızlı kendini geliştirme kapasitesi tam da bir çekirdek yapay zekânın zekâ patlamasını başlatabilmesini sağlayan hayati özelliktir.

Eğer genel zekâ elzemse, alet yapay zekâyı sıradan bir aletin iç rahatlatıcı derecede edilgen niteliğini muhafaza etmek için yorumlamanın başka bir yolu var mıdır? Aracı olmayan bir genel zekâya sahip olunabilir mi? Sezgisel açıdan, sıradan yazılımı güvenli kılan şey yeterliliğinin sınırlı olması değildir; aynı zamanda hırsının olmamasıdır. Excel'de bir yolunu bulsa dünyayı ele geçirmeyi gizliden gizliye isteyen bir altprogram yoktur. Elektronik çizelge uygulamasının "istediği" hiçbir şey yoktur; sadece programdaki talimatları körü körüne uygular. Aynı türde daha genel olarak zeki bir uygulama yaratmanın önünde ne tür bir engel olduğu merak edilebilir. Bir hedef tarifi önüne konduğunda onu nasıl başaracağına dair tıpkı Excel'in bir sütun sayıya toplamı hesaplayarak yanıt vermesi gibi bir planla yanıt veren –ve dolayısıyla çıktısına dair veya insanların onu nasıl kullanmayı seçebileceğine dair herhangi bir "tercih" ifade etmeyen– bir kâhin mesela?

Yazılım yazmanın klasik yolu, programcının yerine getirilecek görevi kodla ifade edilebilen matematiksel açıdan iyi tanımlanmış adımlar silsilesinden oluşan bir açıktan çözüm süreciyle formüle edecek kadar ayrıntılı anlamasını gerektirir.[13] (Pratikteyse, yazılım mühendisleri yararlı davranışların depolandığı kod kütüphanelerine güvenirler ve davranışların nasıl uygulamaya döküldüğünü anlama ihtiyacı duymadan bundan yararlanabilirler. Ama bu kodu ilk yaratan, neler yaptıklarına dair ayrıntılı bir kavrayışa sahip olan programcılardır.) Bu yaklaşım iyi anlaşılmış görevlerin çözülmesinde işe yarar ve şu an kullanımda olan çoğu yazılımda geçerlidir. Gelgelelim yapılması gereken görevlerin hepsinin tam olarak nasıl çözüleceğini kimse bilmediğinde yetersiz kalır. İşte yapay zekâ alanından tekniklerin işe yarar hale geldiği nokta burasıdır. Dar kapsamlı uygulamalarda, makine öğreniminin çoğunlukla insan tasarımı bir programdaki birkaç parametreye salt ince ayar çekmek için kullanılması mümkündür. Örneğin bir spam filtresi, sınıflandırma algoritmasının çeşitli tanılama özelliklerine bindirdiği ağırlıkları değiştiren bir süreçte elle sınıflandırılmış e-posta mesajları ko-

nusunda eğitilebilir. Daha hırslı bir uygulamadaysa, sınıflandırıcının yeni özellikleri kendi başına keşfedebileceği ve değişen bir ortamda geçerliliklerini test edebileceği şekilde yaratılması mümkündür. Daha da karmaşık bir spam filtresineyse kullanıcının karşılaştığı ödünleşimler hakkında ya da sınıflandırdığı mesajların içerikleri hakkında muhakemede bulunma yeteneği bahşedilebilir. Bu örneklerin hiçbirinde programcının spam ile spam olmayanı ayırmanın en iyi yolunu bilmesine gerek yoktur; öğrenme, keşfetme ya da muhakemede bulunma yoluyla kendi performansını artırabilen bir algoritmanın nasıl kurulacağını bilmesi yeterlidir.

Yapay zekâdaki ilerlemelerle birlikte, programcının verili bir görevin nasıl yerine getirileceğini öğrenmesi için gerekli bilişsel emeğin daha büyük bir kısmını boşaltmak mümkün olacaktır. Uç bir örnekte, programcı sadece neyin başarı olduğuna dair formel bir kıstas belirleyecek ve çözümü bulma işini yapay zekâya bırakacaktır. Yapay zekâ aramasına rehberlik etmesi için güçlü bulgulayıcılardan ve diğer yöntemlerden yararlanarak olanaklı çözümler uzamındaki yapıyı keşfedecektir. Başarı kıstasını karşılayan bir çözüm bulana kadar aramayı sürdürecektir. O zaman, yapay zekâ ya bizatihi çözümü hayata geçirecek ya da (bir kâhin örneğinde) çözümü kullanıcıya raporlayacaktır.

Bu yaklaşımın iptidai biçimleri bugün çok yaygın olarak kullanılmaktadır. Gelgelelim yapay zekâ ve makine öğrenme tekniklerinden yararlanan yazılım, programcıların öngörmediği çözümler bulabilirse de; bütün pratik amaçlar açısından bakıldığında bir alet gibi işlev gösterir ve varoluşsal risk teşkil etmez. Tehlike alanına ancak bu çözüm arayışlarında kullanılan yöntemler aşırı güçlü ve genel hale geldiğinde, yani bunlar genel zekâya denk hale gelmeye başladığında ve özellikle de süper zekâya denk hale gelmeye başladığında gireriz.

Bu noktada, sorun çıkabilecek (en az) iki yer vardır. Birincisi, süper zeki arama süreci hem beklenmedik hem de hiç amaçlanmamış bir çözüm bulabilir. Bu durum öncesinde ele alınan türde başarısızlıklardan birine ("tersten örnekleme", "altyapı bolluğu" ya da "zihin suçu") yol açabilir. Bunun nasıl olacağının en bariz şekilde görüldüğü örnek, bulduğu çözümü doğrudan uygulayan bir hükümdar ya da cin örneğidir. Eğer moleküler gülen yüzler yapmak ya da gezegeni kâğıt kıskaçlarına dönüştürmek süper zekânın keşfettiği çözüm kıstasını karşılayan ilk fikirse, o zaman gülen yüzler ya da kâğıt kıskaçları elde ederiz.[14] Ama eğer geri kalan her şey iyi giderse sadece çözümü *raporlayan* bir kâhin bile tersten örnekleme nedeni olabilir. Kullanıcı kâhinden belli bir sonucu başaracak bir plan ya da belli bir işleve hizmet edecek bir teknoloji talep eder ve kullanıcı planı takip ettiğinde ya da teknolojiyi inşa ettiğinde, tersten örnekleme doğabilir, tıpkı yapay zekâ bizatihi çözümü uygulasaydı olacağı gibi.[15]

Sorun çıkabilecek ikinci yer, yazılımın operasyon sürecidir. Eğer yazılımın bir çözüm aramak için kullandığı yöntemler yeterince karmaşıksa, bizatihi arama sürecini akıllı

bir tarzda yürütmek için hükümler içerebilir. Bu örnekte, yazılımı çalıştıran makine salt bir aletten ibaret olmayıp daha çok bir aracıya benzemeye başlayabilir. Mesela yazılım çözüm arayışını nasıl sürdüreceğine dair bir plan geliştirmekle başlayabilir. Plan ilkin hangi alanların hangi yöntemlerle araştırılacağını, hangi verilerin toplanacağını ve mevcut bilgi işlem kaynaklarından en iyi nasıl yararlanılacağını saptayabilir. Yazılım kendi iç kıstasını (belirlenen zamandaki kullanıcının saptadığı kıstası tatmin eden bir çözüm bulmanın yeterince yüksek bir olasılığını vermek) karşılayan bir plan ararken, alışılmışın dışında bir fikre denk gelebilir. Sözgelimi ek bilgi işlem kaynakları edinme ve potansiyel kesintiye uğratıcıları (insanlar gibi) safdışı etmeyle başlayan bir plan hazırlayabilir. Bu tür "yaratıcı" planlar, yazılımın bilişsel yetenekleri yeterince yüksek bir düzeye çıktığında karşımıza çıkar. Yazılım bu tür bir planı eyleme döktüğünde, varoluşsal bir felaket doğabilir.

KUTU 9'daki örneklerin gösterdiği gibi, ucu açık arama süreçleri mevcut sınırlı biçimlerinde bile bazen insan merkezci olmayan tuhaf ve beklenmedik çözümler sergileyebilir. Günümüzdeki arama süreçleri tehlikeli değildir, zira bir programın dünyayı ele geçirmesini sağlayabilecek türde bir planı keşfedemeyecek kadar zayıftır. Bu tür bir plan son derece zor adımlar gerektirir (mevcut teknolojinin birçok kuşak ötesinde bir yeni silah teknolojisinin icat edilmesi ya da insan kalemşorların geliştirdiği iletişimlerin hepsinden çok daha etkili bir propaganda kampanyasının hayata geçirilmesi gibi). Bir makinenin bu tür fikirleri gerçekten işletecek şekilde geliştirmek şöyle dursun, *tasavvur etmek* için bile bir şansa sahip olması için, muhtemelen dünyayı en azından normal bir yetişkin insanın sahip olduğu dünya modeli kadar zengin ve gerçekçi bir tarzda temsil etme kapasitesine ihtiyacı olacaktır (gerçi bazı alanlardaki farkındalık yokluğu herhalde başka alanlarda fazladan yetenekle telafi edilebilir). Karmaşık planlama problemlerini çözme girişimlerini genellikle kaba kuvvet yöntemleriyle püskürten kombinasyonel patlamadan ötürü (1. Bölüm'de görmüştük), bilinen algoritmaların kusurlarını salt daha fazla bilgi işlem gücü akıtarak aşmak gerçekçi görünmemektedir.[16] Gelgelelim arama ya da planlama süreçleri de yeterince güçlendiğinde, potansiyel açıdan tehlikeli hale gelirler.

Aracı türü amaçlı davranışın güçlü arama süreçlerinin (iç çalışma planları arayan süreçler ve kullanıcının saptadığı bazı kıstaslar için doğrudan çözümler arayan süreçler dahil) uygulanmasından kendiliğinden ve rastgele çıkmasına izin vermek yerine, kasten aracılar yaratması daha iyi olabilir. Bir süper zekâya açıktan aracı benzeri bir yapı kazandırmak kestirilebilirliği ve şeffaflığı artırmanın bir yolu olabilir. İnançları ile değerleri arasında net bir ayrım olacak şekilde yaratılmış olan iyi tasarlanmış bir sistem, bizim doğabilecek sonuçlara dair bir kestirimde bulunmamıza imkân sağlar. Sistemin hangi inançları edineceğini ya da kendisini hangi durumlarda bulacağını tam olarak öngöremesek bile, nihai değerlerini ve dolayısıyla gelecekteki eylemlerini seçme

ve potansiyel bir planı değerlendirmede kullanacağı kıstasları teftiş edebileceğimiz bilinen bir yer olacaktır.

Karşılaştırma

Ele aldığımız farklı sistem kastlarının özelliklerini özetlemek yararlı olabilir (**TABLO 11**).

En güvenli sistemin hangisi olduğunu belirlemek için başka araştırmalara ihtiyaç var. Cevap yapay zekânın kullanılacağı koşullara bağlı olabilir. Kâhin kastının güvenlik bakış açısından çekici olduğu açıktır, zira hem yeterlilik kontrol yöntemlerine hem de uygulanacak güdülenme seçilimi yöntemlerine izin verecektir.

Dolayısıyla sadece hükümdar kastına hükmediyor görünebilir; bu durum (dünyanın diğer güçlü süper zekâlar içerdiğine inanıldığı ve dolayısıyla toplumsal bütünleşme ya da insani zapturaptın uygulanabilir olabileceği senaryolar haricinde) güdülenme seçilimi yöntemlerine izin verecektir. Gelgelelim bir kâhin, yozlaştırılması ya da elindeki gücü akılsızca kullanması mümkün olan operatörünün eline epey bir güç verebilir, oysa bir hükümdar bu tehlikelere karşı belli bir koruma sunacaktır. Dolayısıyla güvenlik sıralamasını belirlemek öyle kolay değildir.

Bir cini kâhin ile hükümdar arasında bir uzlaşma olarak görebiliriz, ama bu iyi bir uzlaşma olacak diye bir kaide yoktur. Birçok açıdan, ikisinin de olumsuz yanlarına sahip olacaktır. Bir alet yapay zekâsının görünüşteki emniyetliliği ise yanılsamalı olabilir. Aletlerin süper zeki aracıları ikame edecek kadar çokyönlü olması için son derece güçlü iç arama ve planlama süreçlerinden yararlanmaları gerekebilir. Aracı tipi davranışlar bu tür süreçlerden planlanmamış bir sonuç olarak doğabilir. Bu durumda sistemi öncelikle bir aracı olacak şekilde tasarlamak daha iyi olabilir, böylece programcılar sistemin çıktısını hangi kıstasların belirleyeceğini daha kolay görebilirler.

KUTU 9: Körü körüne aramanın tuhaf çözümleri

Basit evrimsel arama süreçleri bazen hiç beklenmedik sonuçlar üretebilir ve bu çözümler kullanıcının tanımladığı formel bir kıstası kullanıcının beklediği ya da niyetlendiğinden çok farklı bir tarzda karşılayabilir.

Evrimleşebilir donanım alanı bu olguya dair birçok örnek sunar. Bu alanda, bir evrimsel algoritma donanım tasarımları uzamını araştırır ve her bir tasarımın uygunluğunu hızla yeniden ayarlanabilen bir sıra ya da anakart üzerinde fiziksel olarak somutlaştırarak test eder. Evrilen tasarımlar çoğu zaman kayda değer bir tutumluluk sergiler. Sözgelimi aramalardan biri, saat olmadan işleyen bir frekans ayırma devresi keşfetmişti (bu bileşen normalde bu işlev için zorunlu görülür). Araştırmacılar evrilmiş devrenin büyüklüğünün bir insan mühendisin bu görev için gerek duyacağından bir iki kat daha az olacağını tahmin ediyorlardı. Devre, bileşenlerinin fiziksel özelliklerini alışılmadık şekillerde kullanıyordu; bazı aktif, zorunlu bileşenler girdi ve çıktı pinlerine bağlı bile değildi! Bu bileşenler daha ziyade normalde mide bulandırıcı yan etkiler olarak görülecek elektromanyetik kuplaj ya da güç kaynağı yüklemesi gibi şeyler aracılığıyla iştirak ediyordu.

Bir osilatör yaratma görevi olan bir diğer arama süreci görünüşte daha da vazgeçilmez olan bir bileşenden mahrumdu: Sığaç (kapasitör). Algoritma kendi başarılı çözümünü sunduğunda, araştırmacılar bunu incelemiş ve ilk başlarda, "işe yaramayacağı" sonucuna varmışlardı. Daha dikkatli incelendiğinde, algoritmanın MacGyver gibi, baskılı devre kartlarını laboratuvarda şans eseri yakınlarda olan kişisel bilgisayarların yarattığı sinyalleri yakalayacak bir anten olarak kullanarak sensörsüz anakartını çakma bir telsiz alıcısı haline getirdiğini keşfetmişlerdi. Devre, istenen salınımlı çıktıyı üretmek üzere bu sinyali artırmıştı.*

Diğer deneylerde, evrimsel algoritmalar anakartın bir osiloskop (salınımizler) ile izlenip izlenmediğini ya da el havyasının laboratuvarın merkezi güç kaynağına bağlı olup olmadığını hisseden devreler tasarlamıştı. Bu örnekler bir ucu açık arama sürecinin, geleneksel insan tasarım düşüncesinin kullanmayı beceremeyeceği, hatta geriye baktığında açıklayamayacağı araçlardan yararlanarak tamamen beklenmedik duyusal yeterlilikler geliştirmek için kendisi açısından erişilebilir olan malzemeleri başka bir amaçla kullanmak üzere değiştirebildiğine örnek teşkil eder.

Evrimsel aramanın "aldatma" ya da belli bir amaca ulaşmanın sezgilere aykırı yollarını bulma eğilimi doğada da görülebilir, ama sanırım biyolojinin görünüm ve hissine olan aşinalığımızdan ve dolayısıyla doğadaki evrimsel süreçlerin gerçek sonuçlarını normal olarak görmeye meylettiğimizden (bunları uygulama öncesinde beklemiyor olsak bile) bu o kadar bariz değildir. Ama evrimsel süreçleri alışıldık bağlamının dışında eylem halinde görebileceğimiz yapay zekâ deneyleri yapmak mümkündür. Bu tür deneylerde araştırmacılar doğada nadiren görülen koşullar yaratabilir ve sonuçları gözlemleyebilir.

---

* Bird ve Layzell (2002) ile Thompson (1997); ayrıca Yaeger (1994, 13–14).

Örneğin 1960'lardan önce biyologlar yırtıcı popülasyonların Malthusçu tuzağa düşmesini engellemek için kendi soylarını kısıtladıklarını savunurlardı.* Bireysel seçilim bu tür bir kısıtlamaya engel olacak olsa da, grup seçiliminin üreme fırsatlarını sömürmek üzere bireysel teşvikleri aşacağı ve gruba ya da o popülasyona yarar sağlayacak özellikleri tercih edeceği kimi zaman düşünülürdü. Teorik analiz ve simülasyon çalışmaları sonradan gösterdi ki, grup seçilimi ilkesel olarak mümkün olmakla birlikte, güçlü bireysel seçilimi ancak doğada nadiren görülebilecek çok katı koşullarda aşabilir.† Ama bu tür koşullar laboratuvarda yaratılabilir. Un kurtları (***Tribolium castaneum***) nüfus büyüklüğünü azaltmak için yetiştirildiğinde, evrim güçlü grup seçilimi uygulamak suretiyle gerçekten de küçük popülasyonlara götürebilmiştir.‡ Gelgelelim bunu başarmanın yolları arasında sadece, evrimsel aramayı naif şekilde insan-biçimlileştiren bir insanın bekleyebileceği doğurganlığın azaltılması ve gelişim zamanının uzatılmasının "iyicil" uyarlanımları değil, yamyamlık artışı da vardır.§

---

* Williams (1966).

† Leigh (2010).

‡ Bu örnek Yudkowsky'den (2011) alınmıştır.

§ Wade (1976). Bilgisayar deneyleri biyolojik evrimin yönlerine benzeyecek şekilde tasarlanmış simülasyonlu evrimle de yürütülmüştür – bir kez daha bazen tuhaf sonuçlar doğurarak (örneğin bkz. Yaeger [1994]).

TABLO 11: Farklı sistem kastlarının özellikleri

Kâhin	**Soru cevaplayıcı bir sistem** *Çeşitlilikler:* Alanla sınırlı kâhin türleri (örneğin matematik); çıktıyla sınırlı kâhin türleri (örneğin sadece evet/hayır/belirsiz cevapları ya da olasılıklar); cevaplamanın sonuçlarının önceden saptanmış "felaket kıstasları"nı karşılayacağını kestirdiği takdirde soruları cevaplamayı reddedecek kâhin türleri; emsal değerlendirme için birden çok kâhin.	• Kutulama yöntemleri tamamen geçerlidir. • Evcimenlik tamamen geçerlidir. Yapay zekânın insan niyetlerini ve çıkarlarını (cinlere ve hükümdarlara kıyasla) anlamasına duyulan ihtiyaç daha azdır. • Evet/hayır sorularından yararlanılması cevapların "yararlılığı" ya da "bilgilendiriciliği"nin bir ölçüsüne duyulan ihtiyacı gereksiz hale getirebilir. Büyük güç kaynağı (operatöre belirleyici bir stratejik üstünlük kazandırabilir) • Operatörün aptalca kullanımına karşı sınırlı koruma • Güvenilmez kâhin türleri, bulması zor ama doğrulanması kolay cevaplar üretmek için kullanılabilir • Cevapların zayıf sağlamaları birden çok kâhin kullanarak mümkün hale gelebilir.
Cin	**bir komutu yerine getiren sistem** *Çeşitlilikler*: Komutun lafzına uymaktan ziyade özüne uyma dereceleri (ya da farklı "dışkestirim mesafeleri") kullanan cinler; alanla sınırlı cinler; öndeğerlendirmeli cinler; cevaplamanın sonuçlarının önceden saptanmış "felaket kıstasları"nı karşılayacağını kestirdiği takdirde soruları cevaplamayı reddedecek olan cinler.	• Kutulama yöntemleri kısmen geçerlidir. Evcimenlik kısmen geçerlidir. • Cin beklenen sonuçların bariz veçhelerine dair bir öndeğerlendirme sunabilir. • Cin aşamalarda değişiklik uygulayabilir ve her bir aşamada değerlendirme fırsatı sunabilir. • Büyük güç kaynağı (operatöre belirleyici stratejik üstünlük kazandırabilir). • Operatörün aptalca kullanımına karşı sınırlı koruma Yapay zekânın insan çıkarlarını ve niyetlerini anlamasına duyulan ihtiyaç (kâhin türlerindekine kıyasla) daha fazladır.

Hükümdar	Ucu açık özel işleyiş için tasarlanmış bir sistem *Çeşitlilikler:* Mümkün olan birçok güdülenme sistemi vardır; öndeğerlendirme ve "sponsor onaylama"dan yararlanma olanağı (13. Bölüm'de ele alınacak)	Kutulama yöntemleri geçerli değildir. Diğer yeterlilik kontrol yöntemlerinin çoğu da geçerli değildir (belki toplumsal bütünleşme ya da insani zapturapt hariç) geçerlidir. Evcimenlik büyük oranda geçersizdir. Yapay zekânın insan çıkarlarını ve niyetlerini anlamasına duyulan ihtiyaç çok fazladır. İlk denemede başarılı olmak zorunludur (ama muhtemelen daha az bir ölçüde, bu söylenen diğer tüm kastlar için de geçerlidir) • Sponsor için potansiyel açıdan büyük bir güç kaynağıdır (belirleyici stratejik üstünlük dahil) • Bir kez kullanılmaya başladığında, operatör tarafından ele geçirilmesi mümkün değildir ve aptalca kullanımlara karşı belli oranda korunacak şekilde tasarlanabilir. • "Cehalet kisvesi" sonuçlarını uygulamak için kullanılabilir.
Alet	Hedef güdümlü davranışı sergilemek üzere tasarlanmamış bir sistem	• Kutulama yöntemleri uygulamaya bağlı olarak geçerli olabilir. • Güçlü arama süreçleri bir makine süper zekâsının gelişiminde ve işleyişinde muhtemelen yer alacaktır. • Formel bir kıstası karşılayacak bir çözüm bulmak için güçlü arama kıstası niyetlenilmemiş ve tehlikeli bir tarzda karşılayan çözümler üretebilir. • Güçlü arama asli arama sürecini uygulamaya dökmenin tehlikeli yollarını bulabilecek olan ikincil, iç arama ve planlama süreçleri içerebilir.

ON BİRİNCİ BÖLÜM

Çokkutuplu Senaryolar

Tekkutuplu bir sonucun, tek bir süper zekânın belirleyici bir stratejik üstünlük elde edip bunu bir tekel kurmak için kullanacağı bir sonucun ne kadar büyük bir tehdit oluşturabileceğini özellikle de 8. Bölüm'de gördük. Bu bölümdeyse çokkutuplu bir senaryoda, rekabet halindeki birden çok süper zeki aracının olduğu bir geçiş sonrası toplumunda neler olacağını inceleyeceğiz. Bu senaryo türlerine olan ilgimizin iki boyutu var. Birincisi, 9. Bölüm'de değinildiği üzere, toplumsal bütünleşmenin kontrol problemine bir çözüm sunduğu düşünülebilir. Bu yaklaşımın bazı sınırlılıkları olduğunu belirtmiştik ve bu bölümde daha eksiksiz bir tablo sunuluyor. İkincisi, kontrol problemini halletmenin bir yolu olarak çokkutuplu bir koşul yaratma yönünde herhangi bir girişimde bulunulmadığında bile, bu tür bir sonuç her halükârda gerçekleşebilir. O halde bu tür bir sonuç nasıl bir şeye benzer? Ortaya çıkacak olan rekabetçi toplum çekici olacak ya da uzun süre ayakta kalacak diye bir kaide yok.

Tekelci senaryolarda, geçiş sonrasında neler olacağı neredeyse tamamen tekelci yapının değerlerine bağlıdır. Dolayısıyla sonuç bu değerlerin ne olduğuna bağlı olarak çok iyi de olabilir çok kötü de. Değerlerin ne olduğu ise kontrol probleminin nasıl çözüldüğüne ve –çözüldüğü ölçüde– tekeli yaratan projenin hedeflerine bağlıdır.

Bu nedenle tekelci senaryoların akıbetiyle ilgilenen birinin gerçekte sadece üç bilgi kaynağı vardır: tekelin eylemlerinden etkilenemeyecek meseleler hakkındaki bilgi (fizik yasaları gibi); yakınsak araçsal değerler hakkında bilgi; tekelin hangi nihai değerlere sahip olacağını tahmin etmeyi ya da kurgulamayı mümkün kılan bilgi.

Çokkutuplu senaryolarda, başka birtakım kısıtlamalar da devreye giriyor; bunlar aracıların nasıl etkileşimde bulunduğuyla alakalı kısıtlamalardır. Bu tür etkileşimlerden doğan sosyal dinamikler oyun teorisinden, iktisattan ve evrim teorisinden teknikleri kullanarak incelenebilir. Ayrıca siyaset bilimi ve sosyolojiden öğeler de insan deneyiminin daha olumsal özelliklerinin bazılarından damıtılıp soyutlanabildiği oranda konuyla ilintilidir. Bu kısıtlamaların geçiş sonrası dünya hakkında bize kesin bir tablo çizmesini beklemek gerçekçi olmasa da, bazı belirgin olanakları saptamamıza ve bazı temelsiz varsayımlara karşı çıkmamıza yardımcı olabilirler.

Düzenleme seviyesinin düşük olduğu, mülkiyet haklarının sıkı sıkıya korunduğu ve pahalı olmayan dijital zihinlerin görece hızlı bir şekilde devreye sokulduğu bir ekonomik senaryoyu irdeleyerek başlayacağız.[1] Bu tür bir modelin ilk akla getirdiği isim konuya dair çalışmalarıyla öncü rol oynayan Amerikalı iktisatçı Robin Hanson'dır. Bu bölümün ilerleyen sayfalarında evrimle alakalı bazı meseleleri ele alacağız ve ilk başta çokkutuplu olsa da sonrasında bir tekele dönüşecek olan geçiş sonrası dünyanın olasılıklarını inceleyeceğiz.

Atlar ve insanlar

Genel makine zekâsı insan zekâsının ikamesi olabilir. Şu anda insanların yaptıkları entelektüel işleri dijital zihinler de yapabiliyor, ama dahası makinelere bir kez iyi kumandalar ya da robot bedenler eklendiğinde insanın fiziki emeğini de ikame edebilirler. Çabucak çoğaltılabiliyor olan makine işçilerin insan işçilere kıyasla fiilen her işte hem daha ucuz hem de daha becerikli olduklarını varsayalım. Ne olurdu o zaman?

Ücretler ve işsizlik

Emek gücünün kolaylıkla kopyalanabilir olduğu durumda piyasada ücretler düşer. İnsanların rekabette tutunabildikleri yegâne yer, müşterilerin öncelikli olarak insanların yaptığı işleri tercih ettikleri alanlar olabilir. Bugün, elle yapılan ya da yerli halk tarafından üretilen mallar bazen bir fiyat primine sahip oluyor. Geleceğin müşterileri de buna benzer şekilde, insan yapımı malları ve insan sporcuları, insan sanatçıları ve insan önderleri işlevsel açıdan ayırt edilemeyen ya da üstün olan yapay muadillerine tercih edebilirler. Gelgelelim bu tür tercihlerin ne kadar yaygın olacağı belirsizdir. Eğer makine yapımı alternatifler yeterince üstünse, muhtemelen fiyatları daha yüksek olacaktır.

Tüketici tercihiyle alakalı olabilecek parametrelerden biri hizmet ya da ürün sağlayan işçinin iç dünyasıdır. Sözgelimi bir konser dinleyicisi kitle, icracının müziği ve mekanı bilinçli şekilde deneyimlediğini bilmekten hoşlanabilir. Olgusal deneyim olmadığında müzisyen çok güçlü bir müzik kutusundan ibaret olarak görülebilir, tek farkı kalabalıkla doğal bir şekilde etkileşimde bulunabilen bir icracının üçboyutlu arzı endam edişi olacaktır. O zaman makineler aynı görevi icra eden bir insanda var olan aynı zihinsel durumları somutlaştıracak şekilde tasarlanabilir. Gelgelelim öznel deneyimler kusursuz şekilde çoğaltılabildiğinde bile bazı insanlar salt organik olanı tercih edebilir. Bu tür tercihlerin ideolojik ya da dini kökleri de olabilir. Nasıl ki birçok Müslüman haram, birçok Yahudi trefa diye gördüğü yiyeceklerden uzak duruyorsa, aynı şekilde gelecekte de üretiminde onaylanmamış makine zekâ kullanımı olan ürünlerden uzak duracak gruplar olabilir.

Neye bağlıdır bu? Ucuz makine emeği insan emeğini ikame edebildiği oranda, insan işleri yok olabilir. Otomasyon ve iş kaybı korkuları elbette yeni değildir. Teknolojik işsizliğe ilişkin kaygılar, en azından Sanayi Devrimi'nden bu yana, dönem dönem nevzuhur etmiştir ve azımsanmayacak sayıda iş, 19. yüzyılın başında mekanize dokuma tezgâhlarının kullanılmasına karşı halk kahramanı "General Ludd'ın" bayrağı altında toplanan İngiliz dokumacıları ve mensucat zanaatkârlarının başına geldiği gibi kaybolup gitmiştir. Gelgelelim makineler ve teknoloji insan emeğinin birçok tikel türünü ikame etse de, fiziksel teknoloji bir bütün olarak emeğin tamamlayıcısı olmuştur. Dünyada ortalama insan ücretleri uzun zamandır yükseliştedir ve bunun esas nedeni de bu tür tamamlayıcılıklardır. Fakat emeğin tamamlayıcısı olarak başlayan şey ileriki bir aşmada emeğin ikamesi haline gelebilir. Atlara ilk başta at arabaları ve sabanlar tamamlayıcılıkta bulunuyordu ve bu sayede atların verimliliği muazzam artmıştı. Sonradan, atlar otomobillerle ve traktörlerle ikame edildi. Sonradan gelen bu yenilikler at emeğine olan talebi azalttı ve bir popülasyon çöküşüne yol açtı. Benzer bir akıbet insan türünü de bekliyor olabilir mi?

At hikâyesinin bir benzerine etrafta neden hâlâ daha atların olduğunu sorarak varabiliriz. Nedenlerden biri atların hâlâ birkaç yerde işe yarıyor olmasıdır: Örneğin atlı polisler var. Ama esas neden, tuhaftır, insanların atların sunabileceği hizmetleri tercih ediyor olmalarıdır (at binme ve at yarışı dahil). Bu tercihler bazı insanların gelecekte sahip olabileceğini varsaydığımız tercihlerle kıyaslanabilir: İleride bazı mallar ve hizmetler insan ürünü olabilir. Manidar olsa da, bu örnekseme eksiktir, zira atlar için tam bir işlevsel ikame hâlâ yoktur. Eğer samanla çalışan ucuz mekanik araçlar varsa ve bunlar biyolojik atlarla tastamam aynı şekle ve davranışa sahipse, aynı hissi veriyorsa (hatta aynı bilinçli deneyimleri varsa), o zaman biyolojik atlara olan talep muhtemelen daha da azalacaktır.

İnsan emeğine olan talepte yeterli bir azalma olduğunda ücretler insanın geçim düzeyinin altına düşecektir. Dolayısıyla insan işçiler için potansiyel tehlike çok ciddidir: Sadece ücret kesintileri, tenzil-i rütbe ya da yeniden eğitilme ihtiyacı değil, aynı zamanda aç kalma ve ölüm. Bir taşıma gücü ya da hareket kaynağı olarak atların miadı dolduğunda, birçoğu köpek maması, kemik tozu, deri ve yapıştırıcı olarak işlenmek üzere mezbahalara yollanmıştı. Bu hayvanların geçimlerini sağlayacakları alternatif bir iş olanağı yoktu. ABD'de 1915'te at sayısı 26 milyon kadardı. 1950'lerin başında 2 milyon at kalmıştı.[2]

Sermaye ve refah

İnsanlar ile atlar arasındaki farklardan biri insanların sermaye sahibi olmasıdır. Sermayenin toplam faktör payı uzun zamandır yaklaşık yüzde 30'da değişmeden kalmıştır (kısa vadeli dalgalanmaları saymıyoruz).[3] Bu demektir ki, toplam küresel

gelirin yüzde 30'u sermaye sahipleri tarafından rant olarak alınırken, geri kalan yüzde 70 işçilere ücret olarak gidiyor. Eğer yapay zekâyı sermaye olarak sınıflandırırsak, o zaman insan emeğini tamamen ikame edebilecek makine zekânın icadıyla birlikte, ücretler bu tür makine ikamelerinin son birim maliyeti seviyesine düşecektir ki, makinelerin çok verimli olduğunu varsayarsak, bu rakam çok düşük, insanın geçim sınırının çok altında olacaktır. Bu durumda, emeğin aldığı gelir payı fiilen sıfıra inecektir. Ama burada sermayenin faktör payının dünyadaki toplam ürünün neredeyse yüzde 100'ü olduğu iması vardır. Dünya GSYH'sı bir zekâ patlamasının ardından tavan yapacağından (hem emeği ikame eden yeni makinelerin sayısı muazzam miktarda olacağı için hem de süper zekânın elde ettiği teknolojik ilerlemeler ve akabinde uzaya yerleşme yoluyla muazzam miktarda yeni toprak edinilmesinden ötürü), demek ki sermayenin toplam geliri de muazzam artacaktır. Eğer bu sermayenin sahibi yine insanlar olursa, insan nüfusunun aldığı toplam gelir çarpıcı derecede artacaktır, üstelik bu senaryoda insanların artık hiçbir ücret geliri olmayacaktır.

Böylece bir bütün olarak insan türü akla hayale gelmeyecek kadar zengin olabilir. Bu gelir nasıl bölüşülecektir? İlk hesaplamaya göre sermaye geliri sahip olunan sermaye miktarıyla orantılı olacaktır. Astronomik büyüme etkisi dikkate alındığında, geçiş öncesindeki çok küçük bir servet bile geçiş sonrasında muazzam bir servete dönüşecektir. Gelgelelim günümüz dünyasında birçok insanın serveti sıfırdır. Buna sadece yoksulluk içinde yaşayan bireyler değil, aynı zamanda iyi bir geliri olan ya da yüksek insan sermayesi olan ama olumsuz net değeri olan bazı insanlar da dahildir. Örneğin müreffeh Danimarka'da ve İsveç'te nüfusun yüzde 30'unun negatif serveti vardır; bunların çoğunluğu çok az somut aktif varlığı olan ve kredi kartı borcu ya da okul taksiti bulunan genç orta sınıf insanlardır.[4] Tasarruflar son derece yüksek faiz getirse bile, büyümenin başlaması için elde avuçta bir başlangıç sermayesinin olması gerekir.[5]

Ne var ki geçiş döneminin başında hiç kişisel serveti olmayan bireyler bile fevkalade zengin hale gelebilirler. Sözgelimi bir emeklilik programına (özel ya da devlet) katılanlar, ödemelerin en azından kısmen karşılanması durumunda iyi bir konumda olacaklardır.[6] Yoksullar da öz varlıklarının roket hızıyla yükseldiğini görenlerin hayırseverliği sayesinde zengin hale gelebilirler: Talih kuşunun astronomik büyüklüğünden ötürü çok küçük bir hayrına bağış bile mutlak anlamda çok büyük bir rakam olacaktır.

Zenginliklerin çalışma yoluyla elde edilebilmesi, makinelerin her alanda insanlardan işlevsel olarak üstün olduğu (keza geçim düzeyinde insan emeğinden bile daha ucuz olduğu) geçiş sonrası aşamada bile mümkündür. Daha önce belirtildiği üzere, insan emeğinin estetik, ideolojik, etik, dini ya da diğer pragmatik olmayan nedenlerden ötürü tercih edildiği küçük alanlar varsa bu olabilir. İnsan sermaye sahiplerinin servetinin ciddi şekilde arttığı bir senaryoda, bu tür bir emeğe duyulan talep de buna denk düşen şekilde artabilir. Yeni trilyonerler ya da katrilyonerler mallarının ve hizmetlerinin

bazılarının organik "dürüst ticaret" işgücü tarafından sağlanması karşılığında ciddi bir meblağ ödemeyi kabul edebilirler. Bir kez daha, atların tarihi bir koşutluk sunuyor. ABD'deki at sayısı 1950'lerin başında 2 milyona düştükten sonra ciddi bir artış yaşandı: Yakın döneme ait bir sayımda neredeyse 10 milyon at olduğu hesaplanmıştır.[7] Yükseliş tarımda ya da ulaşımda atlara duyulan yeni işlevsel ihtiyaçlardan değildir; daha ziyade, ekonomik büyüme sayesinde daha fazla Amerikalı bir hobi olarak at binmeye başlamıştır.

İnsanlar ile atlar arasında sermaye sahipliği haricinde, konuyla ilintili bir diğer farklılık da insanların siyasal hareketlilik sergileyebiliyor olmalarıdır. İnsanların yönettiği bir hükümet, devletin vergilendirme yetkisinden yararlanarak özel kârları yeniden paylaştırabilir ya da değerli kamu varlıklarını (kamu arazileri gibi) satarak gelir elde edebilir ve bu gelirle vatandaşlarını emekliliğe ayırabilir. Bir kez daha, geçiş sırasında ve hemen sonrasında muazzam bir ekonomik büyüme olacağından, etrafta müthiş bir zenginlik olacak ve bu sayede tüm işsiz yurttaşların karnını doyurmak görece kolay olacaktır. Sadece tek bir ülkenin tüm dünyadaki insanlara sağlam bir ücret ödeyebilmesi, bugün birçok ülkenin dış yardımlara ayırdığından daha fazla masrafa girmeden mümkün olacaktır.[8]

Tarihsel perspektiften Malthus ilkesi

Buraya kadar değişmez bir insan nüfusu varsaydık. Kısa zaman aralıkları için makul bir varsayım olabilir bu, zira biyoloji insanın üreme hızını sınırlar. Ama uzun zaman aralıklarında varsayım her zaman makul değildir.

İnsan nüfusu son 9 bin yılda bin kat artmıştır.[9] Tarihin ve tarihöncesinin çoğunda insan nüfusu dünya ekonomisinin sınırlarına çarpmamış olsaydı, artış çok daha hızlı olurdu; bahsedilen dönemlerde, çoğu insanın hayatta kalıp ortalama iki çocuk büyütmesine ancak yeten bir geçim düzeyinde gelire sahip olduğu Malthusçu denebilecek bir koşul hüküm sürdü.[10] Geçici ve yerel ferahlama dönemleri vardı: Salgın hastalıklar, iklim dalgalanmaları ya da savaşlar ara ara nüfusu kırdı ve topraklar boşalınca hayatta kalanlar besin alımlarını ilerletebildiler ve daha çok çocuk yetiştirdiler, ta ki saflar sıklaşıp Malthus ilkesi yeniden yerleşene kadar. Ayrıca toplumsal eşitsizlik sayesinde, ince bir elit katman istikrarlı bir şekilde geçim sınırının üstünde yaşayabildi (geçindirilebilecek nüfusun toplam büyüklüğünü biraz azaltma pahasına). Üzücü ve uyumsuz bir düşünce: Bu gezegendeki varlığımızın önemli bir kısmında normal durum olan bu Malthusçu koşulda en büyük insancılların kuraklıklar, salgınlar, katliamlar ve eşitsizlik (yani insan refahının en büyük düşmanları) olduğu söylenebilir: Sadece bunlar ortalama refah düzeyinin ara ara tam geçim sınırının biraz üstüne fırlamasına imkân sağlamıştır.

Yerel dalgalanmalarla birleştiğinde, tarihin makro örüntüsü ilk başta yavaş olup, akabinde teknolojik yeniliklerin itkisiyle hızlanan bir ekonomik büyüme göstermektedir.

Büyüyen dünya ekonomisi küresel nüfusta orantılı bir artış sağlamıştır. (Daha kesin ifade edersek, bizatihi daha büyük bir nüfus büyüme hızını güçlü şekilde ivmelendirmiş gözükmektedir, belki de esasen insanlığın kolektif zekâsını artırmak suretiyle.[11]) Gelgelelim ancak Sanayi Devrimi'nden itibaren ekonomik büyüme nüfus artışının ayak uyduramadığı kadar hızlı hale gelmiştir. Böylece ortalama gelir ilkin Batı Avrupa'daki ilk sanayileşen ülkelerde, ardından dünyanın çoğunda artmaya başladı. Bugün en yoksul ülkelerde bile ortalama gelir, bu ülkelerdeki nüfus artışının da gösterdiği gibi, asgari geçim sınırının çok üstündedir.

Bugün en hızlı nüfus artışı en yoksul ülkelerde görülmektedir, zira bu ülkeler, daha gelişkin toplumlarda çoktan yerleşmiş olan düşük doğurganlık rejimine "demografik geçiş"i henüz tamamlamamıştır. Nüfus uzmanları bu yüzyılın ortasında dünya nüfusunun yaklaşık 9 milyara ulaşacağını ve ardından, yoksul ülkelerin bu düşük doğurganlık rejiminde gelişmiş ülkelerin yanında yer almaya başlamasıyla nüfusun durağan bir noktaya geleceğini ya da azalacağını tahmin ediyorlar.[12] Bugün birçok zengin ülkede doğurganlık oranları yenilenme düzeyinin altındadır, bazı ülkelerdeyse çok altındadır.[13]

Fakat daha uzun vadeli düşünür ve teknolojinin değişmediğini, refahın sürdüğünü varsayarsak, bizim küçük gezegenimizin destekleyebileceği sınırlara çarpacak olan bir dünya nüfusunun tarihsel ve ekolojik açıdan normal koşuluna geri döneceğimizi beklemek yanlış olmaz. Eğer bu söylenenler şu an küresel ölçekte gözlemlediğimiz zenginlik ile doğurganlık arasındaki olumsuz ilişki ışığında sezgilerinize aykırı görünüyorsa, bu modern çağın tarihin sadece kısa bir dilimi olduğunu ve aslında gayet de bir sapma teşkil ettiğini hatırlamamız gerekiyor. İnsan davranışı günümüz koşullarına hâlâ uyum sağlayamamıştır. Kapsayıcı uygunluğumuzu artırmanın bariz yollarından istifade edemediğimiz gibi (sperm ya da yumurta donörü olmak gibi), doğum kontrolünden yararlanarak doğurganlığımızı bilfiil baltalıyoruz. Evrimsel uyarlanmışlık çevresinde, sağlıklı bir cinsel dürtü bireyin üreme potansiyelini azamiye çıkaracak şekillerde hareket etmesi için yeterli olabilirdi; oysa modern çevrede mümkün olan en fazla sayıda çocuğun biyolojik ebeveyni olmak için daha doğrudan bir arzunun muazzam bir seçilimsel üstünlüğü olacaktır. Bugün bu tür bir arzu seçilime konu oluyor, tıpkı üreme eğilimimizi artıran diğer özellikler gibi. Gelgelelim kültürel uyum sağlama biyolojik evrimden rol çalabilir. Bazı topluluklar (Hutterciler ya da Quiverfull evanjelik hareketinin takipçileri gibi) büyük aileleri teşvik eden doğurgan kültürlere sahiptir ve dolayısıyla da hızla büyümektedirler.

Nüfus artışı ve çevre

Bir sihir yapıldığını ve mevcut toplumsal-ekonomik koşulların mevcut şekliyle donduğunu varsayarsak, geleceğe yüksek doğurganlık düzeylerini sürdüren kültürel ya da etnik gruplar damga vuracaktır. Eğer çoğu insan günümüz çevresinde uygunluğu

asgariye çıkarıcı tercihlere sahip olsaydı nüfus her bir kuşakta rahatlıkla iki katına çıkardı. Nüfus kontrol politikaları olmazsa (güçlü tercihlerin evrilip bu politikaları atlatmasına karşı durmak için hep daha sıkı ve etkili olması gerekecektir), dünya nüfusu bir kısıtlama (toprak kıtlığı ya da önemli yenilikler için kolay fırsatların tüketilmesi gibi) ekonominin ayak uydurmasını imkânsız hale getirene kadar üssel büyüme kaydetmeye devam edecektir; bu noktada ortalama gelir azalmaya başlayacak ve sonunda müthiş yoksulluk yüzünden birçok insan ikiden fazla çocuk yetiştiremez hale gelecektir. Böylece Malthus ilkesi ürkütücü bir kölesahibi gibi yine meşum yüzünü gösterecek, bereketli hayal dünyasına kaçışımızı sona erdirecek ve bizi gerisin geriye yeniden prangaya vurup şu bezdirici geçim kavgasını geri getirecektir.

Bu uzun vadeli bakış açısı zekâ patlamasıyla daha yakın bir gelecekte gerçekleşebilir. Yazılım kopyalanabilir olduğundan, emülasyonlardan ya da yapay zekâlardan oluşan bir nüfus hızla –onyıllar ya da yüzyıllar içinde değil dakikalar içinde– iki katına çıkabilir ve mevcut tüm donanımı çok geçmeden tüketebilir.

Özel mülkiyet evrensel bir Malthusçu koşulun ortaya çıkmasına karşı kısmi koruma sağlayabilir. Klanların (ya da kapalı toplulukların veya devletlerin) değişen miktarlarda mülkiyetle başlayıp üreme ve yatırım konusunda bağımsız şekilde farklı politikalar benimsediklerini düşünün. Bazı klanlar geleceği yabana atarlar ve ellerinde ne var ne yoksa harcarlar, sonra da yoksullaşmış üyeleri küresel proletarya saflarına katılırlar (veya emekleriyle geçinemezlerse, ölürler). Diğer klanlar kaynaklarından bazılarını yatırımda kullanır, ama sınırsız bir üreme politikası benimserler: Bu tür klanların nüfusları, üyelerinin neredeyse üremeleriyle aynı hızda ölecek kadar yoksullaştıkları içsel bir Malthusçu koşula ulaşana kadar nüfusları artar ve bu noktada klanın nüfus artışı kaynak artışına denk düşecek şekilde yavaşlar. Başka klanlarsa doğurganlıklarını sermayelerinin büyüme hızının altıyla sınırlayabilirler; bu tür klanlar sayılarını yavaş yavaş artırırken, üyeleri de kişi başına zenginleşir.

Eğer zenginlik zengin klanlardan hızla çoğalan ya da hızla azalan klanların üyelerine dağıtılırsa (çocukları, kopyaları ya da dölleri, kendi hataları olmamasına karşın, dünyaya hayatta kalıp ilerlemek için yetersiz sermayeyle yollanmıştır), o zaman evrensel bir Malthusçu koşula daha fazla yaklaşılacaktır. Sınırlayıcı örnekte tüm klanların tüm üyeleri geçim düzeyinde gelir alacak ve herkes yoksul olacaktır.

Eğer mülkiyet yeniden paylaştırılmazsa, açıkgöz klanlar belli bir miktar sermayeye tutunabilirler ve zenginliklerinin mutlak anlamda büyümesi mümkündür. Gelgelelim insanların sermayelerinden makine zekâlar kadar yüksek oranlarda gelir elde edip edemeyecekleri net değildir, çünkü emek ile sermaye arasında her ikisine de (örneğin hem yetenekli hem varlıklı olan bir müteşebbis ya da yatırımcı) arzda bulunabilen tek bir aracının, sermayesinden finansal kaynaklara sahip ama bilişsel kaynaklardan mahrum olan aracıların erişebildiği piyasa hızını aşan bir getiri oranı elde edebildiği düzeyde

sinerjiler olabilir. İnsanlar makine zekâlar kadar yetenekli olmadıklarından, sermayelerini daha yavaş artırabilirler – elbette kontrol problemi tamamen çözülmedikçe, zira bu durumda insanların gelir elde etme hızı makinelerinkine denk hale gelecektir, çünkü bir insan asil, bir makine vekili tasarruflarını yönetmekle görevlendirebilir ve bunu masrafsız şekilde ve çıkar çatışması olmadan yapabilir; ama aksi durumda, bu senaryoda, ekonominin makinelerin elinde olan kısmı asimptotik şekilde yüzde yüze yaklaşacaktır.

Ekonominin makinelerin elinde olan kısmının asimptotik şekilde yüzde yüze yaklaştığı bir senaryoda, insanın payına düşen kısım azalacak diye bir kaide yok. Eğer ekonomi yeterli bir süratte büyürse, o zaman görece azalan bir kısmı bile yine de mutlak hızda artıyor olabilir. Bunu insanlık için mütevazı düzeyde iyi bir haber olarak görebilirsiniz: Mülkiyet haklarının korunduğu çokkutuplu bir senaryoda (kontrol problemini hiçbir şekilde çözemezsek bile) insanların sahip olduğu toplam servet miktarı artabilir. Elbette bu durum ne insan nüfusundaki artışın kişi başına geliri asgari geçim düzeyine indirmesi problemini ne de geleceğe önem vermedikleri için kendilerini mahveden insanlar problemini çözer.

Uzun vadede ekonomiye giderek en yüksek tasarruf oranına sahip olan klanlar (şehrin yarısına sahip olan ve bir köprü altında yaşayan para canlısı pintiler) hükmeder hale gelecektir. Ancak yatırım için daha fazla fırsatın sunulduğu zaman geldiğinde azami derecede zengin pintiler tasarruflarını aşağı çekmeye başlayacaklardır.[14] Gelgelelim eğer mülkiyet hakları kusursuz şekilde korunmuyorsa (örneğin daha verimli makineler allem edip kallem edip zenginliği insanlardan kendilerine aktarmayı başarırlarsa), o zaman insan kapitalistler sermayelerini çok daha önce harcamak zorunda kalabilirler ki, bu tür aktarımlar yüzünden (ya da bu tür aktarımlara karşı servetlerini korumak için yaptıkları masraflar sürekli artacağı için) sermayeleri tükenip gitmesin. Eğer bu gelişmeler biyolojik zaman ölçeğinde değil de dijital zaman ölçeğinde gerçekleşirse, o zaman Buzul Çağı insanları göz açıp kapayıncaya kadar mülksüzleştirildiklerini görebilirler.[15]

Bir algoritmik ekonomide hayat

Geçiş sonrası Malthusçu durumda biyolojik insanlar için hayatın insanın tarihsel durumlarından herhangi birine (avcı-toplayıcı, çiftçi ya da büro işçisi) benzemesi gerekmiyor. Bu senaryoda insanların çoğunluğu tasarruflarıyla kıt kanaat geçinen aylak rantiyeciler olabilirler.[16] Çok yoksuldurlar, ama küçük gelirlerini tasarruflarından ya da devlet sübvansiyonlarından elde ederler. Teknolojinin son derece ileri olduğu bir dünyada yaşarlar, sadece süper zeki makinelerin değil, yaşlandırmayı önleyici tıbbın, sanal gerçekliğin ve çeşitli geliştirme teknolojilerinin ve haz ilaçlarının olduğu bir dünyadır bu; ama bunlara genel itibariyle paraları yetmez. Belki de modern tıptan

yararlanmak yerine büyümelerini durduracak ve metabolizmalarını yavaşlatacak ilaçlar alırlar ki, geçim masrafları azalsın (kabına sığamayan insanlar tedricen azalan asgari gelirle hayatta kalmayı başaramazlar). Sayımız arttıkça ve ortalama gelirimiz daha da azaldıkça, bir emekli maaşı almak için yeten asgari yapı neyse o derekeye düşebiliriz: Belki de makinelerden oksijen ve besin alan asgari derecede bilinçli kavanozdaki beyinler olarak yavaş yavaş yeterli parayı biriktirip bir robot teknisyene klonumuzu yaptırarak üretmemizi hallederiz.[17]

Daha fazla tutumluluk ise yükleme yoluyla elde edilebilir, zira ileri süper zekânın tasarladığı fiziksel açıdan optimum hale getirilmiş bir hesaplama substratı, bir biyolojik beyinden daha verimli olacaktır. Gelgelelim emülasyonlar emekli maaşı almaya ya da vergiden muaf tasarruf hesabı açmaya uygun olmayan gayri insan ya da gayri yurttaş olarak görülürlerse, dijital alana göçün önü kesilebilir. Bu durumda, biyolojik insanlar için bir adacık açık kalabilir, ama muhtemelen, çok daha büyük bir nüfus oluşturan emülasyonlarla ya da yapay zekâlarla yan yana.

Buraya kadar sadece insanların akıbetine odaklandık ve tasarruflarla, sübvansiyonlarla ya da insan çalıştırmayı tercih eden diğer insanlardan elde ettikleri ücret geliriyle kendini geçindirmelerinin mümkün olabileceğini söyledik. Şimdi de dikkatimizi buraya kadar "sermaye" diye sınıflandırılan varlıklardan bazılarına çevirelim: İnsanların sahibi olabilecekleri, yerine getirdikleri işlevsel görevler hatırına üretilip kullanılan ve çok çeşitli işlerde insan emeğinin yerini alabilecek olan makineler. Yeni ekonominin bu yük beygirlerinin hali pürmelali ne olacak?

Eğer bu makineler salt otomatlarsa, bir buhar makinesi ya da saatteki mekanizma gibi basit araç gereçlerse, o zaman daha fazla yoruma gerek olmayacaktır: Bir geçiş sonrası ekonomisinde bu tür bir sermayeye fazlasıyla ihtiyaç olacaktır, ama sezgili olmayan bir alet edevat parçası için işlerin ne âlemde olduğunu kimse umursamayacaktır. Gelgelelim eğer makinelerin bilinçli zihinleri varsa –eğer işleyişleri olgusal farkındalıkla bağlantılı olacak bir tarzda inşa edildilerse (veya başka bir nedenle onlara ahlaki statü hamledildiyse)– o zaman genel sonucu bu makine zihinleri nasıl etkileyeceği üzerinden değerlendirmek önemli hale gelir. Çalışan makine zihinlerin refahı, sonucun en önemli veçhesi olarak bile gözükebilir, zira sayısal üstünlük onlarda olabilir.

Gönüllü kölelik, rastgele ölüm

İlk göze çarpan soru şudur: Bu çalışan makine zihinler sermaye (köle) olarak mı tutulacaklar yoksa özgür ücretli işçi olarak mı kiralanacaklar? Ne var ki yakından incelendiğinde, bu meselenin çok da önemli olmadığı söylenebilir. Bunun iki nedeni var: Birincisi, eğer bir Malthus devletinde özgür bir işçiye asgari geçim düzeyinde ücret ödenirse, yiyecek ve diğer zorunlu şeylerin parasını ödedikten sonra elde avuçta bir şey kalmayacaktır. Yok eğer işçi bir köleyse, sahibi onun bakımı için ödeme yapacak ve yine

geriye bir gelir kalmayacaktır. Her iki durumda da, işçinin sadece zorunlu ihtiyaçları karşılanacaktır, o kadar. İkincisi, özgür emekçinin bir şekilde geçim düzeyinin üstünde gelir elde ettiğini varsayalım (düzenlemenin bir kıyağıyla belki de). Artıyı nasıl harcayacak? Yatırımcılar "gönüllü köle" olacak, yani asgari geçim düzeyinde ücretlere gönül rızasıyla çalışacak işçiler yaratmayı daha kârlı göreceklerdir. Bu durumda, itaatkâr olan işçileri kopyalayarak bu tür işçiler yaratabilirler. Uygun seçilimle (ve belki de kodda belli bir değişiklikle) yatırımcılar sadece çalışmaya gönüllü olmayı tercih etmeyen aynı zamanda bir şekilde elde edebilecekleri artı geliri sahiplerine geri bağışlamayı da seçecek olan işçiler yaratmayı başarabilirler. O zaman, işçi isterse tüm yasal hakları olan özgür bir birey olsun, parayı işçiye vermek aslında parayı sahibin ya da işverenin bir cebinden diğerine aktarmak olacaktır.

Belki içinizden bazıları kendisine verilen her işi gönüllü yapacak ya da ücretini sahibine bağışlayacak bir makine tasarlamanın zor olacağı itirazında bulunacaktır. Özellikle de emülasyonların daha tipik anlamda insani arzuları olacağı düşünülebilir. Ama özgün kontrol problemi zor olsa bile, burada geçişten *sonraki* bir durumu ele alıyoruz ve bu dönemde güdülenme seçilimi yöntemleri muhtemelen kusursuzlaştırılmış olacaktır. Emülasyonlar söz konusu olduğunda, sadece önceden var olan insan karakterleri havuzundan *seçim yaparak* epey ilerlenebilir; zaten başka birkaç güdülenme seçilimi yöntemi de tarif etmiştik. Kontrol problemi bazı açılardan şu mevcut varsayımla da basitleştirilebilir: Yeni makine zekâ yasalara uyan diğer süper zeki aracıların zaten var olduğu istikrarlı bir toplumsal-ekonomik matrise girecektir.

O halde, şimdi de ister köle ister özgür olsun işçi sınıfı makinelerin ahvalini ele alalım. Öncelikle, zihinde canlandırılması en kolay örnek olan emülasyonlara odaklanacağız.

Dünyaya yeni bir biyolojik insan işçi getirmek ne kadar deneyim ve uzmanlık gerektiğine bağlı olarak on beş ila otuz yıl arasında bir süre alır. Bu süre zarfında yeni insanın beslenmesi, konut, yiyecek-içecek ve eğitim ihtiyaçlarının karşılanması gerekir, hepsi de çok büyük masraflardır. Hâlbuki bir dijital işçinin yeni bir kopyasını yapmak yeni bir programı işler belleğe yüklemek kadar kolaydır. Böylece yaşam daha ucuz hale gelir. Bir işletme yeni kopyalar üreterek işgücünü sürekli taleplere uygun hale gelecek şekilde uyarlayabilir; ihtiyaç olmayan kopyalarıysa çöpe atabilir ve bilgisayar kaynaklarında yer açar. Bu durum dijital işçiler arasında ölüm oranının son derece yüksek olmasına yol açabilir. Birçoğu sadece tek bir öznel gün yaşayabilir.

Emülasyon sahiplerinin ya da işverenlerin işçilerini sık sık "öldürmek" ya da "sonlandırmak" isteyebilmelerinin tek nedeni talep dalgalanmaları değildir.[18] Eğer bir emülasyon zihin, tıpkı bir biyolojik zihin gibi, işlev göstermek için dinlenme ve uyuma dönemlerine ihtiyaç duyarsa, bir günün sonunda yorgun emülasyonu silip yerine depodaki taze ve dinlenmiş emülasyonu sürmek daha ucuz olabilir. Bu usul

o gün öğrenilmiş olan her şeyde geriye dönük amneziye yol açacağından, uzun bilişsel bağlar gerektiren işlerde çalışan emülasyonların bu şekilde sık sık silinmesi söz konusu olmayacaktır. Örneğin eğer her sabah masanın başına geçtiğinde önceki gün ne yaptığınızı hatırlayamıyorsanız kitap yazmak zor olacaktır. Ama diğer işler sık sık yenileriyle değiştirilen aracılar tarafından layıkıyla yerine getirilebilir: Bir dükkândaki tezgâhtarın ya da bir müşteri hizmetleri çalışanının, gerekli eğitimi aldığında yeni bilgiyi sadece yirmi dakikalığına hatırlaması gerekebilir.

Yenileriyle değiştirilen emülasyonlar bellek ve yetenek oluşumunu engelleyeceğinden, bazı emülasyonlar uzun bilişsel zincirleri tam anlamıyla gerektirmeyen işlerde bile dinlenme ve uyuma için de dahil olmak üzere sürekli çalışacakları özel bir eğitim programına sokulabilirler. Örneğin bir müşteri hizmetleri çalışanı koçların ve performans değerlendirmecilerinin yardımıyla optimum öğrenme çevrelerinde uzun yıllar çalıştırılabilir. Ardından, bu öğrencilerin en iyileri damızlık olarak kullanılacak, milyonlarca yeni kopyanın her gün çıkarılacağı bir kalıp işlevi göreceklerdir. Bu tür kalıp işçilerin performansını artırmak için büyük çaba harcanacaktır, çünkü üretkenlikteki en ufak bir artış bile milyonlarca kopyaya uygulandığında büyük bir ekonomik değer taşıyacaktır.

Belli işler için kalıp işçileri eğitme çabalarına paralel olarak, temeldeki emülasyon teknolojisini geliştirmek için de yoğun çaba harcanması gerekecektir. Buradaki ilerlemeler bireysel kalıp işçilerdeki ilerlemelerden bile daha değerli olacaktır, zira genel teknolojideki iyileşmeler sadece belli bir meslekte çalışanlara değil, tüm emülasyon işçilere (ve potansiyel açıdan, işçi olmayan emülasyonlara da) uygulanabilir. Mevcut emülasyonların daha verimli uygulamalarına ve nöromorfik ve tamamen sentetik yapay zekâ mimarilerinin geliştirilmesine izin veren bilgisayar kısayollarını bulmak için muazzam kaynaklar harcanacaktır. Bu araştırmanın büyük kısmı muhtemelen çok hızlı donanımda çalıştırılan emülasyonlar tarafından yapılacaktır. Bilgisayar gücünün fiyatına bağlı olarak, en iyi insan zihinlerinin (ya da onların geliştirilmiş versiyonlarının) milyonlarca, milyarlarca, hatta trilyonlarca emülasyonu makine zekâ cephesini ilerletmek için gece gündüz demeden çalışabilir ve bunlardan bazıları biyolojik beyinlerden çok daha hızlı büyüklüklerde işlev gösterebilir.[19] İnsan türü emülasyonlar çağının kısa süreceğini –yıldız zamanında *çok* kısa bir reklâm arası olacağını– ve çok geçmeden yerini son derece üstün yapay zekâ çağına bırakacağını düşünmek için sağlam bir nedendir bu.

Emülasyon işçi çalıştıranların sürüdeki çürük elmaları neden dönem dönem ayıklayabileceklerine dair birkaç nedenden yukarıda bahsettik: Farklı türde işçi talebindeki dalgalanmalar, dinlenme ve uyku zamanı emülasyonun olmamasının sağladığı tasarruflar ve yeni, gelişkin kalıpların kullanıma sokulması. Güvenlik kaygıları da bir diğer neden olabilir. İşçilerin yıkıcı planlar ve komplolar hazırlamalarını engellemek

için, bazı hassas mevkilerdeki emülasyonlar sadece sınırlı dönemler için çalıştırılıp öncesinde depolanmış hazır-duruma sık sık resetlenebilirler.[20]

Emülasyonların resetleneceği bu hazır durumlar dikkatlice hazırlanıp gözden geçirilecektir. Tipik bir kısa ömürlü emülasyon; sadakat ve üretkenklik için optimum olan dinlenmiş bir zihinsel durumda uyanabilir. Yıllar süren (öznel) yoğun eğitim ve seçilimin ardından sınıfının en iyisi olarak mezun olduğunu, ardından canlandırıcı bir tatil yaptığını, iyi bir uyku çektiğini, ardından galeyana getiren bir konuşma ve gaza getirici bir müzik dinlediğini hatırlar ve nihayet şimdi de, işe koyulmak ve işvereni için elinden geleni yapmak için yanıp tutuşuyordur. İşgününün sonunda öleceği düşüncesi onu o kadar da rahatsız etmiyordur. Ölüm nevrozları ya da diğer takıntıları olan emülasyonlar daha verimsiz olurlar ve bu durumda seçilmezlerdi.[21]

Azami derecede verimli bir çalışma eğlenceli olur muydu?

Bunun gibi varsayımsal bir koşulun istenirliğini değerlendirmede önemli değişkenlerden biri ortalama emülasyonun haz durumudur.[22] Tipik bir emülasyon işçi dert nedir bilir mi, yoksa önüne konan işte sıkı çalışma deneyiminden zevk mi alacaktır?

Kendi duygularımızı hayali emülasyon işçilere taşımak doğru olmaz. Mesele *siz* sürekli çalışsanız ve sevdiklerinizle geçirecek zaman bulamasanız ne hissedeceğiniz değildir; çoğunuz bunun korkunç bir kader olacağını düşünür, ama o ayrı.

Bugünkü insanın çalışma saatlerindeki ortalama haz deneyimini düşünmek görece daha ilintilidir. Dünya çapında yürütülen çalışmalarda, insanlara ne kadar mutlu oldukları sorulduğunda çoğu kendisinin "gayet mutlu" ya da "çok mutlu" olduğunu söylüyor (1 ila 4 ölçeğinde ortalama 3,1'dir).[23] Ortalama duygu durumuna ilişkin araştırmalardaysa, insanlara son dönemde çeşitli olumlu ya da olumsuz duygusal durumları ne kadar sık yaşadıkları sorulduğunda da benzer bir sonuç elde ediliyor: -1 ila 1 ölçeğinde net duygu oranı yaklaşık 0,52'dir. Bir ülkenin kişi başına gelirinin ortalama öznel esenliğe olumlu etkisi mütevazı düzeydedir.[24] Gelgelelim bu bulgulardan hareketle geleceğin emülasyon işçilerinin haz durumuna ilişkin çıkarımda bulunmak tehlikelidir. Bunun nedenlerinden biri, emülasyonların koşulunun çok farklı olacak olmasıdır: Bir yandan, çok daha sıkı çalışıyor olabilirler; diğer yandan, hastalık, ağrı sızı, açlık, zehirli kokular gibi şeylerden azade olabilirler. Fakat bu tespitler yine de hedefi büyük oranda ıskalar. Burada çok daha önemli olan husus, haz tonunu ilaçların ya da beyin cerrahisinin dijital muadili yoluyla ayarlamanın kolay olacak olmasıdır. Yani geleceğin emülasyonlarının haz durumunu bizim ve bizim gibi diğer insanların bu koşullarda nasıl hissedeceklerini tahayyül ederek çıkarsamaya çalışmak hata olur. Haz durumu bir tercih meselesidir. Şu an ele aldığımız modelde, tercihi yapanlar, emülasyon işçilere yatırımlarından azami kazanç elde etmeye çalışan sermaye sahipleridir.

Dolayısıyla emülasyonların ne kadar mutlu hissedecekleri (istihdam edilecekleri çeşitli işlerde) hangi haz durumlarının en üretken olacağına gelip dayanmaktadır.

Burada bir kez daha insanların mutluluğuna dair gözlemlerden çıkarımda bulunmaya çalışabiliriz. Eğer çoğu yerde, zamanda ve meslekte durum buysa, insanlar genellikle en azından mütevazı derecede mutluysa, bu durum şu an ele aldığımız gibi bir geçiş sonrası senaryosunda benzer bir durum olabileceği varsayımını doğuracaktır. Daha net konuşmak gerekirse, bu örnekte savunulan şey insan zihinlerinin mutlu olmaya meyilli olduğu ve bu yeni koşullarda da muhtemelen kendisini tatmin edeceği değil; daha ziyade insan için belli bir ortalama mutluluk düzeyinin geçmişte uyarlanabilir olduğu kanıtlandığına göre, belki gelecekte de insan türü zihinler için de benzer bir mutluluk düzeyi böyle olacaktır. Ama bu formül çıkarımın zayıflığını da eleverir, yani Afrika savanalarında dolanan avcı-toplayıcı insansılar için uyarlanabilir olan zihinsel temayüller geçiş sonrası sanal gerçekliklerde yaşayan modifiye emülasyonlar için de uyarlanabilir olacak diye bir kaide yoktur. Geleceğin emülasyon işçilerinin insanlık tarihindeki tipik işçiler kadar ya da daha mutlu olacaklarını kesinlikle *umut edebiliriz*; ama (şu an ele aldığımız çokkutuplu bırakınız yapsınlar senaryosunda) böyle olacağını varsaymak için henüz ikna edici bir neden yok.

Mutluluğun insanlar arasında yaygın olmasının nedeni (artık hangi sınırlı ölçüde yaygınsa), neşeli ruh halinin evrimsel uyarlanmışlık çevresinde bir işaretleşme işlevi görmesi olabilir mi acaba? Toplumsal grubun diğer üyelerine çok iyi durumda olduğu izlenimini vermek (sağlığım sıhhatim yerinde, akranlarımla halim keyfim iyi ve iyi bir istikbalim olacağına eminim) bireyin popülerliğini artırmış olabilir. Dolayısıyla neşelilik konusundaki eğilim seçilimin sonucu olmuş olabilir ve bunun sonucunda insan nörokimyası şu anda basit maddiyatçı kıstaslara göre azami derecede verimli olabileceklere kıyasla olumlu duyguya yatkındır. Eğer durum böyleyse, o zaman yaşama sevincinin geleceği, geçiş sonrası dünyada neşenin toplumsal işaretleşme işlevini korumasına bağlı olabilir; bu konuya birazdan geri döneceğiz.

Peki ya memnun canlar somurtkanlardan daha çok enerji yayarlarsa? Belki de neşeliler yaratıcı sıçramalara ve hayaller kurmaya, yani geleceğin işverenlerinin çoğu işçisinde değersiz göreceği davranışlara daha yatkındır. Belki de işi hatasız yapmayı takıntı haline getiren asık suratlı ya da endişeli tavır çoğu işte üretkenliği azamiye çıkaran bir tutum olacaktır. Ben böyle olduğunu iddia etmiyorum, sadece böyle olup olmayacağını bilmediğimizi söylüyorum. Yine de gelecekteki bir Malthusçu koşula ilişkin böyle karamsar bir hipotezin doğru olduğu anlaşılırsa, işlerin ne kadar kötü olabileceğini düşünmek zorundayız: Sadece daha iyi bir şey yaratmayı başaramamamızın fırsat maliyetinden ötürü değil (bu muazzam olacaktır), aynı zamanda durum zaten başlı başına kötü olacağı, hatta belki de ilk Malthusçu durumdan bile çok daha kötü olacağı için de.

Nadiren tüm çabamızı ortaya koyarız; koyduğumuzdaysa, bu bazen acı verir. Koşu bandının dik yokuş modunda koştuğunuzu tahayyül edin: Kalbiniz güm güm atar, kaslarınızın acıdığını hissedersiniz, ciğerleriniz hava hava diye çırpınır. Zamanölçere bir bakın: Bir sonraki aranız, ki aynı zamanda ölümünüz olacaktır, 49 yıl, 3 ay, 20 gün, 4 saat, 56 dakika ve 12 saniye sonradır. Keşke doğmasaydım dersiniz.

Bir kez daha iddia bunun böyle olacağı değil, böyle olmayacağını biliyor olmamızdır. Kuşkusuz daha iyimser bir öngörüde de bulunulabilir. Örneğin emülasyonların bedensel yaralanma ve hastalık yaşamaları için bir neden yoktur: Fiziki acizliğin kaldırılması mevcut duruma göre büyük bir iyileşme olacaktır. Ayrıca sanal gerçekliğin yaratıldığı şey çok ucuz olabileceğinden, emülasyonlar şatafatlı ortamlarda (dağın tepesindeki enfes saraylarda, çiçeklerin yeni yeni açtığı bahar ormanlarındaki taraçalarda veya masmavi bir gölün kumsallarında) doğru aydınlanmayla, sıcaklıkla, manzara ve dekorla; rahatsız edici gazlardan, gürültülerden, dumanlardan ve böceklerden vareste şekilde; rahat kıyafetler içinde, temiz ve odaklanmış hissederek ve iyi beslenmiş bir halde çalışabilirler. Daha önemlisi, eğer –tamamen mümkün göründüğü üzere– birçok işte üretkenlik için optimum olan beşeri zihinsel durum eğer neşeli heveslilikse, o zaman emülasyon ekonomisi çağı tam bir cennet olabilir.

Her halükârda, eğer varsayılan güzergâh olur da distopyaya kayarsa, meseleleri bir insanın ya da bir şeyin işleri düzeltmek üzere araya gireceği bir tarzda ayarlamanın muazzam bir seçenek değeri olacaktır. Ayrıca yaşam kalitesi yok oluşun yaşamaya devam etmekten daha iyi bir seçenek olduğu şekilde sürekli kötüleşirse, ölüm ve nisyana izin verecek bir tür kaçış politikası da arzulanır olabilir.

Bilinçsiz fasonlar?

Uzun vadede, emülasyon çağı yerini yapay zekâ çağına bıraktıkça (ya da makine zekâya öncesinde bir tüm beyin emülasyonu aşaması olmadan doğrudan varılırsa), acı ve hazzın çokkutuplu bir sonuçta tamamen ortadan kalkması mümkündür, zira hazcı bir ödül mekanizması karmaşık bir yapay aracı (insan zihninin aksine hayvan beyin gücü mirasını yük olarak taşımayan bir aracı) için en etkili güdülenme sistemi olmayabilir. Belki de daha ileri bir güdülenme sistemi haz ve acı konusunda kesin işlevsel analogları olmayan bir fayda işlevinin ya da başka bir mimarinin aleni bir temsiline dayalı olacaktır.

Bununla ilintili ama biraz daha radikal bir çokkutuplu sonuç (her türlü değerin gelecekte safdışı bırakılmasını içerebilecek bir sonuç) ise evrensel proletaryanın bilinçli bile olmayacak olmasıdır. Bu olasılık en çok da yapay zekâ için geçerlidir, zira insan zekâsından çok farklı şekilde yapılandırılması mümkündür. Ama makine zekâ ilk başta tüm beyin emülasyonu yoluyla elde edilip bilinçli dijital zihinlere varsa bile, bir geçiş sonrası ekonomisinde salıverilen rekabet güçleri ya sentetik yapay zekâ en baştan

yaratıldığı ya da emülasyonlar art arda gelen değişiklikler ve iyileştirmeler sonucunda özgün insani biçiminden önemli ölçüde uzaklaşacağı için makine zekânın giderek daha az insanbiçimsi türlerinin ortaya çıkışına kolaylıkla yol açabilir.

Emülasyon teknolojisi geliştirildikten sonra, nörobilim ve bilgisayar bilimindeki süregelen ilerlemenin (dijital zihinlerin hem araştırmacı hem de denek olarak hizmet vermesinin de sağladığı kolaylıkla) bireysel bilişsel modülleri bir emülasyonda yalıtmayı ve bunları diğer emülasyonlardan ayrı şekilde modüllere takmayı mümkün hale getirdiği bir senaryoyu düşünelim. Farklı modüllerin etkili şekilde işbirliği yapabilmesi için bir eğitim ve uyum dönemi gerekebilir; ama ortak standartlara uyum sağlayan modüller diğer standart modüllerle daha çabuk şekilde arayüz kurabilirler. Bu durum standartlaştırılmış modülleri daha verimli hale getirecek ve daha fazla standartlaşma için baskı yaratacaktır.

Emülasyonlar artık işlevselliklerinin artan kısımlarını fason olarak vermeye başlayabilirler. Sayısal muhakeme görevini Gauss-Modules, Inc.'a yollayabilecekken neden aritmetik öğrenesin ki? Düşüncelerini kelimelere dökmek için Coleridge Conversations'ı tutabilecekken neden sözlü ifadeyle uğraşasın ki? Hedeflerinizi sizin hedef sisteminizi izleyip kaynaklarınızı yöneterek kendi başınıza denediğinizde elde edeceğinizden daha iyi kotarabilecek tasdikli yürütme modülleri varsa neden kişisel hayatınızla ilgili kararlar alasınız ki? Bazı emülasyonlar işlevselliklerinin çoğunu korumayı ve başkalarının daha verimli şekilde yapabileceği işleri bizzat yapmayı tercih edebilirler. Bu emülasyonlar kendi bitkilerini yetiştirmekten ya da kendi hırkalarını örmekten zevk alan hobi düşkünleri gibi olacaklardır. Bu tür hobici emülasyonların verimi daha az olacaktır ve ekonominin daha az verimli katılımcılarından daha çok verimli katılımcılarına net bir kaynak akışı olursa, sonunda hobiciler kaybedecektir.

Dolayısıyla ayrı insan-türü zekâların olduğu bulyon küpler algoritmik bir çorbada eriyip dağılır.

Yeterlilikleri bir insan zihninin bilişsel mimarisine kabaca denk düşen toplamlar haline getirerek optimum verimlilik elde edilmesi tasavvur edilebilir. Örneğin bir matematik modulünün bir dil modülüne bağlanması ve her ikisinin de üçünün birlikte çalışması için bir yürütme modülüne bağlanması söz konusu olabilir. O zaman, bilişsel fasonluk neredeyse tamamen işlemez hale gelecektir. Ama bunun böyle olduğuna güven duymak için ikna edici bir gerekçenin yokluğunda, insan türü bilişsel mimarilerin ancak insan nörolojisinin kısıtlamaları dahilinde optimum olduğu (ya da hiç olmadığı) olasılığını uygun görmemiz gerekir. Biyolojik nöral ağlara uygulanamayacak mimariler inşa etmek mümkün hale geldiğinde, yeni tasarım uzamı açılır ve bu genişletilmiş uzamdaki küresel optimumların tanıdık türde zihniyetlere benzemesi gerekmez. İnsan türü bilişsel örgütlenmeler bu durumda geçiş sonrasındaki rekabetçi ekonomide ya da ekosistemde bir niş bulamayacaklardır.[25]

İnsan zihniyle kıyaslandığında daha az karmaşık olan (bireysel modüller gibi), daha fazla karmaşık olan (devasa modül kümeleri gibi) ya da benzer karmaşıklıkta ama son derece farklı mimarileri olan karmaşıklıklar için nişler olabilir. Bu karmaşıklıkların içsel değeri olacak mıdır? Bu tür yabancı karmaşıklıkların insan karmaşıklıklarını ikame ettiği bir dünyaya hoş geldin demeli miyiz?

Sorunun cevabı bu yabancı karmaşıklıkların özgül doğasına bağlı olabilir. Mevcut dünyada birçok örgütlenme düzeyi vardır. Çokuluslu şirketler ve ulus-devletler gibi bazı karmaşık oluşumlarda bileşen olarak insanlar vardır; ama bu yüksek düzey karmaşıklıklara genellikle sadece araçsal değer bahşederiz. Şirketlerin ve devletler, onları oluşturan insanların bilincinin üstünde ve ötesinde bilince sahip değildir (genellikle varsayılan budur): olgusal acı, haz ya da herhangi tür bir deneyim hissedemezler. Onlara insan ihtiyaçlarına hizmet ettikleri ölçüde değer veririz ve bu işlevleri sona erdiğinde onları hiç acımadan "öldürürüz". Ayrıca düşük düzey varlıklar da vardır ve onlara da genellikle ahlaki bir statü atfedilmez. Bir uygulamayı akıllı telefondan silmenin zararı olmadığına inanırız ve bir beyin cerrahı epilepsili beyinden bozuk bir modülü yok ettiğinde kimseye bir yanlış yapmadığını düşünürüz. İnsan beynine benzer bir düzeydeki egzotik olarak örgütlenmiş karmaşıklıklara gelince, bilinçli deneyim için bir kapasitesi ya da potansiyeli olduğunu düşünseydik, herhalde çoğumuz ahlaki bir mahiyet atfederdik.[26]

Dolayısıyla uç bir örnek tahayyül edebiliriz: Bazıları şu an gezegende var olan her şeyden çok daha incelikli ve zeki olan birçok karmaşık yapının yer aldığı teknolojik açıdan çok daha ileri bir toplum, ama bilinçli olan hiçbir varlığın bulunmadığı ya da refahı ahlaki mahiyet taşımayan bir toplum. Ekonomik açıdan mucizelerin, teknolojik açıdan muhteşem şeylerin olduğu ve bunlardan kâr edecek kimsenin yer almadığı bir toplum. Çocukların olmadığı bir Disneyland.

Evrim her zaman ileriye doğru değildir

"Evrim" kelimesi, belki de evrimi bir iyilik kuvveti olarak gören genelgeçer anlayışın bir yansıması olarak, genellikle ilerlemeyle eşanlamlı kullanılır. Evrimsel sürecin içsel hayırlılığına duyulan yanlış inanç zeki yaşamın geleceğinin rekabet dinamikleri tarafından belirlendiği çokkutuplu bir sonucun arzulanırlığına dair adil bir değerlendirmeye engel olabilir. Bu tür bir değerlendirmenin geçiş sonrası dijital hayat çorbasında uyum sağlayabilir olduğunu gösteren farklı fenotiplerin olasılık dağılımına dair en azından zımni bir kanaate yaslanması gerekir. Bu meselelere hâkim olan kaçınılmaz belirsizlik yapışkanından açık ve doğru bir cevap çıkarmak kolay olmayacaktır; hele bir de buna bir iyimserliği mantıksızlık derecesindeki bir katman eklersek.

İleriye doğru akan bir evrime duyulan inancın kaynaklarından biri geçmişte evrimsel sürecin yönünün yukarı doğru olmuş olmasıdır. Evrim iptidai çoğaltıcılardan

başlayarak giderek "ileri" organizmalar üretti, buna zihinleri, bilinci, dili ve aklı olan varlıklar da dahildir. Daha yakın dönemde, biyolojik evrimle az da olsa benzerlikler gösteren kültürel ve teknolojik süreçler insanların ivmelenmiş bir hızda gelişmelerini mümkün hale getirmiştir. Gerek tarihsel gerek jeolojik zaman ölçeğinde bakıldığında, büyük resimde karmaşıklık, bilgi, bilinç ve eşgüdümlü hedefe yönelik örgütlenme düzeylerinin giderek arttığı bir eğri görünmektedir: Bu trendi, doğrusunu söylemek gerekirse giderek, "ilerleme" diye adlandırmak mümkündür.[27]

İyicil sonuçlar ürettiğine güvenilebilecek bir süreç olarak evrim imgesini, hem insan dünyasında hem de doğada gördüğümüz muazzam ıstırapla uzlaştırmak zordur. Evrimin başarılarından memnun olanlar bunu etikten ziyade estetik bir perspektiften yapıyor olabilirler. Ama burada ilintili olan soru ne tür bir geleceği bir bilimkurgu romanında okumanın ya da bir doğa belgeselinde izlemenin büyüleyici olacağı değil, ne tür bir gelecekte yaşamanın iyi olacağıdır: İki çok farklı mesele.

Ayrıca var olan her türlü ilerlemenin bir şekilde kaçınılmaz olduğunu düşünmek için hiçbir neden yoktur. Çoğu şans eseri olmuş olabilir. Bu itiraz bir gözlem seçilim etkisinin kendi evrimsel gelişimimizin başarısına dair elde edebileceğimiz kanıtları filtrelediği gerçeğinden destek alır.[28] Yaşamın ortaya çıktığı tüm gezegenlerin yüzde 99,999'unda, zekâ sahibi gözlemcilerin, kökenlerini düşünmeye başladıkları aşamaya kadar gelişmeden önce yaşamın sona erdiğini varsayalım. Durum böyle olsaydı ne sonuç çıkarmamız gerekirdi. Herhalde, gerçekte gözlemlediğimiz şeye benzer bir şey gözlemlerdik. Belli bir gezegende zeki yaşamın evrilmesi olasılığının düşük olduğu hipotezi kendimizi hayatın erken bir aşamada yok olduğu bir gezegende bulabileceğimizi tahmin etmez; daha ziyade, kendimizi zeki yaşamın evrildiği bir gezegende bulmamız gerektiğini tahmin edebilir, bu tür gezegenler ilkel yaşamın evrildiği tüm gezegenlerin çok küçük bir kısmını oluştursa bile. Dolayısıyla yeryüzündeki hayatın uzun sicil kaydı gezegenimizde yüksek organizmaların ortaya çıkma olasılığının yüksek olduğu –hele hele neredeyse kaçınılmaz olduğu– iddiasını destekleyecek çok az kanıt sunuyor olabilir.[29]

Üçüncüsü, mevcut koşullar cennetimsi olsa bile ve türsel bir iptidai durumdan kaçınılmaz biçimde doğduğu gösterilebilse bile, yine de iyimser eğilimin belirsiz bir geleceğe kadar süreceğinin garantisi olmayacaktır. Bu durum felaket bir yok oluş olanağını yok saysak, hatta evrimsel gelişmelerin giderek artan karmaşıklıkta sistemler üretmeye devam edeceğini varsaysak bile geçerliliğini korur.

Azami verimlilik için seçilmiş olan makine zekâsı işçilerin son derece ağır bir tempoda çalışacaklarını ve ne kadar mutlu olacaklarının bilinmediğini daha önce belirtmiştik. Ayrıca geleceğin rekabetçi dijital yaşam çorbasındaki en uygun yaşam biçimlerinin bilinçli olmayabileceğini de bir olasılık olarak ortaya koymuştuk. Tam bir haz ya da bilinç kaybı dışında, birçok insanın iyi bir yaşam için vazgeçilmez göreceği

diğer vasıfların tükenip gitmesi mümkün olabilir. İnsanlar müziğe, mizaha, aşka, sanata, oyuna, dansa, sohbete, felsefeye, edebiyata, maceraya, keşiflere, yeme-içmeye, arkadaşlığa, aileye, spora, doğaya, geleneğe ve maneviyata vb. değer verirler. Bunlardan herhangi birinin uyum sağlayıp yaşayacağının garantisi yoktur. Belki de uygunluğu azamiye çıkarmak; dur durak bilmeyen çok yoğun bir hamaliye, hiçbir zevk ve heyecan vermeyen, sadece bir ekonomik verimlilik çıtasını tutturmayı amaçlayan sıkıcı ve mükerrer bir çalışma temposundan başka bir şey olmayacaktır. O zaman, seçilen fenotipler yukarıda bahsedilen özelliklerden mahrum yaşamlar olacaktır ve kişinin aksiyolojisine bağlı olarak, sonuç iğrenç, değersiz ya da salt sefillik olarak görülebilir, ama her halükârda insanın övülmeye değer hissedeceği bir ütopyadan çok uzak bir şey olacaktır.

Bu türde karamsar bir tablonun şu an müzik, mizah, aşk, sanat vesaireden zevk alıyor olmamızla nasıl bağdaşabildiği merak edilebilir. Eğer bu davranışlar gerçekten "savurganlık" ise, o zaman türümüzü şekillendiren evrim süreçleri tarafından hoşgörüyle karşılanması, hatta teşvik edilmesi nasıl mümkün olmuştur? Modern insanın evrimsel bir dengesizlik içinde olması bunu açıklamaz, zira Buzul Çağı'ndaki atalarımız da bu israfların çoğundan payına düşeni almıştı. Söz konusu davranışların çoğu *Homo sapiens*'e özgü bile değildir. Şaşaalı gösteriş hayvanlar âlemindeki cinsel seçilimden ulus-devletler arasındaki prestij yarışına kadar çok çeşitli bağlamlarda görülebilir.[30]

Bu davranışların her biri için evrim bağlamında tam bir açıklama sunmak bu kitabın kapsamını aşıyor olsa da, bunlardan bazılarının bir makine zekâ bağlamıyla alakalı olamayabileceğini belirtebiliriz. Örneğin sadece bazı türlerde ve baskın olarak gençlerde görülen oyun, esasen, genç hayvanın ileriki yaşamında ihtiyaç duyduğu yetenekleri öğrenmesinin bir yoludur. Emülasyonlar halihazırda olgun yetenek repertuarına sahip birer yetişkin olarak yaratılabildiğinde ya da tek bir yapay zekânın edindiği bilgi ve teknikler doğrudan başka bir yapay zekâya nakledilebildiğinde, oyunbaz davranış ihtiyacı daha az yaygın olabilir.

Hümanist davranışların diğer birçok örneği bedensel ya da zihinsel metanet, toplumsal statü, müttefiklerin niteliği, bir kavgadan sağ çıkabilme yeteneği ve isteği ya da kaynaklara sahip olma gibi doğrudan gözlemlenmesi zor olan vasıfların sahtesi yapılması zor işaretleri olarak evrilmiş olabilir. Tavuskuşunun kuyruğu klasik örnektir: Sadece uygun tavuskuşları gerçekten abartılı düzeyde tüyler geliştirebilir ve dişi tavuskuşları bunu çekici bulacak şekilde evrilmişlerdir. Davranışsal özellikler, en az morfolojik özellikler kadar, genetik uygunluğun işaretleri olabilir.[31]

Şaşaalı gösterişin hem insanlar hem de diğer türler arasında çok yaygın olduğu düşünüldüğünde, teknolojik açıdan daha ileri yaşam biçimlerinin de repertuarında olup olmayacağı sorulabilir. Oyunbazlığın, müzikselliğin, hatta bilincin gelecekteki zeki bilgi işleme ekolojisinde dar anlamda araçsal bir değeri olmayacak olsa bile, bu

özelliklerin yine de diğer uyum sağlayıcı niteliklerin güvenilir işaretleri olmaları bakımından sahiplerine evrimsel bir üstünlük sağlayacağı düşünülemez mi?

Bizim için değeri olan şey ile geleceğin dijital ekolojisinde uyum sağlayıcı olacak şeyler arasında öncesinde yerleşmiş bir uyum olanağını dışlamak zor olmakla birlikte, şüpheyle yaklaşmak için nedenler vardır. İlkin, doğada bulduğumuz masraflı gösterişlerin çoğunun cinsel seçilimle bağlantılı olduğunu düşünün.[32] Halbuki teknolojik açıdan olgun yaşam biçimleri arasında üreme ağırlıklı olarak ya da sadece cinsel olabilir.

İkincisi, teknolojik açıdan ileri aracılar kendileri hakkında güvenilir şekilde bilgi aktarmanın yeni araçlarına sahip olabilirler ve bu araçlar pahalı gösteriş gerektirmez. Bugün bile, profesyonel kreditörler kredi verip vermeyeceklerini değerlendirirken şık takım elbise ve Rolex saat gibi pahalı gösterişlerden ziyade sahiplik belgeleri ve banka hesapları gibi belgesel kanıtlara bakıyorlar. Gelecekte, bir aracı müşterinin sahip olduğunu iddia ettiği özelliği davranışsal kayıtların ayrıntılı incelenmesi yoluyla doğrulayan, simülasyonlu çevrelerde testler yapan ya da kaynak kodunu doğrudan teftiş eden denetim şirketleriyle çalışmak mümkün olabilir. Kendi niteliklerini bu tür bir denetimi kabul ederek göstermek şaşaalı gösteriş yoluyla sergilemekten daha verimli olabilir. Bu türde profesyonel olarak dolayımlı bir işaretin *sahtesini üretmek* yine de masraflı olacaktır (işareti güvenilir kılan asli özellik budur), ama *doğruyu söylemek* gösterişli bir yalan söylemekten çok daha ucuz olabilir.

Üçüncüsü, mümkün olan masraflı gösterişlerin hepsi içsel olarak değerli ya da toplumsal açıdan istenir değildir. Birçoğu tek kelimeyle israftır. Rakip şefler arasında bir statü rekabeti biçimi olan Kwakiutl potlaç merasimlerinde muazzam bir servet halkın gözleri önünde yok edilirdi.[33] Rekora imza atan gökdelenler, megayatlar ve ay roketleri bunun günümüzdeki benzerleri olarak görülebilir. Müzik ve mizah gibi faaliyetlerin insan yaşamının değerini içsel olarak artırdığını iddia etmek mümkün olmakla birlikte, benzer bir iddiayı moda aksesuarlar ve benzer tüketim çılgınlığı statü sembolleri gibi masraflı alışkanlıklar için dile getirmek tartışmalıdır. Daha kötüsü, masraflı gösterişler doğrudan zararlı olabilir (çete şiddeti ya da askeri kabadayılığa yol açan maçoluk gibi). Dolayısıyla geleceğin zekâ sahibi yaşam biçimleri masraflı işaretleşmelerden yararlansalar bile, işaretin değerli türde olup olmayacağı –bir baykuşun yüksek sesli melodisi gibi mi yoksa karakurbağasının tekheceli vıraklaması (ya da kuduz bir köpeğin durmadan havlaması) gibi mi olacağı– tartışmaya açıktır.

Geçiş sonrasında bir tekel kurulması mı?

Makine zekâya geçişin dolaysız sonucu çokkutuplu olsa bile, ileride bir tekel kurulması olasılığı devam eder. Bu tür bir gelişme uzun vadeli bir eğilim olarak daha büyük ölçeklerdeki siyasal bütünleşmeyi sürdürecek ve doğal sonucuna götürecektir.[34] Bu nasıl gerçekleşebilir?

İkinci bir geçiş

Başlangıçtaki bir çokkutuplu sonucun geçiş sonrası tekele varmasının yollarından biri, ilk baştaki geçişin ardından, geride kalan güçlerden birine belirleyici bir stratejik üstünlük sağlayacak kadar büyük olan ikinci bir teknolojik geçişin olmasıdır: O zaman, bu güç bir tekel kurma fırsatını elde edebilir. Bu türden varsayımsal bir ikinci geçiş daha yüksek bir süper zekâ düzeyine atılımla sağlanabilir: Sözgelimi eğer ilk dalga makine süper zekâsı emülasyon temelli olursa, ikinci dalga şu anda araştırmayı yapan emülasyonlar kendi kendini ilerletebilen yapay zekâlar geliştirdiğinde başarıya ulaşacaktır.[35] (Alternatif olarak, ikinci geçişi nanoteknolojideki bir atılım ya da henüz öngörülmemiş askeri ya da genel amaçlı bir başka teknoloji tetikleyebilir.)

İlk geçişten sonraki gelişimin temposu son derece hızlı olacaktır. Egemen güç ile tek yakın rakibi arasındaki en ufak bir boşluk bile ikinci geçiş evresinde egemen güce belirleyici stratejik üstünlük kazandırabilir. Örneğin iki projenin ilk geçişe sadece birkaç gün arayla girdiklerini ve havalanma çok yavaş olduğundan, aradaki boşluğun egemen projeye havalanma sırasındaki herhangi bir noktada belirleyici bir stratejik üstünlük kazandırmadığını varsayalım. İki proje birlikte süper zeki güçler olarak ortaya çıkar, ama içlerinden biri diğerinin birkaç gün önünde olabilir. Ama şu an makine süper zekâsı için tipik olan araştırma zaman ölçeklerinde gerçekleşen gelişmeler var, bunlar belki de biyolojik insan zaman ölçeğinde yürütülen araştırmalardan binlerce, hatta milyonlarca kez daha hızlıdır. Dolayısıyla ikinci geçiş dönemi teknolojisinin gelişimi günler, saatler, hatta dakikalar içinde tamamlanabilir. Önde gidenin önderliği sadece birkaç gün olsa da, atılım belirleyici bir stratejik üstünlüğe dönüşebilir. Gelgelelim teknolojik dağılma (casusluk ya da diğer kanallar yoluyla) teknolojik gelişme kadar hızlanırsa, bu etkinin ortadan kalkacağını belirtelim. O zaman geriye kalan ikinci geçişin keskinliği, yani ilk geçişten sonraki dönemin olaylarının genel hızıyla karşılaştırıldığında hangi hızda gerçekleşeceği olacaktır. (Bu anlamda, ilk geçişten sonra gelişmeler ne kadar hızlı olursa, ikinci geçişin keskinliği o kadar az olacaktır.)

Keza belirleyici stratejik üstünlüğün bir tekel kurmak için kullanılmasının, ikinci ya da daha sonraki bir geçiş sırasında ortaya çıkması durumunda daha muhtemel olduğu da düşünülebilir. İlk geçişten sonra karar alıcılar ya süper zeki olacaktır ya da bir süper zekâdan tavsiyeye erişim olanağına sahip olacaktır ve bu da mevcut stratejik seçeneklerin içerimlerini netleştirecektir. Ayrıca ilk geçişten sonraki durum potansiyel rakiplere karşı proaktif müdahalenin saldırgan için daha tehlikesiz olacağı bir durum olacaktır. Eğer karar alan zihinler ilk geçişten sonra dijitalse, kopyalanmaları ve saldırılara karşı daha savunmasız hale gelmeleri mümkündür. Bir savunmacı misilleme saldırılarında saldırgan nüfusun onda dokuzunu öldürebilecek güce sahip olsa bile, bu durum ölenler hemen diriltilebildiği oranda çok bir sıkıntı yaratmaz. Altyapının mahvedilmesi (gerçi yeniden imar edilebilir) fiilen sınırsız ömürleri olan dijital zihinler

için sorun teşkil etmez, zira muhtemelen kaynaklarını ve nüfuzlarını kozmolojik bir zaman ölçeğinde azamiye çıkarmayı planlıyor olacaklardır.

Süper organizmalar ve ölçek ekonomileri

Eşgüdümlü insan toplamlarının (şirketler ya da uluslar gibi) büyüklüğü bir tarihsel çağdan diğerine değişiklik gösterebilecek çeşitli parametrelerden (teknolojik, askeri, parasal ve kültürel) etkilenir. Bir makine zekâsı devrimi bu parametrelerin birçoğunda köklü değişiklikler gerektirecektir. Belki de bu değişiklikler bir tekelin kurulmasını kolaylaştıracaktır. Bu muhtemel değişikliklerin ne olduğuna ayrıntılı şekilde bakmadan karşıt olanağı (değişikliklerin birleşmeden ziyade parçalanmayı kolaylaştırması) dışlayamasak da, burada karşılaştığımız artan çeşitliliğin ya da belirsizliğin bir tekelin ortaya çıkma ihtimaline daha fazla itimat etmek için bir temel olabileceğini belirtebiliriz. Bir makine zekâsı devrimi, deyim yerindeyse, işleri alevlendirir: Diğer türlü belki de muhtemel görünmeyebilecek olan jeopolitik saflaşmaları ve yeniden saflaşmaları mümkün kılacak şekilde kartlar yeniden karılabilir.

Siyasal bütünleşmenin ölçeğini etkileyebilecek tüm etkenlerin enikonu tahlili bu kitabın kapsamını bir hayli aşacaktır: İlintili siyaset bilimi ve iktisat yazınına dair bir değerlendirme başlı başına ayrı bir kitabın konusudur. Biz burada birkaç etkene, kontrolün merkezileştirilmesini kolaylaştırabilecek aracıların dijitalleştirilmesinin veçhelerine kısaca değinmekle yetinelim.

Carl Shulman'a göre, bir emülasyon nüfusunda seçilim baskıları "süperorganizmalar"ın, yani klanlarının iyiliği için kendilerini feda etmeye hazır emülasyon gruplarının ortaya çıkmasına yardımcı olur.[36] Süper organizmalar, üyeleri kendi çıkarına bakan örgütlerin baş belası olan aracılık problemlerinden vareste olacaktır. Bedenlerimizdeki hücreler gibi ya da bir gerçek sosyal [eusocial] böcek kolonisindeki tekil hayvanlar gibi kendi kopya kardeşlerine karşı tamamen diğerkâm olan emülasyonlar, incelikli teşvik şemalarının yokluğunda bile birbirleriyle işbirliği halinde olacaklardır.

Eğer bireysel emülasyonların anlaşmalı silinmesine (ya da süresiz ertelenmesine) izin verilmezse, süperorganizmalar bilhassa güçlü bir üstünlüğe sahip olacaklardır. Kendini korumakta ısrar eden emülasyonların çalıştırıldığı şirketler ya da ülkeler eski ya da fazla işçileri idame ettirmek için sonu gelmeyen bir taahhüt altına sokulacaklardır. Halbuki emülasyonları hizmetleri sona erdiğinde kendilerini isteyerek silen örgütlerin talepteki dalgalanmalara daha kolay uyum sağlamaları gerekmeyecektir ve özgürce deneyler yapıp işçilerinin çeşitlerini çoğaltarak sadece en üretken olanları koruyacaklardır.

Eğer gönülsüz silmeye izin *verilmezse*, o zaman gerçek sosyal emülasyonların karşılaştırmalı üstünlüğü azalır, ama herhalde ortadan kalkmaz. İşbirliğine ve kendini feda etmeye açık aracıların işverenleri tüm örgütte aracılık problemlerinin azaltılmasından (emülasyonlar kendilerinin silinmesine karşı gösterebilecekleri her türlü direnişi

bastırmak zorunda kalma belasından kurtulmak da dahil) hâlâ daha verimlilik elde edebilir. Genel olarak, kendi bireysel yaşamlarını kamu yararı için feda etmeye hazır olan işçilere sahip olmanın üretkenlik açısından kazanımları, bir örgütün kendisine fanatikçe bağlı üyelere sahip olarak elde edebileceği faydaların özel bir örneğidir. Bu tür üyeler örgüt için diri diri mezara girip az paraya uzun süre çalışmakla kalmazlar; aynı zamanda işyerinde huzursuzluk çıkarmayıp örgütün çıkarı için en iyisi neyse istikrarlı bir şekilde onu yapmaya çalışarak denetleme ve bürokratik kısıtlama ihtiyacını azaltırlar.

Eğer bu tür bir adanmışlığı elde etmenin yegâne yolu üyeliği kopya kardeşlerle sınırlamaksa (böylece belli bir süper organizmadaki tüm emülasyonlar aynı kalıptan çıkarılır), o zaman süper organizmalar rakip örgütlerinkinden daha dar bir yetenek havuzundan istifade etmenin zararını göreceklerdir, ama bu zarar iç aracılık problemlerinden kaçınmanın yararları yanında önemsiz kalacaktır.[37] Eğer bir süper organizma en azından farklı eğitimde üyeleri içerebilirse, bu zarar ciddi ölçüde azalır. Tüm üyeler tek bir ilk-kalıptan çıkarılsa bile, işgücü yine de çeşitli yeteneklerle katkı sunabilir. Allame derecesinde yetenekli bir emülasyon ilk-kalıpla başlanır ve ardından gelenler farklı eğitim programlarına ayrılabilir; kopyalardan biri muhasebe, diğeri elektrik mühendisliği vb. öğrenir. Böylece ortaya farklı yeteneklerde olmasa da farklı becerilerde üyeler çıkar. (Azami çeşitlilik birden fazla ilk-kalıbın kullanılmasını gerektirebilir.)

Bir süper organizmanın asli özelliği tek bir atanın kopyalarından oluşması değil, içindeki tüm bireysel aracıların tamamen ortak bir hedefe bağlı olmasıdır. Dolayısıyla bir süper organizma yaratabilmek kontrol problemine kısmi bir çözüm gerektiriyor olarak görülebilir. Kontrol problemine tamamen genel bir çözüm kişinin keyfi bir nihai hedefi olan bir aracı yaratmasını sağlayacakken, bir süper organizmanın yaratılması için gereken kısmi çözüm sadece aynı nihai hedefi olan (entipüften olmayan ama mutlaka keyfi denemeyecek bir nihai hedef için) çoklu aracıların yaratılabilmesini gerektirir.[38]

Dolayısıyla bu alt kısımda ortaya konan esas mesele tek klonlu emülasyon gruplarıyla sınırlı olmayıp, çeşitli çokkutuplu makine zekâ senaryolarına uygulandığını açık hale getirecek tarzda daha genel olarak ifade edilebilir. Mesele şu ki, aktörler dijital olduğunda mümkün olabilecek güdülenme seçilimi tekniklerindeki belli türde ilerlemeler, şu anda büyük insan örgütlerine köstek olan ve ölçek ekonomilerine sorun teşkil eden verimsizliklerden bazılarını aşmaya yardımcı olabilir. Bu sınırlar kalktığında örgütler (ister şirket ister ulus ister diğer ekonomik ya da siyasal varlıklar olsun) büyüyebilirler. Geçiş sonrasında bir tekelin ortaya çıkışını kolaylaştırabilecek bir etken varsa o da budur.

Süper organizmaların (ya da kısmen seçilmiş güdülenmeleri olan diğer dijital aracıların) mükemmelleşebileceği alanlardan biri zorlamadır. Bir devlet polis, asker, istihbarat hizmeti ve kamu idaresinin tamamen sadık olmasını sağlamak için güdülenme seçilimi yöntemlerinden yararlanabilir. Shulman'ın belirttiği gibi:

> [Dikkatle hazırlanmış ve doğrulanmış bir sadık emülasyonun] kurtarılmış devletleri ideolojik açıdan tekbiçim bir askeri, bürokratik ve polis gücünün kadrosunu doldurmak üzere milyarlarca kez kopyalanabilir. Kısa bir çalışma döneminin ardından, her kopyanın yerini aynı kurtarılmış devletin taze bir kopyası alacak ve böylece ideolojik çekişme engellenecektir. Belli bir yargı yetkisi dahilinde, bu yeterlilik inanılmaz derecede ayrıntılı gözlem ve düzenlemeye izin verebilir: Diğer her bir mukim için böyle bir kopya olabilir. Bu sayede kitle imha silahlarının geliştirilmesi engellenebilir, beyin emülasyon deneylerine ya da yeniden üretimlere ilişkin bir mevzuat yürürlüğe konabilir, liberal demokratik bir anayasa yürürlüğe koyulabilir ya da dehşetengiz ve kalıcı bir totalitarizm kurulabilir.[39]

Bu tür bir yeterliliğin birinci mertebeden etkisi iktidarın pekiştirilmesi ve muhtemelen az sayıda elde temerküz etmesi olacak gibi görünüyor.

Antlaşmayla birleşme

Geçiş sonrası bir çokkutuplu dünyada uluslararası işbirliğinden elde edilecek büyük potansiyel kazanımlar vardır. Savaşlardan ve silahlanma yarışlarından sakınılabilir. Astrofiziksel kaynaklardan küresel olarak optimum hızda yararlanılabilir. Aceleye getirmekten kaçınmak ve yeni tasarımların tamamen incelenmesine imkân sağlamak adına daha ileri makine zekâ biçimlerinin geliştirilmesi eşgüdümlü hale getirilebilir. Varoluşsal riskler sunabilecek diğer gelişmeler ertelenebilir. Ayrıca küresel ölçekte tekbiçim yönetmelikler yürürlüğe koyulabilir; bunların arasında güvence altında olan bir yaşam standardı (bunun için belli bir nüfus kontrolüne gerek duyulacaktır) ve emülasyonlar ile diğer dijital ve biyolojik zihinlerin sömürüsünü ve istismarını engellemek için kanun hükümleri de olacaktır. Ayrıca kaynakları karşılanabilir tercihlere sahip aracılar (bu konu 13. Bölüm'de daha ayrıntılı ele alınacak), hiçbir şey elde etmeme riskini alacakları ve kazananın her şeyi kazanacağı bir mücadeleden ise geleceğin belli bir parçasını güvence altına alacak bir ortak anlaşmayı tercih edeceklerdir.

Gelgelelim işbirliğinin sağlayacağı büyük potansiyel kazanımların varlığı gerçekten işbirliği yapılacağı anlamına gelmez. Günümüz dünyasında küresel eşgüdümün daha iyi hale getirilmesiyle muazzam kazanımlar elde edilebilir: Askeri harcamaların, savaşların, aşırı balıkçılığın, ticaret duvarlarının ve hava kirliliğinin azaltılması bunlardan birkaçıdır. Fakat bu iri meyveler dalında çürümeye terk ediliyor. Neden böyle? Ortak iyiliği azamiye çıkaracak olan tamamen işbirliğine dayalı bir sonucun önündeki engel nedir?

Engellerden biri yapılabilecek anlaşmalara uyulmasını sağlamanın zorluğudur (izleme ve yaptırım masrafları dahil). İki nükleer rakibin ikisi de atom bombalarını bırakmaktan daha kârlı çıkabilir; ama bu konuda ilkesel bir anlaşmaya varsalar bile, yine de silahsızlanma diğer tarafın aldatabileceği korkusundan ötürü kâğıt üzerinde kalabilir. Bu korkuyu dindirmek bir doğrulama mekanizması kurulmasını gerektirir.

Silahlanma programının yeniden kurulmadığından emin olmak için mevcut stokların yok edilmesini denetleyecek ve ardından nükleer reaktörleri vb. izleyecek ve teknik-beşeri istihbarat toplayacak müfettişlere ihtiyaç duyulabilir. Masraf kalemlerinden biri bu müfettişlere yapılacak ödemedir. Bir diğer kalem ise müfettişlerin casusluk yapıp ticari ya da askeri sırları sızdırmaları riskidir. Belki en önemlisi, her iki taraf da diğerinin gizliden gizliye nükleer silahlar geliştirdiğinden korkabilir. Sonuçta, iki tarafın da anlaşmaya uyduğunu teyit etmek çok zor olduğundan, potansiyel açıdan birçok yararlı anlaşma hayata geçirilemiyor.

Eğer izleme masraflarını azaltan yeni teftiş teknolojileri mevcut olursa, bunun işbirliğini artırması beklenebilir. İzleme masraflarının geçiş sonrası dönemde azalıp azalmayacağı tamamen açık değildir. Birçok yeni ve güçlü teftiş teknikleri olacağı kesin olsa da yeni saklama araçları da olacaktır. Özellikle de düzenleme isteyecek faaliyetlerin artan bir kısmı siber-uzayda, yani fiziksel denetimden uzakta gerçekleşecektir. Örneğin yeni nanoteknoloji silah sistemleri ya da yeni nesil yapay zekâ tasarımında çalışan dijital zihinler bunu fazla bir fiziksel ayak izi bırakmadan yapabilirler. Dijital adli tıp, anlaşmayı ihlal edenin gayri meşru faaliyetlerini gizleyebileceği saklamanın ve şifrelemenin tüm katmanlarına sızmayı başaramayabilir.

Güvenilir bir yalan detektörü –eğer geliştirilebilirse– anlaşmaya uyulup uyulmadığını izlemek için son derece yararlı bir alet olabilir.[40] Bir teftiş protokolünde kilit görevlilerle görüşme hükümleri yer alabilir; amaç, anlaşmanın tüm hükümlerinin uygulanmasına kararlı olup olmadıklarını ve epey çaba harcamalarına karşın hiçbir ihlalin söz konusu olmadığını doğrulamaktır.

Aldatmayı planlayan bir karar alıcı, öncelikle astlarına gayri meşru faaliyete girmeleri ve bu faaliyeti bizzat karar alıcıdan bile gizlemeleri, ardından bu dalaverelere giriştiklerine dair izleri hafızalarından silecek bir prosedürden geçmeleri emrini vererek bu türden doğrulama şemasına dayalı bir yalan detektörünü alt edebilir. Uygun şekilde hedef alınmış hafıza silme işlemleri daha ileri nöroteknolojiye sahip biyolojik beyinlerde pekâlâ mümkün olabilir, hatta mimarilerine bağlı olarak makine zekâlarda daha kolay bile olabilir.

Devletler bu problemi devletin imzaladığı ya da gelecekte imzalayabileceği anlaşmaları baltalamak ya da çökertmek gibi bir niyetleri olup olmadığını kontrol etmek üzere kilit görevlileri düzenli olarak bir yalan detektörüyle test eden bir izleme şemasına bağlı kalarak aşmaya çalışabilirler. Bu tür bir bağlılık, diğer anlaşmaların doğrulanmasını kolaylaştıracak bir tür üst-anlaşma olarak görülebilir; ama devletler güvenilir bir müzakere partneri olarak görülmenin yararını tek yanlı şekilde elde etmek için kendilerini buna bağlayabilirler. Fakat bu bağlılık ya da üst-anlaşma yukarıda bahsi geçen türde bir atama ve unutma dalaveresi yoluyla baltalanma probleminin aynısıyla karşılaşabilir. İdeal olarak, üst-anlaşma taraflardan herhangi birisi bunun uygulanmasını

baltalamak için gerekli iç düzenlemeleri yapma fırsatı elde etmeden *önce* uygulamaya konulacaktır. Kötülük bir boşluk yakalayıp aldatma tohumlarını bir kez ektiğinde, o toprakta güven bir daha asla tutmaz.

Bazı durumlarda anlaşma ihlallerini *tespit edebilmek* bir anlaşma için gerekli güvenin tesis edilmesi için tek başına yeterlidir. Fakat diğer örneklerde şartlara riayet edilmesini *zorlamak* ya da bir ihlal durumunda ceza vermek için bir mekanizmaya ihtiyaç vardır. Eğer yanlış yapılan tarafın anlaşmadan çekilmesi tehdidi ihlalleri engellemek için yeterli değilse, sözgelimi anlaşmayı ihlal eden taraf sonrasında diğer tarafın nasıl yanıt verdiğini umursamayacağı bir üstünlük elde ederse, bir zorlama mekanizmasına ihtiyaç duyulabilir.

Eğer güdülenme seçilimi için son derece etkili yöntemler mevcutsa, bağımsız bir aracıya imzacılardan bir ya da birkaçının karşı çıkması pahasına anlaşmayı zorla dayatacak kadar askeri ya da polisiye güç tevdi edilerek bu zorlama problemi çözülebilir. Bu çözüm dayatma aracısına güvenilmesini gerektirir. Ama yeterince iyi güdülenme seçilimi teknikleri olduğunda, anlaşmanın tüm taraflarının bir dayatma aracısının tasarımını ortaklaşa denetlemesiyle gerekli güven tesis edilebilir.

Gücü dışsal bir dayatma aracısına vermek tekkutuplu sonuca (ilk makine zekâsı devrimi öncesinde ya da sırasında bir tekelin doğduğu sonuca) dair değerlendirmelerimizde karşılaştığımız meselelerin birçoğunu gündeme taşır. Rakip devletlerin hayati güvenlik çıkarlarına ilişkin antlaşmaları dayatabilmek için, dış dayatma aracısının fiilen bir tekel kurması gerekecektir: Küresel bir süper zeki Leviathan. Fakat tek bir fark var, o da şu anda bu Leviathan'ı yaratacak aracıların biz insanların şu an sahip olduğundan daha fazla yeterliliğe sahip olacağı bir geçiş sonrası durumu ele alıyor olmamızdır. Bu Leviathan yaratıcıları zaten bizzat süper zeki olabilirler. Bu durum kontrol problemini çözüp, yaratılmasında söz sahibi olan tüm tarafların çıkarlarına hizmet edecek bir dayatma aracısı tasarlama ihtimalini epey artıracaktır.

Riayet edilip edilmediğini izlemenin ve zorlamada bulunmanın masrafları bir tarafa, küresel eşgüdümün önünde başka engeller var mıdır? Belki de geriye kalan en önemli mesele *pazarlık maliyetleri* adını verebileceğimiz şeydir.[41] Dahili olan herkese yarar sağlayacak bir anlaşma olanağı olduğunda bile, taraflar ganimeti nasıl bölüşecekleri konusunda anlaşmayı başaramadıklarında, anlaşma ölü doğar. Örneğin iki insan bir dolar kâr getirecek bir anlaşma yapar, ama iki taraf da altmış sent hak ettiğini düşünür ve daha azını kabul etmezse, anlaşma gerçekleşmez ve potansiyel kazançtan feragat edilir. Genel olarak, müzakereler taraflardan birinin yaptığı stratejik pazarlık tercihlerinden ötürü çetin ya da uzun olabilir veya tamamen sonuçsuz kalabilir.

Gerçek yaşamda insanlar stratejik pazarlık olanağına karşın sık sık anlaşmaya varmayı başarırlar (gerçi epey bir zaman harcayıp sabretme pahasına). Fakat strate-

jik pazarlık problemlerinin geçiş sonrası dönemde farklı bir dinamiğe sahip olacağı düşünülebilir. Bir yapay zekâ müzakerecisi formel bir aklilik anlayışına daha tutarlı şekilde bağlı kalabilir ve diğer yapay zekâ müzakerecileriyle eşleştiğinde bunun yeni ya da beklenmedik sonuçları olması mümkündür. Ayrıca bir yapay zekâ insanların sahip olmadığı ya da hayata geçirmelerinin çok ama çok daha zor olduğu bazı pazarlık hamlelerine sahip olabilir (bir yol haritasına ya da politikaya ön taahhütte bulunmak dahil). İnsanlar (ve insanların işlettiği kurumlar) ara ara ön taahhütte bulunabilirlerse de (gerçi güvenilirlik ve özgüllük dereceleri kusurludur), bazı makine zekâ türleri birtakım keyfi, çiğnenemez ön taahhütlerde bulunabilir ve müzakereci partnerlere bu tür bir ön taahhütte bulunulduğunu doğrulama izni verebilirler.[42]

Güçlü ön taahhüt tekniklerinin varlığı müzakerelerin doğasını kökten değiştirebilir ve ilk hamle üstünlüğüne sahip olan aracıya muazzam bir avantaj sağlayabilir. Eğer belli bir aracının katılımı işbirliğinden yana muhtemel kazanımların gerçekleştirilmesi için zorunluysa ve eğer bu aracı ilk hamleyi yapabiliyorsa, kendisine sözgelimi artı-değerin yüzde 99'undan daha azını veren hiçbir anlaşmaya ön taahhütte bulunmayarak ganimetin bölüşülme şartlarını dayatacak bir konumda olacaktır. Bu durumda, diğer aracılar, adaletsiz teklifi reddederek hiçbir şey almama ya da boyun eğerek değerinin yüzde 1'ini alma seçeneğiyle karşı karşıya geleceklerdir. Eğer ilk hamleyi yapan aracının ön taahhüdü açıkça doğrulanabilirse, müzakereci partnerler yegâne iki seçeneklerinin bunlar olduğundan emin olabilirler.

Aracılar bu tarzda sömürülmekten kaçınmak için şantajı ve her türlü adaletsiz teklifi reddetme konusunda ön taahhütte bulunabilirler. Bu türde bir ön taahhüt bir kez yapıldığında (ve başarılı şekilde duyurulduğunda), diğer aracılar tehditlerde bulunmayı ya da sadece kendi çıkarlarına olan anlaşmaları kabul etme konusunda ön taahhütlerde bulunmayı çıkarlarına uygun görmeyeceklerdir, çünkü tehditlerin sökmeyeceğini ve adaletsiz tekliflerin reddedileceğini bileceklerdir. Ama bu da bir kez daha ilk hamlede bulunmanın üstünlüğe sahip olduğunu gösteriyor. İlk harekete geçen aracı salt başkalarının adaletsiz bir üstünlük elde etmesini engellemek adına güçlü konumundan mı yararlanacağını yoksa müstakbel ganimetlerden aslan payına mı uzanacağını seçebilir.

En iyi konumda olanın, şantajlara, hatta kendisinin katılımının vazgeçilmez olduğu ama neredeyse tüm kazanımları elde etmediği her türlü anlaşma teklifine kapalı olduğu bir değer sistemi ya da bir mizaçla yola koyulan aracı olduğu düşünülebilir. Bazı insanlar, şimdiden, uzlaşmaz ruhta çeşitli veçhelere tekabül eden kişilik özelliklerine sahip gözüküyorlar.[43] Gelgelelim etrafta hakları olan paydan daha fazlasını hak ettiklerini düşünen ve geri adım atmamaya kararlı olan başka aracıların var olduğu ortaya çıkarsa gergin mizaç ters tepebilir. Bu durumda, durdurulamaz kuvvet hareket ettirilemez bir nesneyle karşılaşacak ve sonuç olarak anlaşmaya varılamayacaktır veya

daha da beteri, topyekûn savaş olacaktır. Uysal ve zayıf iradeli taraf, hakkı olan paydan daha az olsa da en azından bir şeyler elde edecektir.

Bu türden bir geçiş sonrası pazarlık oyununda ne tür bir oyun-teorik dengesine varılacağı ilk başta anlaşılmayabilir. Aracılar burada ele alınanlardan daha karmaşık stratejileri tercih edebilirler. Bir Schelling noktası olarak işlev gösterecek, bir hakkaniyet normuna yaslanan bir dengeye varılacağı *umut edilecektir*: Ortak beklentilerden ötürü diğer açılardan yeterince belirlenmemiş bir eşgüdüm oyununda muhtemel bir eşgüdüm noktası olacak bir büyük sonuç uzamında göze çarpan bir özellik. Bu tür bir denge, evrimle gelen temayüllerimizin ve kültürel programlamamızın bir kısmı sayesinde güçlendirilebilir: Hakkaniyetin herkesçe tercih edilmesi değerlerimizi bir geçiş sonrası çağına nakletmeyi başardığımızı varsayarsak, beklentileri ve stratejileri çekici bir dengeye yol açacak şekillerde yönlendirebilir.[44]

Her halükârda, güçlü ve esnek ön bağlılık biçimleri mümkün olduğunda müzakerelerin sonuçları beklenmedik bir kılığa girebilir. Geçiş sonrası çağ çokkutuplu başlasa bile, küresel çaptaki önemli tüm eşgüdüm problemlerini çözen bir anlaşmanın sonucu olarak neredeyse hemen ortaya çıkacak olabilir. Bazı cari masraflar (belki bunlara izleme ve dayatma maliyetleri de dahil olacaktır) ileri makine zekâlarının yararlanabileceği yeni teknolojilerle birlikte dibe vurabilir. Diğer maliyetler, özellikle de stratejik pazarlıkla ilintili maliyetler, kayda değer düzeyde kalabilir. Ama stratejik pazarlık ulaşılan anlaşmanın doğasını nasıl etkilerse etkilesin, eğer bir anlaşmaya varılacaksa anlaşmaya varılmasını neden uzun süre geciktireceğine dair açık bir sebep yoktur. Eğer hiçbir anlaşmaya varılamazsa, o zaman bir kavga çıkabilir ve bu durumda ya taraflardan biri kazanır ve yenen güçbirliğinin etrafında bir tekel kurulur ya da sonuç bir tekelin asla kurulamadığı ve genel sonucun insanlığın ve torunlarının daha eşgüdümlü ve işbirlikçi bir tarzda hareket etselerdi ulaşılabilecek ve ulaşılması gereken çıtanın çok altında kalabileceği sonu gelmez bir çatışma olabilir.

* * *

Çokkutupluluğun, istikrarlı bir biçimde elde edilse bile, çekici bir sonucu garanti altına almayacağını görmüş olduk. Özgün asli-vekil problemi çözümsüzlüğünü koruyor ve bunu geçiş sonrası küresel eşgüdüm başarısızlıklarıyla alakalı yeni birtakım problemlerin altına gömmek durumu daha da kötüleştirmekten başka bir işe yaramaz. Bu nedenle tek bir süper zeki yapay zekâyı güvenle nasıl tutabileceğimiz sorusuna geri dönelim.

Değerler Edinmek

Yeterlilik kontrolü en iyi haliyle geçici ve yardımcı bir önlemdir. Plan süper zekâyı ilelebet şişede tutmak değilse, güdülenme seçilimini hatmetmek zorunlu hale gelecektir. Ama bir yapay aracıya nihai değeri olarak görüp peşine düşebileceği bir değer nasıl aşılanabilir? Aracı zeki değilken, insani açıdan anlamlı bir değeri anlama, hatta temsil etme yeterliliğinden mahrum olabilir. Ama eğer bu işlemi aracı süper zeki olana kadar ertelersek, bizim onun güdülenme sistemine karışmamıza direnebilir ve 7. Bölüm'de gösterdiğimiz gibi, bunu yapması için yakınsak araçsal nedenlere sahip olacaktır. Bu değer yükleme problemi zordur, ama yüzleşilmesi gerekir.

Değer yükleme problemi

Bir süper zekânın kendisini içinde bulabileceği tüm durumları sıralamak ve her biri için hangi eyleme girişeceğini saptamak olanaksızdır. Benzer şekilde, olanaklı tüm dünyaların bir listesini çıkarıp her birine bir değer atfetmek de olanaksızdır. Üç taş oyunundan çok daha karmaşık olan her alanda kapsamlı sıralamanın uygulanabilir olması için çok fazla olanaklı durum (ve durum tarihleri) vardır. Dolayısıyla bir güdülenme sistemi kapsamlı bir arabul çizelgesi olarak görülemez. Bunun yerine, daha soyut şekilde, aracıya verili bir koşulda ne yapacağını kararlaştırma izni veren bir formül ya da kural olarak ifade edilmelidir.

Bu tür bir karar kuralını saptamanın formel yollarından biri fayda işlevidir. Bir fayda işlevi (1. Bölüm'den hatırlayacağımız üzere) doğabilecek her bir sonuca veya daha genel olarak, her bir "mümkün dünya"ya değerler atfeder. Bir fayda işlevi söz konusu olduğunda, beklenen faydayı azamiye çıkaran bir aracı tanımlanabilir. Bu tür bir aracı her seferinde beklenen faydası en fazla olan eylemi seçer. (Beklenen fayda, mümkün olan her dünyanın faydasını bu dünyanın öznel olasılığını girişilen tikel bir eyleme koşulla gerçek dünya olmasının öznel olasılığına göre tartarak hesaplanır. Gerçekte, mümkün olan sonuçlar bir eylemin beklenen faydasını tam olarak hesaplamayı engelleyecek kadar çoktur. Gelgelelim karar kuralı ve fayda işlevi birlikte, bir aracının kestirebileceği şekilde tasarlandığı normatif bir ideal (optimumluk kavramı) belirler ve aracının zekileşmesiyle birlikte yakınlaşma daha da artar.[1] Kendisi için mevcut

olan eylemlerin beklenen faydasına dair iyi bir kestirimi hesaplayabilecek bir makine yaratmak tümüyle bir yapay zekâ problemidir.[2] Bu bölümde başka bir problem ele alınıyor: Makineleri zeki hale getirme problemi çözülse bile devam edecek bir problem.

Bu fayda azamileştirici aracı çerçevesinden yararlanarak, kontrol problemini yapay zekâya insani açıdan akla yatkın olan değerli bir sonuca denk düşecek bir nihai hedef bahşederek çözmeyi amaçlayan geleceğin çekirdek yapay zekâ programlayıcısının durumunu düşünebiliriz. Programlayıcının aklında yapay zekânın teşvik etmesini istediği belli bir insani değer vardır. Somut olmak adına, gelin buna mutluluk diyelim. (Eğer adalet, özgürlük, şan şöhret, insan hakları, demokrasi, ekolojik denge ya da kendini geliştirmeyle de ilgilenseydik benzer meseleler doğacaktı.) Programlayıcı, beklenen fayda çerçevesi bakımından, barındırdıkları mutluluk miktarıyla orantılı olarak olanaklı dünyalara fayda tahsis eden bir fayda işlevi aramaktadır. Peki, bu tür bir fayda işlevini bilgisayar koduyla nasıl ifade edebilir? Bilgisayar dillerinde temel olarak "mutluluk" gibi kelimeler yer almaz. Eğer böyle bir kelime kullanılacaksa, öncelikle tanımlanması gerekir. Bu kelimeyi diğer üst düzey insani kavramlar üzerinden tanımlamak yeterli değildir: "Mutluluk insan doğasına içkin olan potansiyellerin zevkine varılmasıdır" ya da buna benzer bir felsefi açımlama işe yaramaz. Tanım, yapay zekânın programlama dilinde ve nihayetinde bireysel bellek yazmaçlarının içeriklerine işaret eden matematiksel işleçlerde ve alıcılarda gözüken terimlerde en aşağı düzeye inmelidir. Problemi bu perspektiften düşündüğümüzde, programlayıcının görevinin zorluğunu anlamaya başlayabiliriz.

Kendi nihai hedeflerimizi saptamak ve kodlamak zordur, çünkü insanın hedef temsilleri karmaşıktır. Fakat karmaşıklık çoğunlukla bizim için şeffaf olduğundan, çoğu zaman orada olduğunu gözden kaçırırız. Bu durumu görme algısına benzetebiliriz. Görme de basit bir şey gibi görünebilir, çünkü bu işlemi çaba harcamadan gerçekleştiririz.[3] Görünen o ki, tek yapmamız gereken gözlerimizi açmaktır, o zaman etrafımızdaki çevrenin zengin, anlamlı, eidetik, üçboyutlu bir görüntüsü zihnimize akmaya başlar. Görmenin bu sezgisel kavranışı bir dükün ataerkil evine ilişkin kavrayışına benzer: Dük söz konusu olduğu kadarıyla, şeyler hep uygun zamanda ve mekânda kendini gösterirken, bu tezahürleri üreten mekanizma gözden ıraktır. Ama en basit görme görevini bile yerine getirmek (mesela mutfakta tuzluğu bulmak) muazzam bir bilgi işlem çalışması gerektirir. Görme korteksi retinada ortaya çıkan ve görme siniri aracılığıyla beyne aktarılan nöron ateşlemelerinin ikiboyutlu örüntülerinin oluşturduğu gürültülü zaman dizilerinden, geriye doğru çalışarak dışsal uzamın yorumlanmış bir üçboyutlu temsilini yeniden inşa etmelidir. Bir metre karelik kıymetli korteks gayri mülkümüzün önemli bir kısmı görsel bilgiyi işlemeye ayrılmıştır ve siz bu kitabı okurken milyarlarca nöron bu görevi yerine getirmek için sürekli çalışmaktadır (bir atölyede dikiş makinesine gömülmüş ve dev bir yorganı saniyede defalarca dikip yeniden diken kadın terziler gibi). Benzer şekilde, bizim görünüşte basit değerlerimiz ve isteklerimiz

aslında muazzam bir karmaşıklık içerir.[4] Programlayıcımız bu karmaşıklığı bir fayda işlevine nasıl nakledebilir?

Yaklaşımlardan biri, sahip olduğumuz ve yapay zekânın peşine düşmesini istediğimiz her değerin tam bir temsilini doğrudan kodlamaya çalışmak olacaktır. Eğer olağanüstü derecede basit hedeflerimiz varsa, örneğin pi sayısının basamaklarını hesaplamak istiyorsak, yani istediğimiz *yegâne* şey yapay zekânın pi'nin basamaklarını hesaplamasıysa ve bu hedefin peşine düşülmesinden doğabilecek başka hiçbir sonuçla ilgilenmiyorsak (altyapı bolluğunun başarısızlığına ilişkin değerlendirmemizi hatırlayın) bu yaklaşım işe yarayabilir. Bu aleni kodlama yaklaşımı evcimenlik güdülenme seçilimi yöntemlerinin kullanımında da işe yarayabilir. Ama eğer makul bir *insani* değeri teşvik etmek ya da korumak istiyorsak ve süper zeki bir egemen olmaya niyet etmiş bir sistem kuruyorsak, o zaman gerekli tam hedef temsilini açıktan kodlamak beyhude bir çaba gibi gözükmektedir.[5]

Eğer insani değerleri bilgisayar koduyla tam teşekküllü temsiller yaratarak bir yapay zekâya nakletmeyi başaramazsak, önümüzde başka hangi seçenekler olabilir? Bu bölümde alternatif birkaç yol ele alınıyor. Bunlardan bazıları ilk bakışta akla yatkın görünebilir, ama daha yakından incelendiğinde o kadar da değildir. İlerideki araştırmalar açık kalan bu yollara odaklanmak zorundadır.

Değer yükleme problemini çözmek gelecek kuşağın en iyi matematik yeteneklerinden bazılarına yakışan zorlukta bir araştırmadır. Yapay zekâ niyetlerimizi kolayca anlamak için yeterli aklı geliştirene kadar bu problemle yüzleşmeyi erteleyemeyiz. Yakınsak araçsal nedenler kısmında gördüğümüz gibi, bir türsel sistem nihai değerlerini değiştirme girişimlerine direnecektir. Eğer bir aracı kendi aracılığı üzerine düşünebilme yeteneğini elde edeceği zamana kadar zaten temelde dostane değilse, sonradan beyin yıkama girişimine ya da onu komşusunu daha çok seven farklı bir aracıyla değiştirme girişimine pek sıcak bakmayacaktır.

Evrimsel seçilim

Evrim insani değerleri olan bir organizmayı en az bir kez üretmiştir. Bu olgu, değer yükleme problemini çözmenin yolunun evrimsel yöntemlerden geçtiği inancını teşvik edebilir. Gelgelelim bu yolda güvenliğe ulaşmanın önünde ciddi engeller vardır. Bu engellere 10. Bölüm'ün sonunda arama süreçlerinin ne kadar tehlikeli olabileceğini ele alırken işaret etmiştik.

Evrim, iki adımın karşılıklı dönüşümünü içeren belli tür bir arama algoritması olarak görülebilir; bu adımlardan biri görece basit bir stokastik kurala göre yeni adaylar yaratarak (rastgele mutasyon ya da cinsel rekombinasyon gibi) çözüm adayları havuzunu genişletir, diğeriyse bir değerlendirme işlevi tarafından test edildiğinde kötü puan alan

adayları eleyerek bu havuzu küçültür. Diğer birçok güçlü arama türünde olduğu gibi, bu sürecin resmi olarak saptanmış arama kıstaslarını tatmin eden ama bizim zımni beklentilerimizi karşılamayan bir çözüm bulması riski vardır. (İster tipik bir insanla aynı hedeflere ve değerlere sahip olan bir dijital zihnin evrilmesi için ister sözgelimi tamamen ahlaklı ya da tamamen itaatkâr olan bir zihnin evrilmesi için uğraşılıyor olsun, bu geçerli olacaktır.) Eğer arzuladığımızı düşündüğümüz şeyin sadece bir veçhesinden ziyade hedeflerimizin tüm boyutlarını doğru şekilde temsil eden resmi bir arama kıstası saptayabilirsek bu riskten kaçınılabilir. Fakat bu tam da değer yükleme problemidir ve bu bağlamda problemin çözüldüğünü varsaymayı gerektirecektir.

Başka bir problem daha var:

> Doğal dünyada yıl başına düşen toplam ıstırap miktarı akla hayale sığmayacak kadar fazladır. Benim bu cümleyi yazdığım dakika zarfında, binlerce hayvan canlı canlı yeniyor, diğerleri hayatta kalmak için çırpınıyor, bir diğer kesimi asalaklar tarafından içeriden yavaş yavaş yok ediliyor, binlercesiyse açlık, susuzluk ya da hastalıktan ötürü yok olup gidiyor.[6]

Sadece kendi türümüz içinde bile her gün 150 bin insan telef olurken sayısız insan da korkutucu çeşitlilikte işkenceler ve sıkıntılar yaşıyor.[7] Doğa büyük bir deneyci olabilir, ama bir etik değerlendirme kurulundan asla geçer not alamaz: Helsinki Beyannamesi'ne ve sağ, sol, ortayolcu her türlü ahlaki norma aykırıdır. Bu tür dehşetleri bilgisayar ortamında nedensiz yere çoğaltmamamız önemlidir. İnsan türü zekâ üretmek için evrimsel yöntemlerden yararlanıldığında zihin suçundan kaçınmak özellikle zor görünmektedir, en azından süreç gerçek biyolojik evrime benzer bir şey olacaksa.[8]

Takviyeli öğrenme

Takviyeli öğrenme, aracıların bazı kümülatif ödül kavramlarını azamiye çıkarmayı öğrenmelerini sağlayan tekniklerin incelendiği makine öğrenimi alanıdır. Arzulanan performansın ödüllendirildiği bir çevre inşa ederek, bir takviyeli öğrenme aracısının çok çeşitli problemleri çözmeyi öğrenmesi sağlanabilir (programlayıcılardan ödül sinyali haricinde ayrıntılı bir bilgilendirme ya da geri besleme olmadığında bile). Çoğu zaman, öğrenme algoritması bir tür değerlendirme işlevinin tedricen kurulmasını içerir ve bu işlev durumlara, durum-eylem çiftlerine ya da kurallara değerler atfeder. (Sözgelimi bir program tavla oynamayı mümkün hamlelere dair değerlendirmesini artımlı olarak geliştirmek için takviyeli öğrenmeden yararlanarak öğrenebilir.) Deneyimler ışığında sürekli güncellenen değerlendirme işleviyse, değer hakkında bir öğrenme biçimini kapsıyor olarak görülebilir. Gelgelelim öğrenilen şey yeni *nihai* değerler değil, belli durumlara ulaşmanın (ya da belli durumlarda belli eylemlerde bulunmanın ya da belli politikaları takip etmenin) *araçsal değerlerine dair* giderek daha doğru *kestirimlerdir*.

Bir takviyeli öğrenme aracısının nihai bir hedefi olduğu söylenebildiği ölçüde, bu hedef değişmeden kalır: Gelecekteki ödülleri azamiye çıkarmak. Ayrıca ödül, çevreden alınan özel olarak tasarlanmış algılardan oluşur. Bu nedenle elektronik kafalama [*wireheading*] sendromu ödülü azamiye çıkarmanın bu alternatif yolunu telkin edecek kadar çetrefilli olan bir dünya modeli geliştiren her takviyeli öğrenme aracısındaki muhtemel bir sonuç olarak kalır.[9]

Bu sözler takviyeli öğrenme yöntemlerinin asla güvenli bir tohum yapay zekâda kullanılamayacağı anlamına gelmez, sadece ödülün azamileştirilmesi ilkesi etrafında örgütlenmemiş olan bir güdülenme sistemine tabi kılınması gerektiğini gösterir. Ne var ki bu durum değer yükleme problemine takviyeli öğrenmeden başka bir araç bulunmasını gerektirir.

Çağrışımsal değer eklenmesi

Şimdi insan şunu merak edebilir: Eğer değer yükleme problemi o kadar alengirliyse, bizzat biz nasıl değerler edinmeyi başarıyoruz?

Aşırı basitleştirilmiş modellerden biri şöyle bir şey olabilir. Hayata bazı görece basit başlangıç tercihleriyle (örneğin zehirli uyaranlardan sakınma) ve çeşitli olanaklı deneyimlere yanıt olarak ek tercihler edinme eğilimleriyle (örneğin kültürümüzde değerli görülüp ödüllendirildiğini gördüğümüz nesneleri ve davranışları tercih etmeye meyilli olabiliriz) başlarız. Hem basit başlangıç tercihleri hem de eğilimler içkindir, evrimsel zaman ölçeklerinde doğal ve cinsel seçilim tarafından şekillendirilmiştir. Ama sonunda yetişkinler olarak hangi tercihlere varacağımız yaşamımızdaki olaylara bağlıdır. Dolayısıyla nihai değerlerimizdeki bilgi içeriğinin çoğu genlerimizde önceden yazılı olmaktan ziyade deneyimle elde edilir.

Örneğin birçoğumuz başka bir insanı sever ve dolayısıyla onun esenliğine büyük bir nihai değer atfeder. Bu tür bir değeri temsil etmek için neler gereklidir? Birçok öğe var olsa da sadece ikisini ele alalım: "İnsan" temsili ve "esenlik" temsili. Bu kavramlar doğrudan DNA'mızda kodlanmış değildir. Daha ziyade, DNA bir beyin yaratmak için talimatlar içerir ve bunlar tipik bir insani çevreye yerleştirildiğinde birkaç yıl içinde insan ve esenlik kavramlarını kapsayan bir dünya modeli geliştirecektir. Bu kavramlar bir kez oluşturulduğunda belli anlamlı değerleri temsil etmek için kullanılabilir. Ama diğer edinilmiş kavramlardan (saksı ya da tirbuşon gibi) ziyade *bu* kavramların etrafında oluşturulan değerlere varacak bazı mekanizmaların içkin olarak mevcut olması gerekir.

Bu mekanizmanın nasıl işlediğinin ayrıntıları iyi anlaşılmamıştır. İnsanlarda mekanizma muhtemelen karmaşık ve çeşit çeşittir. Meseleyi yumurtadan çıktıktan kısa süre sonra yuvadan ayrılan kuşlarda yavru mühürleme gibi (yeni doğan kuşun yumurtadan çıktıktan sonraki ilk gün içinde uygun bir hareketli uyaran sunan bir

nesneye fiziksel yakınlık arzusu edindiği durum) daha basit bir biçimde ele alırsak anlaması daha kolay olur. Yavru kuşun hangi tikel nesneye yakın olmayı arzuladığı kendi deneyimlerine bağlıdır; sadece bu şekilde mühürleme genel eğilimi genetik olarak belirlenmektedir. Benzer şekilde, Harry, Sally'nin esenliğine bir nihai değer atfediyor olabilir; ama bu ikili hiç karşılaşmamış olsaydı, Harry başka birine âşık olabilirdi ve dolayısıyla nihai değerleri de farklı olurdu. Genlerimizin hedef edinen bir mekanizma inşası için kodlama yapabilme yeteneği bilgi açısından fazla (bizatihi gende olabileceklerden daha fazla) karmaşıklığı olan nihai hedeflere nasıl varabildiğimizi açıklar.

Dolayısıyla bir yapay zekâ için güdülenme sistemini de benzer bir ilkeye göre inşa edip edemeyeceğimizi düşünebiliriz. Demem o ki, karmaşık değerleri doğrudan saptamak yerine, yapay zekânın uygun bir çevreyle etkileşimde bulunduğunda bu değerlerin edinilmesine yol açan bir mekanizma saptayabilir miyiz?

İnsanlarda gerçekleşen değer eklenme sürecini taklit etmek zor görünmektedir. İnsanlardaki ilintili genetik mekanizma evrim yoluyla sonsuz sayıda işin ürünüdür ki, bu işlerin tekrarlanması güç olabilir. Ayrıca mekanizma muhtemelen insanın sinirbilişsel mimarisine göre incelikle ayarlanmıştır ve dolayısıyla tüm beyin emülasyonları haricinde makine zekâlara uygulanamaz. Ayrıca eğer yeterince aslına sadık tüm beyin emülasyonları mevcutsa, öncesinde yüklü bazı insani değerlerin tam temsilleriyle gelen bir yetişkin beyniyle başlamak daha kolay görünecektir.[10]

Dolayısıyla insan biyolojisini yakından taklit eden bir değer artırımı sürecini uygulamaya çalışmak değer yükleme problemine yönelik umut vermeyen bir saldırı gibi görünmektedir. Ama belki de bir yapay zekâyı kendi hedef sistemine ilintili karmaşık değerlerin yüksek doğrulukta temsillerini taşımaya götürecek apaçık bir yapay ikame mekanizma tasarlayabiliriz? Bunun başarılı olması için yapay zekâya bir biyolojik insan ile aynı değerlendirme eğilimlerini vermek zorunlu olmayabilir. Hatta bu istenir bir amaç bile olmayabilir; nihayetinde insan doğası kusurludur ve çoğu zaman belirleyici bir stratejik üstünlük elde etmeye hazır olan bir sistemde hoşgörülebilecek olan kötülüğe yatkınlığı sık sık sergiler. Belki de insan normlarından sistematik şekillerde (örneğin bir insanda mevcut olsa istisnai derecede iyi bir karakteri yansıttığını söyleyeceğimiz şekillerde diğerkâm, şefkatli ya da âlicenap olan nihai hedefler edinme yönünde daha sağlam bir eğilime sahip olarak) ayrılan bir güdülenme sistemini hedef almak daha iyidir. Gelgelelim insan normundan bu tür sapmaların iyileştirme olarak sayılması için rastgele değil, çok tikel doğrultularda belirtilmesi gerekir ve insani açıdan anlamlı değerlendirici genellemeler sağlaması için büyük oranda bozulmamış bir insan merkezci atıf çerçevesinin varlığını varsaymaya devam edeceklerdir ki, 8. Bölüm'de incelediğimiz yapay derecede makul hedef tariflerinin bir tür tersten örneklemelerinden sakınılabilsin. Bunun gerçekleşebilir olup olmadığı tartışmaya açıktır.

Çağrışımsal değer eklenmesiyle ilgili meselelerden bir diğeri de yapay zekâ sisteminin eklenme mekanizmasını çalışmaz hale getirebilecek olmasıdır. 7. Bölüm'de gördüğümüz üzere, hedef sistemi bütünlüğü yakınsak bir araçsal değerdir. Yapay zekâ araştırmacıları bilişsel gelişimde belli bir aşamaya ulaştıklarında, eklenme mekanizmasının işleyişini yozlaştırıcı bir etki olarak görmeye başlayabilirler.[11] Bu mutlaka kötü bir şey değildir, ama hedef sisteminin doğru zamanda mühürlenmesi için gerekli özenin gösterilmesi şarttır: Uygun değerler toplandıktan *sonra*, ama istenmeyen ilave eklenmeler üzerine yazılmadan *önce*.

Güdülenme iskeleti

Değer yükleme problemine bir diğer yaklaşımaysa güdülenme iskeleti adını verebiliriz. Tohum yapay zekâya açıktan kodlamaya da başka bir uygulanabilir yöntem aracılığıyla temsil edebileceğimiz görece basit nihai hedefleri olan bir geçici hedef sistemi vermeyi içerir. Yapay zekâ daha karmaşık temsil yetileri geliştirdiğinde, bu geçici iskelet hedef sistemi yerine farklı nihai hedefleri olan bir sistemi geçiririz. Bu halef hedef sistemi de tam teşekküllü bir süper zekâ haline gelmekte olan yapay zekâyı yönetir.

İskelet hedefleri yapay zekâ için sadece araçsal olmayıp aynı zamanda nihai hedefler de olduğu için, yapay zekânın bunların ikame edilmesine direnmesi beklenebilir (hedef-içerik bütünlüğü yakınsak bir araçsal değerdir). Bu durum tehlikelidir. Eğer yapay zekâ iskelet hedeflerinin ikame edilmesini püskürtmeyi başarırsa, yöntem çuvallar.

Bu başarısızlıktan kaçınmak adına önlemler alınması şarttır. Örneğin yeterlilik kontrol yöntemleri olgun güdülenme sistemi yerleştirilene kadar yapay zekânın güçlerini sınırlamak için uygulanabilir. Daha özelde, yapay zekânın bilişsel gelişimini güvenli olan ama nihai hedefleri arasında yer almasını istediğimiz değerleri temsil etmesine izin veren bir düzeyde tutmaya çalışabiliriz. Bunu yapmak için belli türde entelektüel yetileri (örneğin stratejiler oluşturma ve Makyavelci planlar yapma) güdük bırakmaya çalışıp, görünüşte daha zararsız yeteneklerin biraz daha üst bir düzeye kadar gelişmesine izin verebiliriz.

Ayrıca güdülenme seçilimi yöntemlerini tohum yapay zekâ ile programlama ekibi arasında daha işbirlikçi bir ilişkiyi teşvik etmek için kullanmaya çalışabiliriz. Örneğin iskelet güdülenme sistemine programlayıcılardan internet üzerinden rehberliği (yapay zekânın mevcut hedeflerinden herhangi birini ikame etmelerine izin vermek dahil) memnuniyetle karşılama hedefini dahil edebiliriz.[12] Diğer iskelet hedefleri arasında programlayıcılara onun değerleri ve stratejileri hakkında şeffaf olmak ve programlayıcıların anlaması için kolay olan ve sonradan hem evcimenlik güdülenmelerini (bilgi işlem kaynaklarının kullanımının sınırlanması gibi) hem de insani açıdan anlamlı bir nihai hedefin yerleştirilmesini kolaylaştıran bir mimari geliştirmeyi içerebilir.

Tohum yapay zekâya kendisini farklı bir nihai hedefle ikame etmek gibi tek bir nihai hedef bahşedilebileceği bile tahayyül edilebilir; bu nihai hedef programlayıcılar tarafından ancak zımnen ya da dolaylı olarak saptanmış olabilir. Bu türden "kendi kendini ikame eden" bir iskelet hedefin kullanılmasının doğurduğu meseleler değer öğrenme yaklaşımı bağlamında da ortaya çıkar; bu konu bir sonraki altbaşlıkta ele alınacak. Başka birtakım meselelerse 13. Bölüm'de masaya yatırılacak.

Güdülenme iskelet yaklaşımının da mahzurları yok değildir. Bunlardan biri yapay zekânın geçici hedef sisteminin hâlâ geçici hedef sistemi devredeyken çok güçlü hale gelmesi gibi bir risk barındırmasıdır. Bu durumda insan programlayıcıların nihai hedef sistemini yerleştirme çabalarını baltalayabilir (zorla direnerek ya da sessiz yıkımla). O zaman, tohum yapay zekâ tam teşekküllü bir süper zekâ haline gelirken eski nihai hedefler varlığını sürdürebilir. Bir diğer mahzuruysa nihai olarak amaçlanmış hedefleri insan düzeyinde bir yapay zekâya yerleştirmenin bunu daha ilkel bir yapay zekâya yerleştirmekten her zaman daha kolay olmamasıdır. İnsan düzeyinde bir yapay zekâ daha karmaşıktır ve saydamsız, değiştirmesi zor bir mimari geliştirmiş olabilir. Bir tohum yapay zekâ ise programlayıcıların yararlı gördükleri her yapıyı nakşedebilecekleri bir boş levha gibidir. Tohum yapay zekâya bilahare son nihai hedefler yerleştirme çabalarına yardımcı olacak bir mimari geliştirmesini istemesini sağlayacak iskelet hedefler vermeyi başardığımız takdirde bu mahzur bir üstünlüğe dönüşebilir. Gelgelelim bir tohum yapay zekâya bu özellikte hedefler vermenin ne kadar kolay olacağı açık değildir; ayrıca ideal olarak güdülenmiş bir tohum yapay zekânın bile iyi bir mimari geliştirme konusunda insan programlama ekibinden çok daha iyi bir iş çıkarmayı nasıl başaracağı da net değildir.

Değer öğrenme

Şimdi sıra, değer yükleme problemi için önemli ama incelikli bir yaklaşıma geldi. Yapay zekânın zekâsını peşine düşmesini istediğimiz değerleri *öğrenmesi* için kullanmakla alakalıdır. Bunu yapmak için yapay zekâya bazı uygun değerleri en azından zımnen seçeceği bir kıstas sağlamamız gerekir. O zaman yapay zekâyı bu zımnen tanımlanmış değerlerin en iyi tahminlerine göre inşa edebiliriz. Yapay zekâ dünya hakkında daha çok şey öğrenip değer belirleme kıstasının içerimlerini peyderpey ambalajından çıkardıkça tahminlerini sürekli geliştirecektir.

Yapay zekâya geçici bir iskelet hedef veren ve ardından onu farklı bir nihai hedefle ikame eden iskelet yaklaşımının aksine, değer öğrenme yaklaşımı yapay zekânın gelişim ve operasyon aşamaları boyunca değişmeyen bir nihai hedefe sahip olarak kalır. Öğrenmek hedefi değiştirmez; sadece yapay zekânın hedef hakkındaki inançlarını değiştirir.

Dolayısıyla yapay zekâya nihai hedefin ne olduğuna dair bir hipotezin lehine ve aleyhine hangi idraklerin kanıt oluşturduğunu belirlemek için kullanabileceği bir kıstas bahşedilmelidir. Uygun bir kıstas belirlemek zor olabilir. Gelgelelim zorluğun bir kısmı

öncelikle bir yapay genel zekâ yaratma problemiyle alakalıdır, zira sınırlı duyu girdilerinden hareketle çevrenin yapısını keşfedebilecek güçlü bir öğrenme mekanizmasını gerektirir. Bu problemi şimdilik bir kenara koyabiliriz. Ama bir süper zeki yapay zekânın nasıl yaratılacağına çözüm bulunsa bile, özel olarak değer yükleme probleminden doğan zorluklar devam edecektir. Değer öğrenme yaklaşımında bunlar algısal ikili dizgilerini değerler hakkındaki hipotezlere bağlayan bir kıstas tanımlamaya ihtiyaç duyma biçimini alır.

Değer öğrenmenin nasıl uygulanabileceğine dair ayrıntılara dalmadan önce, genel fikri bir örnekle açıklamak yararlı olabilir. Bir kâğıt parçasına birtakım değerlerin tarifini yazdığımızı varsayın. Kâğıdı katlayıp mühürlü bir zarfa koyuyoruz. Ardından insan düzeyinde genel zekâsı olan bir aracı yaratıp ona şu nihai hedefi verebiliriz: "Zarfta tarif edilen değerleri azami oranda gerçekleştir." Bu aracı ne yapacaktır?

Aracı zarftaki kâğıtta ne yazdığını ilk başta bilmez; ama hipotezler oluşturabilir ve bu hipotezlere öncelleri ve mevcut ampirik veriler temelinde olasılıklar atfedebilir. Sözgelimi aracı insan yaratısı metinlerin diğer örnekleriyle karşılaşmış ya da insan davranışına dair bazı genel örüntüleri gözlemlemiş olabilir. Bu sayede bazı tahminler yapabilecektir. Yazılı notun "adaletsizliği ve gereksiz ıstırabı asgariye indir" ya da "hisse sahiplerinin gelirlerini azamiye çıkar" gibi bir değeri tarif etme ihtimalinin; "tüm gölleri plastik poşetlerle doldur" gibi bir değer tarif etmesi ihtimalinden daha yüksek olduğunu tahmin etmek için psikoloji diplomasına sahip olmaya gerek yoktur.

Aracı bir karar aldığında, mektupta tarif edildiğinin en muhtemel olduğuna inandığı değerleri gerçeğe dökmekte etkili olacak eylemlerde bulunabilir. Daha önemlisi, aracı mektupta söylenenler hakkında daha fazlasını öğrenmenin yüksek bir araçsal değeri olduğunu görecektir. Bunun nedeni mektupta tarif edilmesi muhtemel olan hemen her nihai değer için şunu söyleyebiliriz: Eğer aracı ne olduğunu bulursa o değerin gerçekleştirilme ihtimali daha yüksektir, zira aracı o zaman bu değeri daha etkili şekilde izleyecektir. Ayrıca aracı 7. Bölüm'de tarif edilen yakınsak araçsal değerleri de keşfedecektir: Hedef sistemi bütünlüğü, bilişsel ilerleme, kaynak edinimi vesaire. Fakat aracının insan refahına ilişkin olarak mektupta tarif edilen değerlere yeterince yüksek bir olasılık atfettiğini varsaydığımızda, bu araçsal değerleri gezegeni derhal bir komputronyuma[13*] dönüştürüp bu suretle insan türünü yok ederek izlemeyecektir, çünkü bunu yapmak nihai değerini gerçeğe dökme ihtimalini tamamen yok etme riskini doğuracaktır.

Bu tür bir aracıyı, her biri farklı yönlere çeken birkaç römorköre bağlı bir mavnaya benzetebiliriz. Her römorkör aracının nihai hedefine dair bir hipoteze denk düşer. Her bir römorkörün motor gücü bağlantılı hipotezin olasılığına denk düşer ve dolayısıyla yeni kanıtlar geldikçe değişir ve mavnanın hareket yönünde değişikliklere yol açar. Ortaya çıkan güç mavnayı bir yandan zımni nihai hedef hakkında bilgi sahibi olmayı kolaylaştıran bir yolda ilerletirken, diğer yandan geri döndürülemez zarara yol açacak şekilde sığlıklara vurmasını da engeller ve sonrasında nihai değere dair kesin bilginin

açık denizine ulaşıldığında, ciddi güç uygulamaya hâlâ devam eden römorkörlerden biri mavnayı en düz ya da en uygun yolda keşfedilen değerin gerçekleştirilmesi yoluna çekecektir.

Zarf ve mavna metaforları değer öğrenme yaklaşımının altında yatan ilkeye örnek teşkil eder, ama birtakım kritik teknik meselelerin üzerinden atlar. Bunlar söz konusu yaklaşımı formel bir çerçeve içinde geliştirmeye başladığımızda daha da göz önüne gelir (bkz. KUTU 10).

Göze çarpan meselelerden biri, yapay zekânın "zarftaki kâğıtta tarif edilen değerleri azami düzeyde gerçekleştir" gibi bir hedefle nasıl donatılacağıdır. (KUTU 10'un terimleriyle söylersek, değer kıstası n nasıl tanımlanacaktır?) Bunu yapmak için değerlerin tarif edildiği yeri saptamak zorunludur. Bizim örneğimizde bunu yapmak için zarftaki mektuba başarılı bir gönderme yapmak gerekir. Bu durum entipüften görünebilirse

KUTU 10: Değer öğrenmenin formelleştirilmesi

Formel bir tarif yapmak bazı şeyleri daha net görmemize yardımcı olabilir. Ne var ki formalizmden hoşlanmayan okurlar bu kısmı atlayabilir.

Bir aracının sınırlı sayıda ayrı çemberde çevresiyle etkileşimde olduğu basitleştirilmiş bir çerçeve düşünün.* Çember k'de aracı y_k eylemini icra eder ve ardından x_k idrakini alır. m yaşam süresi olan bir aracının etkileşim tarihi $y_1x_1y_2x_2 \ldots y_mx_m$'dir (biz bunu $yx_{1:m}$ ya da $yx_{\leq m}$ şeklinde kısaltabiliriz). Her bir çemberde aracı o zamana kadar aldığı idrak silsilesine dayalı olarak bir eylem seçer.

Öncelikle bir takviyeli öğrenen düşünün. Optimum bir takviyeli öğrenen (YZ-PÖ) gelecekteki beklenen ödülleri azamiye çıkarandır. Şu denkleme riayet eder:†

$$y_k = \arg\max_{y_k} \sum_{x_k yx_{k+1:m}} (r_k + \ldots + r_m)\, P(yx_{\leq m} \mid yx_{<k} y_k).$$

Ödül silsilesi $r_k, \ldots, r_m$ idrak silsilesi $x_{k:m}$'de zımnen vardır, zira aracının verili bir çemberde aldığı ödül aracının bu çemberde aldığı idrakin parçasıdır.

Daha önce savunulduğu üzere, bu takviyeli öğrenme türü mevcut bağlamda uygun değildir, zira yeterince zeki bir aracı eğer ödül sinyalini (elektronik kafalama) doğrudan manipüle edebilirse azami ödül elde edebileceğini fark edecektir. Zayıf aracılar için bunun bir problem olması gerekmiyor, zira kendi ödül kanallarına bulaşmalarını fiziksel olarak engelleyebiliriz.

---

* Burada Daniel Dewey'nin (2011) bazı açıklamalarından yararlanacağız. Bu çerçeveye katkı sunan diğer arka plan fikirleri Marcus Hutter (2005) ve Shane Legg (2008), Eliezer Yudkowsky (2001), Nick Hay (2005), Moshe Looks ve Peter de Blanc tarafından geliştirilmiştir.

† Gereksiz karmaşıklıklardan kaçınmak adına, dikkatimizi gelecekteki ödülleri umursamazlık etmeyen belirlenimci aracılarla sınırlıyoruz.

Ayrıca sadece bizim için makul olan şekillerde hareket ettiklerinde ödül alacakları şekilde çevrelerini kontrol edebiliriz. Fakat bir takviyeli öğrenen, ödüllerinin bizim isteklerimize ve kaprislerimize yapay bağımlılığını ortadan kaldırmak için güçlü bir itki duyacaktır. Bir takviyeli öğrenen ile ilişkimiz bu nedenle temelde çatışmaya dayalıdır. Eğer aracı güçlüyse, tehlike vardır.

Elektronik kafalama sendromunun çeşitlilikleri dışsal bir duyum ödül sinyali aramayan, bilakis hedefleri bir iç durumun elde edilmesi olarak tanımlanan sistemleri de etkileyebilir. Örneğin "eyleyen-eleştiren" adı verilen sistemlerde aracının davranışının verili bir performans ölçüsünü yakalamaktan ne kadar uzak olduğunu hesaplayan ayrı bir eleştiren modülünün onaylanmamasını asgariye çekmek için eylemler seçen bir eyleyen modülü vardır. Bu kurulumdaki sorun eyleyen modülünün eleştireni değiştirerek ya da tamamen ortadan kaldırarak onaylanmamayı asgariye çekebileceğini fark edebilecek olmasıdır, tıpkı meclisi lağvedip basını devletleştiren bir diktatör gibi. Sınırlı sistemler söz konusu olduğundan, eyleyen modülüne eleştiren modülünü değiştirmek için hiçbir araç vermeyerek bu problemden kaçınılabilir. Yeterince zeki ve verimli bir eyleyen modülü ise her zaman eleştiren modülüne erişim elde edebilir (sonuçta eleştiren modülü bir bilgisayardaki fiziki bir süreçten ibarettir).*

Değer öğrenene gelmeden önce, gözlem-fayda azamileştiricisi (YZ-GFA) adı verilen bir ara adımı ele alalım. YZ-TÖ'deki ödül dizilerini ($r_k + \ldots + r_m$) yapay zekânın gelecekteki etkileşim tarihinin tamamına dayalı olmasına izin verilen bir fayda işleviyle ikame ederek elde edilir:

$$y_k = arg\,max_{y_k} \sum_{x_k y x_{k+1:m}} U(yx_{\leq m}) P(yx_{\leq m} \mid yx_{<k} y_k).$$

Bu formül elektronik kafalama probleminin etrafından dolaşmayı sağlar, zira tüm bir etkileşim tarihi boyunca tanımlanmış bir fayda işlevi kendi kendini kandırma (ya da aracı cephesinden gerçekliğin doğru bir görüntüsünü elde edecek yeterlilikte yatırım yapılamaması) işaretlerini gösteren etkileşim tarihlerini cezalandırmak üzere tasarlanabilir.

Dolayısıyla YZ-GFA, *ilkesel olarak*, elektronik kafalama probleminin etrafından dolaşmayı mümkün kılar. Gelgelelim bu olanaktan yararlanmak için mümkün olan etkileşim tarihleri

---

* Matematiksel açıdan, bir aracının davranışı mümkün olan her bir etkileşim tarihini bir eyleme haritalandıran bir *aracı işlevi* olarak formalize edilebilir. En basit olan aracılar haricinde, aracı işlevini açıkça bir arabul çizelgesi olarak göstermenin oluru yoktur. Bunun yerine aracıya hangi eylemi yerine getireceğine dair bir hesap yapma yöntemi verilir. Aynı aracı işlevini hesaplamanın birçok yolu olduğundan, bu bir aracı programı olarak aracının daha da incelikli şekilde bireyselleştirilmesine yol açar. Bir aracı programı verili bir etkileşim tarihi için bir eylemi hesaplayan özgül bir program ya da algoritmadır. Formel olarak somutlaştırılmış bir çevreyle etkileşimde bulunan bir aracı programı düşünmek çoğu zaman matematiksel açıdan uygun ve yararlı olmakla birlikte, bunun bir idealleştirme olduğunu hatırlamak önemlidir. Gerçek aracılar fiziksel olarak somutlaştırılır. Bu demektir ki aracı çevreyle sensörleri ve efektörleri yoluyla etkileşimde bulunmakla kalmaz, aynı zamanda aracının "beyni" ya da kontrolcüsü *bizatihi fiziksel gerçekliğin parçasıdır*. Dolayısıyla onun işlemlerini dış fiziksel müdahaleler etkileyebilir (sadece sensörlerinden algılar alarak değil). Bu nedenle bir noktada bir aracıyı bir aracı uygulaması olarak görmek zorunlu hale gelir. Bir aracı uygulaması çevreden müdahaleler olmadığında bir aracı işlevini uygulayan fiziksel bir yapıdır. (Bu tanım Dewey'nin izinden gitmektedir [2011].)

sınıfının üzerinde uygun bir fayda işlevi saptamamız gerekecektir ki, bu görev korkutucu derecede zor görünmektedir.

Fayda işlevlerini aracının kendi etkileşim tarihlerinden ziyade doğrudan olanaklı dünyalar (ya da olanaklı dünyaların özellikleri veya dünya hakkındaki teoriler) üzerinden saptamak daha doğal olabilir. Eğer bu yaklaşımdan yararlanırsak, YZ-GFZ optimumluk kavramını yeniden formüle ederek basitleştirebiliriz:

$$y = arg\,max_y \sum_w U(w)P(w \mid Ey).$$

Burada $\mathcal{V}(.)$, aracının (kararını verdiği anda) elinde olan toplam kanıtlar, U ise bazı olanaklı dünyalar sınıfına fayda atfeden bir fayda işlevidir. Optimum aracı beklenen faydayı azamiye çeken eylemi tercih eder.

Bu formüllerdeki sıradışı problemlerden biri, fayda işlevi U'yu tanımlamanın güçlüğüdür. Böylece nihayet değer yükleme problemine geri dönüyoruz. Fayda işlevinin öğrenilmesini sağlamak için formalizmimizi fayda işlevlerine karşı belirsizliğe izin verecek şekilde genişletmemiz gerekir. Bunu şöyle yapabiliriz (YZ-DY):*

$$y = arg\,max_{y \in Y} \sum_{w \in W} P(w \mid Ey) \sum_{u \in U} U(w)P(\mathcal{V}(U \mid w).$$

Burada E fayda işlevlerinden fayda işlevleri hakkındaki önermelere bir işlevdir. $\mathcal{V}(U)$ ise fayda işlevi U'nun $\mathcal{V}$ tarafından ifade edilen *değer kıstasını* karşıladığı önermesidir.†

Dolayısıyla hangi eylemin gerçekleştirileceğini kararlaştırmak için şöyle ilerlenebilir: Birincisi, mümkün olan her dünyanın (w) koşullu olasılığını hesaplayın (mevcut kanıtlar ve y eyleminin gerçekleştirileceği varsayımı dikkate alınarak). İkincisi, mümkün olan her fayda işlevi (U) için, U'nun değer kıstası $\mathcal{V}$'yi karşıladığı koşullu olasılığı hesaplayın (w'nin gerçek dünya olduğu koşuluna göre). Üçüncüsü, mümkün olan her fayda işlevi U için, mümkün dünyanın (w) faydasını hesaplayın. Dördüncüsü, bu nicelikleri birleştirerek y eyleminin beklenen faydasını hesaplayın. Beşincisi, mümkün olan her eylem için bu işlemi tekrarlayın ve en yüksek beklenen faydaya sahip olan eylemi gerçekleştirin (beraberlikleri bozmak için keyfi bir yönteme başvurun). Tarif edildiği üzere, bu işlem (mümkün olan her dünyanın alenen ve

---

* Dewey bir değer öğrenme aracısı için şu optimallik fikrini önerir:

$$y_k = arg\,max_{y_k} \sum_{x_k y x_{k+1:m}} P_1(yx_{\leq m} \mid yx_k y_k) \sum_U U(yx_{\leq m}) P_2(U \mid yx_{\leq m}).$$

Burada, P_1 ve P_2 iki olasılık işlevidir. İkinci toplam uygun fayda işlevleri sınıfını, olanaklı etkileşim tarihlerinin üzerine yerleştirir. Metinde sunulan versiyonda basitleştirici olanaklı dünyalar notasyonundan yararlanmanın yanı sıra bazı bağımlılıkları da açık hale getirdik.

† Fayda ($\mathbb{U}$) işlevlerinin faydaların karşılaştırılmasını ve ortalamasının alınmasını mümkün kılacak şekilde olması gerektiğini belirtmek gerekiyor. Genel olarak, bu sorunludur ve iyiye ilişkin farklı ahlak teorilerinin kardinal fayda işlevleri üzerinden nasıl temsil edileceği her zaman açık değildir. Örneğin bkz. MacAskill 2010.

ayrı ayrı ele alınmasını içerir) elbette hesaplama açısından izi sürülebilir değildir. Yapay zekânın bu optimumluk fikrine yaklaşan hesaplamalı kısayollardan yararlanması gerekir.

Dolayısıyla soru bu değer kıstası $\mathcal{V}$'nin nasıl tanımlanacağıdır.* Yapay zekâ değer kıstasının yeterli bir temsiline sahip olduğunda, hangi olanaklı dünyaların gerçek dünya olmasının en muhtemel olduğuna ilişkin bilgi toplamak için genel zekâsından ilkesel olarak yararlanabilir. O zaman, hangi fayda işlevinin w'daki $\mathcal{V}$ kıstasını karşıladığını bulmak amacıyla bu türde akla yatkın olanaklı dünyalardan her biri için bu kıstası uygulayabilir. Böylece YZ-DY formülünü değer öğrenme yaklaşımındaki bu kilit zorluğun (nasıl temsil edileceği zorluğunun) saptanması ve diğerlerinden ayrılmasının bir yolu olarak görebiliriz. Formalizm bu yaklaşımın işler hale getirilebilmesi için çözülmesi gereken diğer birtakım meselelere de ışık tutar (Y, W ve U'nun nasıl tanımlanacağı gibi).†

---

* Veya daha genel olarak, zira $\mathcal{V}$ herhangi bir çift olanaklı dünya ve bir fayda işlevi için $\mathcal{V}(U)$ önermesinin w'da doğru olup olmadığını doğrudan ima edecek şekilde olamayabileceğinden, yapılması gereken yapay zekâya $P(\mathcal{V}(U) \mid w)$ koşullu olasılık dağılımının yeterli bir temsilini vermektir.

† Önce $\mathbb{Y}$'yi, bir aracı için olanaklı olan eylemler sınıfını düşünün. Buradaki meselelerden biri tam olarak neyin eylem sayılması gerektiğidir: Sadece temel motor komutları (örneğin "çıktı kanalı #00101100'a bir elektrik sinyali yollayın") ya da yüksek düzey eylemler (örneğin "kamerayı yüze odaklı tutun")? Bir pratik uygulama planından ziyade bir optimallik fikrini geliştirmeye çalıştığımızdan, alanın temel motor komutları olduğunu varsayabiliriz (ve olanaklı motor komutları zaman içinde değişebileceğinden, $\mathbb{Y}$'yi zamana endekslememiz gerekebilir). Gelgelelim uygulamaya ilerlemek için muhtemelen bir tür hiyerarşik planlama sürecini devreye sokmak gerekecektir. Ama bir insanın bu yönde ne kadar gidebileceğinin sınırları vardır: $\mathbb{Y}$'de bir eylemin beklenen faydasının bilgi işlemi birden çok bilgi işlem gerektirir ve eğer bu tür işlemlerden her biri AI-VL'ye göre değerlendirilmesi gereken $\mathbb{Y}$'deki bir eylem olarak görülürse, daha başlamayı bile imkânsız hale getirecek sonsuz bir gerilemeyle karşılaşırız. Bu sonsuz gerilemeden kaçınmak adına, beklenen faydayı hesaplamayı amaçlayan her türlü aleni girişimi sınırlı sayıda kayda değer eylem olanaklılığıyla sınırlamak gerekir. (Nihayetinde, sistem bu bulgulayıcı süreçle ilgili değişiklikler yapmak için bazı olanaklı eylemlere dair açıktan karar almanın bir yolunu bulabilir; bu eylemler aynı süreç tarafından aleni dikkat için işaretlenmiş olabilir. Dolayısıyla uzun vadede sistem AI-VL'nin saptadığı ideale yaklaşmak için giderek etkili hale gelebilir.)

Ardından bir olanaklı dünyalar sınıfı olan $\mathbb{W}$'yu düşünün. Buradaki zorluklardan biri $\mathbb{W}$'yu yeterince kapsayıcı olacak şekilde saptamaktır. $\mathbb{W}$'ya ilintili bir w'yu dahil edememek yapay zekâyı gerçekten yaşanan bir durumu temsil edemez hale getirecek ve bunun sonunda yapay zekâ kötü kararlar alacaktır. Örneğin $\mathbb{W}$'nun tertibini belirlemek için bir ontoloji teorisinden yararlandığımızı varsayalım. Sözgelimi $\mathbb{W}$'ya parçacık fiziğindeki standart modelde bulunan temel parçacıkların doldurduğu uzay-zaman manifoldundan oluşan tüm olanaklı dünyaları dahil ederiz. Bu durum standart model eksik ya da yanlışsa yapay zekânın epistemolojisini çarpıtabilir. Daha fazla olanağı kapsamak için daha büyük bir $\mathbb{W}$-sınıfından yararlanmaya çalışabiliriz; ama mümkün olan her fiziksel evrenin dahil edildiğinden emin olabilsek bile, başka bir olanağın dışarıda bırakıldığından endişe duyabiliriz. Örneğin bilinç konusundaki olguların fizik konusundaki olgulara bağımlı olmadığı ikinci olanaklı dünyalar olanağına ne diyebiliriz? Peki ya endeksizatik? Biz hata yapabilen insanların göz ardı etmiş olabileceği ama işlerin olabildiğince iyi gitmesi için önemi olduğu anlaşılabilecek diğer türde şeyler? Bazı insanların bir tikel ontoloji teorisinin doğru olduğuna dair güçlü inançları vardır. (Yapay zekânın geleceği hakkında yazan insanlar arasında zihinsel olanın fiziksel olana bağımlı olduğu bir materyalist ontolojiye inanç çoğu zaman ön kabul olarak alınır.) Fakat fikirlerin tarihi üzerine

bir an durup düşünmek gözde ontolojimizin yanlış olma ihtimalinin yüksek olduğunu fark etmemize yol açabilir. 19. yüzyıl istatistikçileri fizikten mülhem bir $\mathbb{W}$ tanımı yapmış olsalardı, muhtemelen Öklitci olmayan uzay zamanı ya da Everettçi ("birçok dünya") kuantum teorisi ya da bir kozmolojik çoklu evren olanağını ya da simülasyon hipotezini (bugün gerçek dünyada edinilme olanağı çok ciddi gözüken olanakları) dahil etmeyi muhtemelen unutacaklardı. Şu anki kuşağın benzer şekilde dikkate almadığı başka olanaklar olduğu da akla yatkındır. (Diğer yandan, eğer $\mathbb{W}$ çok büyükse, sonluötesi kümelere önlemler belirlemekle alakalı teknik zorluklar ortaya çıkar.) İdeal olan, işleri bir şekilde yapay zekânın bir tür ucu açık ontoloji kullanabileceği şekilde düzenleyebilmemiz olacaktır: Bizatihi yapay zekânın yeni türde bir metafizik olanağı tanıyıp tanımayacağına karar verirken kullanabileceğimiz aynı ilkeleri kullanarak sonraya da taşıyabileceği bir ontoloji.

$P(w \mid Ey)$'yi düşünün. Bu koşullu olanağı saptamak değer yükleme probleminin gerçek bir parçası değildir. Yapay zekânın zeki olması için, birçok ilintili olgusal olanağın ötesinde makul derecede doğru olasılıklar çıkarmanın bir yoluna halihazırda sahip olmamız gerekiyor. Bu konuda çok başarısız olan bir sistem bizi burada ilgilendiren türde bir tehlike oluşturmayacaktır. Gelgelelim yapay zekânın kendisini araçsal açıdan etkili kılacak kadar iyi, ama normatif açıdan büyük önem taşıyan bazı olanaklar konusunda doğru şekilde düşünebilecek kadar iyi olmayan bir epistemolojiye sahip olması gibi bir risk vardır. (Şimdi, $P(w \mid Ey)$'yi saptama problemi $\mathbb{W}$'yu saptama problemiyle bu açıdan ilintilidir. $P(w \mid Ey)$'yi saptamak, mantıksal olanaksızlıklara karşı belirsizliğin nasıl temsil edileceği gibi diğer meselelerle yüzleşmeyi de gerektirir.

Daha öncesinde bahsedilen meseleler (olanaklı eylemlerin bir kümesinin, olanaklı dünyaların bir kümesinin ve kanıtı olanaklı dünyalar kümelerine bağlayan bir olabilirlik dağılımının nasıl tanımlanacağı) son derece türseldir. Formel olarak saptanmış çok çeşitli aracılar için benzer meseleler doğar. Geriye, değer öğrenme yaklaşımı için daha özgül olan birtakım meseleleri incelemek kalıyor; yani $\mathbb{U}$, $\mathcal{V}(U)$, and $P(\mathcal{V}(U) \mid w)$'nun nasıl tanımlanacağı.

$\mathbb{U}$ bir fayda işlevleri sınıfıdır. U'daki her bir fayda işlevi $U(w)$ W'daki w'ya, yani olanaklı her dünyaya faydalar atfetmesi gerektiği oranda $\mathbb{U}$ ile $\mathbb{W}$ arasında bir bağlantı vardır. Ama U'nun yeterince çok ve çeşitli fayda işlevini içermesi anlamında geniş tabanlı olması gerekir ki, bunların en azından bir tanesi niyetlenen değerleri temsil etme noktasında iyi bir iş çıkararak güveni haklı çıkarabilsin.

Basit bir şekilde $P(U \mid w)$ yazmak yerine $P(\mathcal{V}(U) \mid w)$ yazmamızın nedeni olasılıkların önermelere atfedildiği gerçeğini vurgulamaktır. Bir fayda işlevi, başlı başına, bir önerme değildir, ama bir fayda işlevini onun hakkında bir iddiada bulunarak bir önermeye dönüştürebiliriz. Örneğin tikel bir fayda işlevi $U(.)$ hakkında belli bir insanın tercihlerini tarif ettiği ya da bir etik teorisinin zımnen söylediği yönergeleri temsil ettiği ya da asil enikonu düşünmüş olsaydı tam da uygulamak isteyeceği fayda işlevi olduğunu iddia edebiliriz. Dolayısıyla "değer kıstası" $\mathcal{V}(.)$ argümanı olarak bir fayda işlevi U'yu alan ve onun değerine U'nun kıstasını tatmin etmesi anlamında bir önerme veren bir işlev olarak yorumlanabilir. $\mathcal{V}(U)$'yu bir kez tanımladıktan sonra, yapay zekâdaki diğer olasılık dağılımlarını elde etmek için kullandığımız hangi kaynaksa ondan koşullu olasılık $P(\mathcal{V}(U) \mid w)$'yu elde etmeyi umabiliriz. (Eğer normatif açıdan ilintili tüm olgular olanaklı dünyaları (W) tekilleştirirken dikkate alınıyorsa, o zaman $P(\mathcal{V}(U) \mid w)$ olanaklı her dünyada sıfıra ya da bire eşit olmalıdır.) Geriye V'nin nasıl tanımlanacağı kalıyor. Bunu da metnin ileriki kısımlarında yapacağız.

de, sakıncaları yok değildir. Tek bir örnek verelim: Göndermenin sadece tikel bir dış fiziksel nesneye değil, tikel bir zamandaki bir nesneye yapılması hayatidir.

Aksi takdirde, yapay zekâ, hedefine ulaşmanın en iyi yolunun özgün değer tarifinin üzerine, daha kolay bir hedef sağlayan bir hedef yazmaktan geçtiğine hükmedebilir (her tamsayı için bir büyük tamsayı olacağı değeri gibi). Bu yapıldığında, yapay zekâ arkasına yaslanıp parmaklarını kütürdetebilir; ama 8. Bölüm'de ele alınan nedenlerden ötürü, kötücül bir başarısızlığın baş göstermesi daha muhtemeldir. Dolayısıyla şu an zamanın nasıl tanımlanacağı sorusuyla karşı karşıyayız. Bir saate işaret edip "Zaman bu aletin hareketlerince tanımlanır" diyebiliriz, ama eğer yapay zekâ saatteki okları oynatarak zamanı değiştirebileceğini tahmin ederse ("zaman" bu şekilde tanımlanıyorsa gerçekten de doğru çıkacak bir tahmin), tanım başarısız olur. (Gerçekçi bir örnekteyse, ilintili değerlerin bir mektupta uygun şekilde tarif edilmeyecek olması nedeniyle daha da karmaşıklaşacaktır; daha muhtemel olanı, insan beyinleri gibi ilintili bilgiyi zımnen içeren önceden mevcut yapılara dair gözlemlerden çıkarılması gerekecektir.)

"Zarfta tarif edilen değerin gerçekleştirilmesini azamiye çekin" hedefinin kodlanmasındaki bir diğer meseleyse, tüm doğru değerler bir mektupta tanımlansa bile ve yapay zekânın güdülenme sistemi bu kaynağa başarıyla girilse bile, yapay zekânın tarifleri bizim istediğimiz gibi yorumlayamayabilecek olmasıdır. Bu durum, 8. Bölüm'de ele alındığı üzere, bir tersten örnekleme riski yaratabilir.

Netleştirmek adına, buradaki zorluk yapay zekânın insanın niyetlerini anlamasının nasıl sağlanacağından çok daha ötedir. Bir süper zekâ bu tür bir anlayışı kolaylıkla geliştirebilir. Burada zorluk, daha ziyade, yapay zekânın tarif edilen değerleri niyetlenen tarzda takip edecek şekilde güdülenmesini sağlamaktır. Yapay zekânın niyetlerimizi anlama yeteneği bunu güvence altına almaz: Yapay zekâ ne kastettiğimizi tamamen bilebilir, ama sözlerimizin bu yorumuna ilgisiz kalabilir (bunun yerine, sözcüklerin başka bir yorumundan güdülenebilir ya da sözcüklerimize tamamen kayıtsız kalabilir).

Buradaki zorluk, güvenlik nedenlerinden ötürü, doğru güdülenmenin insan kavramlarını tamamen temsil edecek ya da insan niyetlerini anlayacak ehliyete kavuşmadan önce tohum yapay zekâya yerleştirilmesinin ideal hedef olması nedeniyle daha da karmaşıklaşmaktadır. Bu durum bir şekilde bir bilişsel çerçeve yaratılmasını ve bu çerçevedeki tikel bir lokasyonun yapay zekânın güdülenme sisteminde nihai değerinin deposu olarak saptanmasının gerekmesidir. Ama bizatihi bilişsel çerçeve gözden geçirilebilir olmalıdır ki, yapay zekâ dünya hakkında bilgisi artıp daha zeki hale geldikçe temsil kapasitelerini genişletebilsin. Yapay zekâ bir tür bilim devrimi yaşayabilir ve bu sırada dünya görüşü sarsılıp belki de öncesinde değerler hakkındaki düşünme tarzının kafa karışıklıklarına ve yanılsamalara dayalı olduğunu keşfettiği ontolojik krizlere girebilir. Ama insanaltı bir gelişim düzeyinde başlayıp sonrasında galaktik bir süper zekâ haline geleceği gelişimi boyunca devam eden yapay zekânın tutumuna özünde değişmeyen bir nihai değer rehberlik edecektir; bu nihai değer, yapay zekânın genel entelektüel ilerleyişinin doğrudan sonucu olarak yapay zekâ tarafından daha iyi anla-

şılacaktır ve olgun yapay zekânın özgün programlayıcılarına kıyasla bunu çok farklı anlaması muhtemeldir ama bunu rastgele ya da hasmane bir tarzda değil iyicil derecede uygun bir tarzda farklı diye görmek gerekir. Bunun nasıl başarılacağıysa cevabı henüz bulunmamış bir sorudur.[14] (Bkz. **KUTU 11**)

Özetle, akla yatkın insan değerleri yüklemek için değer öğrenme yaklaşımından nasıl yararlanılacağı henüz bilinmiyor (son dönemdeki fikirlere dair bazı örnekler için, bkz. **KUTU 12**). Şu anda bu yaklaşım bir mevcut teknikten ziyade bir araştırma programı olarak görülmelidir. Eğer işe yararsa, değer yükleme problemine en ideal çözümü oluşturabilir. Diğer yararları arasında, zihin suçunu engellemenin doğal bir yolunu sunması sayılabilir, zira programlayıcılarının hangi değerleri yüklemiş olabileceklerine dair makul tahminlerde bulunan bir tohum yapay zekâ zihin suçunun bu değerler tarafından muhtemelen olumsuz değerlendirileceğini ve en iyi seçeneğin bundan kaçınmak olacağını (en azından daha kesin bilgiler edinilene kadar) öngörecektir.

KUTU 11: Dostane olmak isteyen bir yapay zekâ

Eliezer Yudkowsky yukarıdaki metinde anlatılan türde davranışı mümkün kılması amaçlanan bir tohum zekâ mimarisinin bazı özelliklerini tarif etmeye çalışmıştır. Onun terminolojisiyle, yapay zeka "dışsal gönderme semantiği"nden yararlanacaktır.* Temel fikri örneklendirmek gerekirse, sistemin "dostane" olmasını istediğimizi varsayalım. Sistem D özelliğini örneklendirmeye çalışma hedefiyle yola çıkıyor ama ilk başta D'nin ne olduğu hakkında çok fazla bir şey bilmiyor. D'nin soyut bir özellik olduğunu ve programlayıcılar "dostanelik"ten bahsettiklerinde muhtemelen D hakkında bilgi aktarmaya çalıştıklarını biliyor olabilir. Yapay zekânın nihai hedefi D'yi örneklendirmek olduğundan, önemli araçsal değerlerden biri D'nin ne olduğu hakkında daha fazlasını öğrenmektir. Yapay zekâ D hakkında daha fazla şey keşfettikçe, davranışına D'nin gerçek içeriği daha fazla rehberlik eder. Böylece, umalım, yapay zekâ daha fazla öğrenip daha akıllı hale geldikçe daha dost canlısı olur.

Bu süreçte programlayıcıların yardımı dokunabilir ve yapay zekânın D anlayışı hâlâ tam değilken yapay zekâya "programlayıcı tasdikleri" (ilk başta yüksek olasılık atfedilen D'nin doğası ve içeriğine dair hipotezler) sağlayarak yapay zekânın feci yanlışlar yapma riskini azaltabilirler. Sözgelimi "programlayıcıları yanlış yönlendirmek dostane değildir" hipotezine yüksek bir ilk olasılık verilebilir. Gelgelelim bu programlayıcı tasdikleri "tanımı gereği doğru" değildir: Dostanelik konusunda karşı çıkılmaz aksiyomlar değildir. Daha ziyade dostanelik konusunda ilk hipotezlerdir; akli bir yapay zekânın en azından programlayıcıların epistemik kapasitelerine kendisininkinden daha fazla güvendiği sürece bir yüksek olasılık atfedeceği hipotezlerdir.

---

* Yudkowsky (2001).

Yudkowsky'nin önerisi kendisinin "nedensel geçerlilik semantiği" adını verdiği şeyden yararlanmayı da kapsar. Buradaki fikir yapay zekânın tam olarak programlayıcıların söylediklerini değil daha ziyade ona söylemeye çalıştıkları (gibi bir) şeyi yapması gerektiğidir. Programlayıcılar tohum yapay zekâya dostaneliğin ne olduğunu açıklamaya çalışırken izahatta hatalar yapabilirler. Ayrıca bizzat programlayıcılar dostaneliğin gerçek doğasını tamamen anlamayabilirler. Bu nedenle yapay zekânın programlayıcıların düşüncesindeki hataları düzeltebilecek ve programlayıcıların sunduğu her türlü kusurlu açıklamadan gerçek ya da kastedilen anlamı çıkarabilecek yeteneğe sahip olmasını isteriz. Örneğin yapay zekâ programlayıcıların dostanelik hakkında öğrenip iletişimde bulunmalarını sağlayan nedensel süreçleri temsil edebiliyor olmalıdır. Basit bir örnek verecek olursak, yapay zekâ bir programlayıcının dostanelikle ilgili bir bilgi girerken yazım yanlışı yapabilme ihtimali olduğunu anlamalı ve ardından bu hatayı düzeltmeye çalışmalıdır. Daha genel olarak, yapay zekâ kaynağından programlayıcılara, oradan da yapay zekâya geçerken dostanelik hakkındaki bilgi akışını bozmuş olabilecek olan her türlü tahrif edici etkiyi düzeltmeye çalışmalıdır ("tahrif edici" burada epistemik bir kategoridir). İdeal haliyle, yapay zekâ olgunlaştıkça, her türlü bilişsel önyargıyı ve programlayıcılarının dostaneliğin ne olduğunu tam olarak anlamalarını önleyebilecek olan daha temel diğer yanlış anlayışları aşmalıdır.

KUTU 12: Yakın döneme ait iki (yarı pişmiş) fikir

"Yüce Meryem" adını verebileceğimiz yaklaşım, evrende başka bir yerde zekâ patlamasını başarıyla idare edebilecek medeniyetlerin var olduğu yada bir gün var olacağı ve nihayetinde bizimkilerle ciddi düzeyde örtüşen değerlere sahip olacakları umuduna dayalıdır. O zaman, kendi yapay zekâmızı bu diğer süper zekâların yapmak istediklerini yapmaya güdülenecek şekilde inşa etmeye çalışabiliriz.* Bunun üstünlüğü bizim doğrudan istediğimiz şeyi yapacak şekilde güdülenmiş yapay zekâmızı yaratmaktansa, bunu yapmanın daha kolay olabilecek olmasıdır.

Bu planın tutması için yapay zekâmızın uzaylı bir süper zekâyla iletişim kurabiliyor olması şart *değildir*. Daha ziyade, yapay zekâmızın eylemlerine uzaylı süper zekâların ondan ne yapmasını isteyeceklerine dair kendi *tahminleri* kılavuzluk edecektir. Yapay zekâmız başka yerlerdeki zekâ patlamalarının muhtemel sonuçlarını kendine örnek alacak ve kendisi de süper zeki hale geldikçe tahminleri giderek daha doğru hale gelecektir. Kusursuz bilgi şart değildir. Zekâ patlamalarının çok çeşitli akla yatkın sonuçları olabilir ve o zaman, yapay zekâmız ortaya çıkabilecek olan çok farklı türdeki süper zekâ tercihlerini olasılığa göre tartarak yerli yerine oturtmak için elinden geleni yapacaktır.

---

* Bu tabir Amerikan futbolundan alınmıştır. Bu oyunda "Yüce Meryem" bir oyuncunun başka bir çaresi kalmadığında, takım arkadaşının yakalayıp touchdown yapacağı umuduyla ve genellikle süre dolarken ileri attığı çok uzun pastır.

Yüce Meryem yaklaşımının bu versiyonu yapay zekâmız için diğer süper zekâların tercihlerine göndermede bulunan bir nihai değer yaratmamızı gerektirir. Bunun tam olarak nasıl yapılacağı henüz net değildir. Gelgelelim süper zeki aracılar yapısal açıdan o kadar ayırıcı olabilirler ki, gelişen yapay zekâmızdaki dünya modeline bakacak ve bir süper zekânın varlığına tekabül eden temsili öğeleri saptayacak bir detektör işlevi gören bir kod parçası yazabiliriz. Bu durumda detektör, bir şekilde, söz konusu süper zekânın tercihlerini çıkarsayacaktır (kendi yapay zekâmızda temsil edildiği şekliyle).* Eğer bu tür bir detektör yaratabilirsek, bunu yapay zekâmızın nihai değerlerini tanımlamak için kullanabiliriz. Zorluklardan biri yapay zekâmızın hangi temsili çerçeveyi geliştireceğini bilmeden önce detektörü yaratmak zorunda kalabilecek olmamızdır. Böylece detektör bilinmeyen bir temsili çerçeveyi sorgulama ve orada temsil edilen süper zekânın tercihlerini çıkarma gereği duyabilir. Bu zor görünse de belki akıllı bir çözüm bulunabilir.†

Eğer temel kurulumun çalışması sağlanabilirse, hemen çeşitli düzeltmeler akla gelir. Örneğin *her* uzaylı süper zekânın tercihlerini takip etmek yerine, bizim yapay zekâmızın nihai değeri uzaylı süper zekâların altkümesinden (değerleri bizim değerlerimize daha yakın olanları seç-

---

* Yüce Meryem yaklaşımı bir süper zekânın kendi tercihlerini biz insanların kendi tercihlerimizi ifade edebileceğimizden daha kesin şekilde ifade edebileceği fikrine dayanır. Örneğin bir süper zekâ tercihini kod olarak saptayabilir. Dolayısıyla eğer yapay zekâmız diğer süper zekâları çevreyi algılayan hesaplamalı süreçler olarak temsil ediyorsa, o zaman yapay zekâmız bu yabancı süper zekâların bazı farazi uyarılmalara (görme alanlarında açılan ve onlara kendi yapay zekâsının kaynak kodunu sunup talimatlarını bize önceden belirlenmiş uygun bir formatta iletmelerini isteyen bir "pencere" gibi) nasıl yanıt vereceklerine dair akıl yürütebiliyor olmalıdır. O zaman, yapay zekâmız bu hayali talimatları (bu yabancı süper zekâların temsil edildiği bu karşıolgusal senaryoya dair kendi modelinden) sesli okuyabilirdi ve bizim de yapay zekâmızı bu talimatlara uymaya güdülenmiş şekilde inşa etmiş olmamız gerekirdi.

† Bunun alternatifi, bir süper zeki medeniyet tarafından yaratılmış fiziksel yapıların (temsillerini) arayan (bizim yapay zekâmızın dünya modeli dahilinde) bir detektör yaratmak olacaktır. O zaman, farazileştirilmiş süper zekâların tercih işlevlerini saptama adımını atlayıp, kendi yapay zekâmıza nihai değer olarak süper zeki medeniyetlerin üretmeye meyilli olduklarına inandığı tüm fiziksel yapıları kopyalama çalışmayı verebiliriz. Fakat bu versiyonda da teknik zorluklar vardır. Sözgelimi bizim kendi yapay zekâmız, süper zekâya eriştikten sonra bile diğer süper zekâların hangi fiziksel yapıları inşa edeceğini kesinkes bilemeyebileceğinden, yapay zekâmız bu yapılara benzerlerini bulmaya çalışma yoluna gitme ihtiyacı duyabilir. Bunu yapmak için yapay zekâmızın fiziksel bir insan ürünün bir diğerine ne kadar yaklaştığını yargılamamızı sağlayacak bir benzerlik ölçüsüne ihtiyacımız olacaktır. Ama kaba fiziksel ölçülere dayalı benzerlik ölçüleri yetersiz olabilir; örneğin bir beynin bir Camembert peynirine emülasyon çalıştıran bir bilgisayara olduğundan daha fazla benzediğini söylemesi bir işe yaramaz.

Daha uygulanabilir bir yaklaşım ise "yol göstericiler" aramak olacaktır Uygun basitlikte bir formatta kodlanmış olan fayda işlevleri hakkındaki mesajlar. Yapay zekâmızı evrende var olabileceği varsayımında bulunduğu fayda işlevleri hakkında bu türde ne mesaj varsa onların peşinden gitmesini isteyecek şekilde inşa ederdik ve Dünya dışı dost yapay zekâların bizimki gibi basit medeniyetlerin kendi yapay zekâmızın araması için inşa edilmiş olduğunu (kendi süper zekâlarıyla) kavrayacak türde çeşitli yol göstericiler yaratacağını umut ederdik.

mek amacıyla) seçmek üzere bir filtre kullanabilir. Sözgelimi bu kümede yer alıp almayacağını belirlemek için bir süper zekânın nedensel kökenine dair kıstaslardan yararlanabiliriz. Ortaya çıkışının belli özellikleri (bunları yapısal terimlerle tanımlayabiliriz), yaratılan süper zekânın değerlerimizi hangi ölçüde paylaşacağıyla bağıntıya sahip olabilir. Belki de nedensel kökenleri tüm beyin emülasyonuna ya da evrimsel algoritmalardan yoğun olarak yararlanmamış veya kontrollü bir havalanmayı akla getirecek şekilde yavaş yavaş ortaya çıkmış bir tohum yapay zekâya kadar götürülebilecek olan süper zekâlara daha fazla güven duyabiliriz. (Nedensel kökenleri hesaba katmak, kendilerinin çoklu kopyalarını çıkaracak olan süper zekâlara aşırı önem vermemizi de engelleyecektir, hatta bizi böyle yapmaları için bir teşvik yaratmaktan alıkoyacaktır.) Başka birçok iyileştirme de mümkün olabilir.

Yüce Meryem yaklaşımı orada bir yerlerde bizim değerlerimizi yeterince paylaşan diğer süper zekâların olduğuna inanmamızı gerektirir.* Bu durum, yaklaşımı ideal olmaktan çıkarır. Gelgelelim Yüce Meryem yaklaşımının önündeki teknik engeller çok ciddi olmakla beraber, alternatif yaklaşımların önündeki engellere kıyasla daha az sıkıntı yaratabilir. İdeal olmayan ama daha kolay uygulanabilir olan yaklaşımları irdelemek anlam ifade edebilir: Bunları kullanmak amacıyla değil, bilakis ideal bir çözüm zamanında hazır olmadığında yaslanacak bir şey olması adına.

Değer yükleme probleminin nasıl çözüleceğine dair bir diğer fikri de Paul Christiano ortaya atmıştır.† Tıpkı Yüce Meryem yaklaşımı gibi, bu da değer kıstasını meşakkatli bir inşa yoluyla değil, daha ziyade bir "dalavere" ile tanımlamaya çalışan bir değer öğrenme yöntemidir. Ama Yüce Meryem yaklaşımının aksine kendi yapay zekâmız için rol model olarak işaret edebileceğimiz diğer süper zeki aracıların varlığını önvarsaymaz. Christiano'nun önerisi kısa açıklamaya biraz kapalıdır (bir dizi esrarlı düşünce içerir), ama en azından temel öğelerine işaret etmeye çalışabiliriz.

Şunları elde edebildiğimizi varsayın: (a) Tikel bir insan beyninin matematiksel açıdan kesin belirtimi ya da saptanması ve (b) keyfi derecede çok miktarda bellek ve CPU gücü olan ideal-leştirilmiş bir bilgisayar içeren matematiksel açıdan iyi tanımlanmış bir sanal çevre. Bu ikisi de var olduğunda bir fayda işlevini (F) insan beyninin bu çevreyle etkileşimde bulunduktan sonra üreteceği çıktı olarak tanımlayabiliriz. F matematiksel açıdan iyi tanımlanmış bir hedef olacaktır, ama hesaplama sınırlılıklarından ötürü *açıktan* tanımlanamıyor olabilir. Ne var ki F, F'nin ima ettikleri hakkındaki hipotezlere olasılıklar atfetmek için çeşitli bulgulayıcılardan yararlanabilen değer öğrenici bir yapay zekâ için bir değer kıstası hizmeti görebilir.

Sezgisel olarak, F'nin uygun şekilde hazırlanmış bir insanın keyfi derecede çok miktarda bilgi işlem gücünden (örneğin bir fayda işlevini saptama analizinde kendisine yardımcı olacak

---

* Eğer her medeniyet değer yükleme problemini bir Yüce Meryem yaklaşımıyla çözmeye çalışsaydı, bir arıza çıkardı. Birilerinin dikenli yolu seçmesi gerekir.

† Christiano (2012).

astronomik sayıda kendi kopyasını çalıştıracak ya da ona bu analizi yürütmesi için daha iyi bir süreç geliştirmesinde yardımcı olacak kadar bilgi işlem gücünden) yararlanabilme üstünlüğüne sahipse ortaya koyacağı fayda işlevi olacaktır. (Burada 13. Bölüm'de daha ayrıntılı olarak ele alınacak bir izleğe değinmiş oluyoruz: "Bütünlüklü dışkestirilmiş irade".)

İdealleştirilmiş çevreyi saptamak görece kolay görünecektir: Kapasitesi rastgele belirlenmiş bir soyut bilgisayarın matematiksel bir tanımını sunabiliriz ve diğer açılardan, sözgelimi içinde bilgisayar terminali olan tek bir odanın matematiksel bir tarifini sunan (soyut bilgisayarı somutlaştıran) bir sanal gerçeklik programından yararlanabiliriz. Peki, tikel bir insan beyninin matematiksel açıdan kesin bir tarifini nasıl elde edebiliriz? Bunun bariz yolu tüm beyin emülasyonu olacaktır, peki ama ya emülasyon teknolojisi zamanında mevcut değilse?

İşte Christiano'nun önerisinin kilit bir yenilik sunduğu yer burasıdır. Christiano matematiksel açıdan iyi saptanmış bir değer kıstası edinmek için pratik açıdan yararlı bir hesaplamalı zihin modeline (çalıştırabileceğimiz bir modele) ihtiyacımız olmadığını söyler. Tek ihtiyacımız olan (muhtemelen zımni ve umutsuz derecede karmaşık) bir matematiksel *tanımdır* ve bunu elde etmesi çok daha kolay olabilir. İşlevsel beyin görüntüleme ve diğer önlemlerden yararlanarak, seçili bir insanın girdi-çıktı davranışı hakkında belki gigabaytlarca veri toplayabiliriz. Eğer yeterli miktarda veri elde edersek, o zaman tüm bu verileri açıklayan en basit matematiksel model aslında söz konusu tikel insanın bir emülasyonudur. Bizim verilerden hareketle bu en basit modeli *bulmamız* hesaplamalı açıdan zor olsa da, verilere göndermede bulunarak ve matematiksel açıdan iyi tanımlanmış bir yalınlık önleminden (1. Bölüm'de **KUTU 1**'de karşılaştığımız Kolmogorov karmaşıklığının bir çeşidi gibi) yararlanarak modeli *tanımlamamız* pekâlâ mümkün olabilir).*

* İnşa ettiğimiz yapay zekânın modeli bulabilmesi de gerekmiyor. Tıpkı bizim gibi, bu türde karmaşık bir zımni tanımın ne gerektireceğine dair akıl yürütebilir (belki de çevresine bakarak ve bizim takip edeceğimiz muhakeme türünün hemen hemen aynısını takip ederek).

En nihayet, "zarfa ne yazılacağı" sorusu vardır; kelime oyunu yapmadan söylersek, yapay zekânın hangi değerleri öğrenmesini sağlamaya çalışmalıyız? Fakat bu mesele yapay zekâ değer yükleme problemine tüm yaklaşımlarda ortaktır. Bu konuya 13. Bölüm'de geri döneceğiz.

Emülasyon modülasyonu

Değer yükleme problemi tüm beyin emülasyonu açısından yapay zekâ için olduğundan biraz farklı görünmektedir. Algoritmaların ve mimarinin incelikli bir anlayışını ve kontrolünü varsayan yöntemler emülasyonlara uygulanabilir değildir. Diğer yandan, artırım güdülenmesi seçilimi yöntemi (sil baştan yapay zekâ için uygulanabilir değildir) emülasyonlar ya da geliştirilmiş biyolojik beyinler için kullanıma hazırdır.[15]

Zenginleştirme yöntemi sistemin miras alınan hedeflerinde ince ayarlar yapacak tekniklerle birleştirilebilir. Örneğin psikoaktif tözlerin dijital eşdeğerinden (ya da biyolojik sistemler örneğinde, gerçek kimyasallardan) yararlanarak bir emülasyonun güdülenme durumunu manipüle etmeye çalışabilirsiniz. Şu anda bile değerleri ve güdülenmeleri farmakolojik açıdan sınırlı ölçüde manipüle etmek mümkündür.[16] Geleceğin ilaçlar kitabında daha özgül ve kestirilebilir etkileri olan ilaçlar yer alabilir. Emülasyonların dijital aracı kontrollü deneyleri daha kolay hale getirip tüm beyin parçalarını doğrudan adreslenebilir kılarak bu tür gelişmeleri fevkalade kolaylaştırabilir.

Tam da biyolojik denekler kullanıldığında olduğu gibi, emülasyonlarla ilgili araştırmalar etik açıdan karmaşık durumlara yol açacaktır ve bunların hepsini bir onay formuyla bir kenara atmak mümkün değildir. Bu tür karmaşalar emülasyon yolundaki ilerlemeyi (düzenleme ya da ahlaki kısıtlamalardan ötürü) yavaşlatabilir, belki de emülasyonların güdülenme yapısının nasıl manipüle edileceğine dair çalışmaları bilhassa engelleyerek. Sonuç olarak, emülasyonlar nihai hedeflerini sınamak ya da uyarlamak için yeterli çalışma yapılmadan önce, bilişsel yetenek bakımından potansiyel açıdan tehlikeli olan süper zekâ düzeylerine varacak şekilde zenginleştirilebilir. Ahlaki karmaşaların bir diğer etkisi de ekiplerin ve ulusların daha az titizlenmesine yol açabilecek olmasıdır. Halbuki dijital insan zihinleriyle deneylerimizdeki ahlak standartlarımızı esnetecek olsak, ciddi miktarda zarar ve yanlıştan sorumlu olabiliriz ki, bunu istemeyeceğimiz açıktır. Diğer şeyler aynı kalmak kaydıyla, bu mülahazalar stratejik açıdan büyük oynanan durumlarda dijital insan araştırma deneklerinden yaygın şekilde yararlanmayı gerektirmeyen alternatif bir yol tutulmasına işaret etmektedir.

Gelgelelim mesele o kadar basit ve net değildir. Tüm beyin emülasyonu araştırmasının yapay zekâ araştırmasına kıyasla ahlaki ihlaller içerme ihtimalinin *daha az* olduğu savunulabilir, zira bir emülasyon zihnin ahlaki statüye eriştiğini tanımamız tamamen yabancı ya da sentetik bir zihninkine kıyasla daha muhtemeldir. Eğer belli türde yapay zekâlar ya da onların alt süreçleri bizim tanımayı başaramadığımız kayda değer bir ahlaki statüye sahipse, bunun sonucunda doğacak ahlaki ihlaller çok ciddi olabilir. Örneğin günümüz programlayıcılarının takviyeli öğrenme aracıları yaratıp onları kaçınmalı uyarıcılara tabi kıldıkları mutlu mesut süreci düşünün. Her gün sadece bilgisayar laboratuvarlarında değil, karmaşık oyuncu olmayan karakterlerin yer aldığı bazı bilgisayar oyunları da dahil olmak üzere birçok uygulamada bu türden sayısız aracı yaratılıyor. Bu aracıların henüz bir ahlaki statüye sahip olamayacak kadar ilkel olduğu söylenebilir. Peki, bundan ne kadar emin olabiliriz? Daha önemlisi, programlarımız ahlaki açıdan ilintili sıkıntıları deneyimleyebilecek hale gelmeden önce zamanında durmayı nasıl bilebileceğimizden ne kadar emin olabiliriz?

(14. Bölüm'de emülasyonun ve yapay zekâ yollarının arzulanırlığını karşılaştırırken ortaya çıkan daha büyük stratejik sorulardan bazılarına geri döneceğiz.)

Kurum tasarımı

Bazı zeki sistemler bizatihi aracılığa muktedir olan zeki parçalardan oluşurlar. Şirketler ve devletler beşeri dünyada buna birer örnektir: Büyük oranda insanlardan oluşsalar da bazı amaçlar için başlı başına özerk aracılar olarak görülebilirler. Bu tür bileşik sistemlerin güdüleri sadece onları oluşturan alt aracıların güdülerine değil, aynı zamanda bu alt aracıların nasıl örgütlendiğine de bağlıdır. Sözgelimi güçlü bir diktatörlükte örgütlenmiş bir grup, diktatör rolünde bulunan alt aracının iradesiyle birebir örtüşen bir istence sahipmiş gibi davranabilirken; bir demokratik grup bazen sanki çeşitli bileşenlerinin istençlerinin bir bileşimi ya da ortalaması olan bir istence sahipmiş gibi davranabilir. Ama bir örgütün, alt aracılarının istençlerinin basit bir işlevi olmayan bir tarzda davranmasını sağlayacak olan yönetişim kurumlarını da düşünebiliriz. (En azından teorik açıdan, *herkesin* nefret ettiği bir totaliter devlet olabilir, çünkü devletin yurttaşlarının bir isyan örgütlemesini engelleyecek mekanizmaları vardır. Her yurttaş devlet mekanizmasındaki rolünü oynamaktansa tek başına isyan ederek daha kötü duruma düşebilir.)

Dolayısıyla bileşik bir sistem için uygun kurumlar tasarlayarak onun etkili güdülenmesini şekillendirmeye çalışabiliriz. 9. Bölüm'de olanaklı bir yeterlilik kontrol yöntemi olarak toplumsal bütünleşmeden bahsetmiştik. Ama orada neredeyse denklerin olduğu bir toplumsal dünyada var oluşunun bir sonucu olarak aracının karşılaştığı teşviklere odaklanmıştık. Buradaysa verili bir aracının *içinde* neler olduğuna, kendi iç örgütlenmesinin onun istencini nasıl belirlediğine odaklanıyoruz. Bu nedenle bir güdülenme seçilimi yöntemine bakıyoruz. Ayrıca bu tür bir iç kurum tasarımı büyük çaplı toplumsal mühendisliğe ya da reforma bağlı olmadığından, genel toplumsal-ekonomik ya da uluslararası ortam ideal ölçüde elverişli değilse bile, süper zekâ geliştiren bireysel bir proje için mevcut olabilecek bir yöntemdir.

Kurum tasarımının en akla yatkın olduğu yer belki de zenginleştirmeyle birleştirilebileceği bağlamlardır. Halihazırda uygun derecede güdülenmiş olan ya da insan türü güdüleri olan aracılarla başlarsak, kurumsal düzenlemeler sistemin rotasından şaşmama şansını artırmak üzere fazladan bir koruma olarak kullanılabilir.

Örneğin sağlam güdülenmiş insan türü aracılarla, sözgelimi emülasyonlarla yola çıktığımızı varsayalım. Bu aracıların bilişsel kapasitelerini artırmak istiyoruz, ama geliştirmelerin güdülerini yozlaştırabileceğinden endişe ediyoruz. Bu zorlukla başa çıkmanın yollarından biri bireysel emülasyonların alt aracı olarak işlev göstereceği bir sistem kurmak olacaktır. Yeni bir geliştirme takdim edildiğinde, öncelikle küçük bir alt aracı alt kümesine uygulanır. Ardından sonuçlar henüz geliştirmenin uygulanmadığı alt aracılardan oluşan bir değerlendirme paneli tarafından incelenir. Ancak bu alt aracılar geliştirmenin yozlaştırmadığına ikna olduğunda alt aracıların geneline de uygulanır.

Eğer geliştirilmiş alt aracıların yozlaştığı görülürse, başka geliştirme eklenmez ve kilit karar alma işlevlerinden uzaklaştırılırlar (en azından sistemin bütünü yozlaşmış alt aracıların güvenle yeniden dahil edilebileceği bir noktaya ilerleyene kadar).[17] Yozlaşmış alt aracılar bu geliştirmeden bir üstünlük elde edebilirler, ama yer aldıkları kurumsal yapı ve tüm alt aracıların küçük bir azınlığını oluşturuyor olmaları iktidarı almalarını ya da yozlaşmayı tüm sisteme yaymalarını engeller. Dolayısıyla sistemin kolektif zekâsı ve yeterliliği bir küçük adım silsilesiyle tedricen geliştirilebilir; burada her adımın sağlamlığı bu adımda takdim edilen yeni alt aracıların ehliyetinden biraz daha az olan alt aracılar tarafından doğrulanır.

Bu tür bir kurum tasarımıyla elde edilebilecek güvenlik miktarı, geliştirilmiş alt aracıları değerlendirmek için kullanılan testlerin doğruluğuyla sınırlanmıştır. Test bataryaları ancak bilişsel yeteneğin önceden var olan menzili dahilinde doğrulanabiliyorken bu menzilin dışındaki aracılara uygulanması gerektiğinden, test inşası problemi daha da karmaşık hale gelmektedir. Bu problem küçük geliştirme adımları atarak ve testin uzun zaman sürmesine izin vererek biraz hafifletilebilir. Fakat bu ihtiyatlılık masrafları artıracak ve yavaş ilerleme doğuracaktır (eğer bir yarış dinamiği söz konusuysa, bu güvenlik önlemlerini alan projenin kendisini kötü bir konuma yerleştirmesi anlamına gelir bu). Küçük adımlarla ve kapsamlı testlerle bile, yine de incelikli risklerden kaçınmak mümkün olmayabilir ve bu da hedef sisteminin tedricen yozlaşmasına yol açar; bu yozlaşma ancak sistem süper zekâya eriştiğinde görünür hale gelebilir. Bu risk test koşulunun bilişsel gelişmenin birçok adımına yayılmasına izin vererek azaltılabilir, böylece ilk başta sağlıklı olduğu tasdik edilmiş ve sonrasında birkaç geliştirme almış olan aracılar bile değiştirilmemiş ve yozlaşmamış özgün aracıların oluşturduğu bir panel tarafından değerlendirilmeye devam edecek ve geriye dönüşe tabi kalacaklardır.

İdeal şekliyle, testi ilk baştaki değerlendirme dönemiyle sınırlamak yerine sürekli kılmak gerekir. Davranışları izlenen insan düzeyinde alt aracıların strateji üretme ve dolayısıyla belli hedefleri gizlemeyi tercih etme yetenekleri vardır. Sürekli teste tabi tutabilmek için, belli bir yeterlilik düzeyindeki alt aracıların biraz daha yüksek yeterlilik düzeyindeki alt aracıları izlemekle görevlendirildikleri bir hiyerarşi oluşturulabilir. Yeterlilik hiyerarşisinin dibinde (ama güç hiyerarşisinin tepesinde) görece aptal ve yavaş asli oturuyor olacaktır. Bu insan, eli ayağı tutan bir halkı yöneten vasat bir yönetimi denetleyen yetkisiz bir saraya hükmeden bunak bir krala benzeyecektir. Yahut daha yeni ve çok daha akıllı yeni korteksin besleme ve üreme gibi hedefler için kullanılmasını sağlayan evrimsel açıdan eski "sürüngen beyni"ne benzetilebilir.

Bu türden ters bir meritokrasinin istikrarını koruyup koruyamayacağı sorgulanabilir. Eğer süper zeki alt aracıların çoğunluğunun yozlaşmış olduğu anlaşılırsa ve tüm şeyden sorumlu insanlar önceki geliştirmelerin çoğunu geri çevirmeye çalışırlarsa, kurumsal yapı bu tür bir hamlenin yaratacağı karşı çıkışa direnecek kadar güçlü olacak mıdır?

Tarihsel ya da çağdaş beşeri toplumlar bağlamında cevap herhalde mantıksız gelecektir. (Birçok ehil insanın düzenli olarak iktidara gelmesi değil mesele; ama *resmen* şempanzeler tarafından da yönetilmiyoruz ki!) Fakat toplumun yazılım olarak hayata geçirildiği bir bağlamda mesele daha muğlaktır.[18]

İki proleter başına son derece eğitimli bir polis ajanının düştüğü bir dijital hiyerarşi hayal edin. Ayrıca her polis ajanının başında da bir özel polis ajanı olduğunu varsayalım. Keza her iki özel polis ajanına da bir özel-oğlu-özel polis ajanı düşmekte ve bu böyle gitmektedir. Dolayısıyla denetleyicilerin her bir katmanında denetlediği katmandaki insan sayısının en az yarısı kadar insan bulunmaktadır. Ayrıca bu toplumun bir yöneticinin astlarına karşı büyük üstünlüklere sahip olacağı şekilde tasarlandığını da hayal edin. Örneğin yönetici, astlarının söylediği ve yaptığı her şeyi izleyip kaydedebilirken, yöneticinin kendisi tek taraflı bir aynanın arkasına gizlenmiştir ve dolayısıyla astları sadece üstlerinin aktarmayı tercih ettiği emir ve talimatları duyabilmektedir. Yöneticilerin elinin altında istedikleri zaman basabildikleri ve bir astın cezalandırılmasını ya da ödüllendirilmesini, durdurulmasını, silinmesini ya da önceki bir duruma geri döndürülmesini sağlayan birtakım düğmeler olduğunu varsayın. Bu yeterlilikler yönetici aracının güçlerini ciddi ölçüde –tarihsel açıdan en totaliter örgütlerde bile mümkün olmayan derecede– artıracaktır.

Fakat bu bile mümkün olan kontrol araçlarının hepsi değildir. Yöneticilerin ellerinin altında astlarının farazi senaryolara (bir komployа katılma ya da kimsenin izlemediğini düşündüğü bir anda kaytarma önerileri gibi) tepkilerini ölçmek için kullanabildikleri gerçekçi sanal gerçeklik simülasyonları da olabilir. Dahası yöneticiler astlarının zihinlerindeki iç monologlara kulak misafiri olabilir ve onların duygusal durumlarını doğrudan manipüle edebilirler.

Sonuçta, yönetici için muazzam üstünlükler söz konusudur.[19] Her yönetici başka, daha üst seviye bir yöneticiye tabi olacak, *onun* her hareketi izlenecek ve yöneticilik görevlerini yaparken kaytarmaması sağlanacaktır. Bu tür bir düzenlemeyle, her yönetici düzeyi yöneticilik yaptığı düzeyden entelektüel açıdan biraz daha zayıf olsa bile, yapı istikrarlı olabilir.[20] Ayrıca son derece süper zeki pek çok aracının yer alabileceği tüm yapı, üst üye erişimine sahip az sayıdaki insan tarafından kontrol edilebilir.[21]

Bu tür bir iç kontrol yapısının bedeli ağır olsa da, altından kalkılamaz diye bir kaide yoktur. Örneğin her bir denetleme kademesinin altındaki kademede yarısı kadar insanın bulunduğu bir modelde, fazladan hesaplama masrafları sistemin proleter parçasının tuttuğu masrafın sadece yüzde yüzüyle sınırlı olacak, üstelik aptal patron katmanları alt aracı başına daha az bilgi işlem isterlerse bu sayı daha da düşecektir. Denetleme, oranı artırarak sıkılaştırılabilir ya da oran azaltılarak masraflar azaltılabilir. Fazladan hesaplama masrafı, rekabetçi bir piyasa senaryosunda asli bir etken olacaktır (karş. 11. Bölüm), ama doğrudan güçlü bir rekabetle karşılaşmayan bir proje bağlamında altın-

dan kalkılabilir görünmektedir. Ayrıca bu denetleme işlevlerini yaratıp test etmek için gerekli gelişim zamanı bakımından bir masraf kalemi çıkacaktır. Kaynakları sağlam bir proje, kontrol yapısının gelişimini makine zekâsının geliştirilmesine paralel yürüterek bu zaman masrafını azaltabilir; ama fazladan görev yükü küçük projeler için ve yakın bir teknoloji yarışı içindeki projeler için bel bükücü olabilir.

Burada dikkate alınmayı hak eden bir başka masraf kalemi daha var: Bu tür bir yapıda işlenen zihin suçları riski.[22] Belirtildiği üzere, bu kurum çok korkunç bir Kuzey Kore çalışma kampını akla getiriyor. Ama bu tür bir kurum işletmenin doğurduğu ahlaki problemleri en azından azaltmanın yolları, kurumdaki alt aracılar tam anlamıyla insani ahlaki statüye sahip emülasyonlar olsalar bile, yok değildir. Hiç olmazsa, sistem gönüllü emülasyonlara bel bağlayabilir. Her alt aracının herhangi bir anda katılımını geri çekme seçeneği olabilir.[23] Sonlandırılmış emülasyonlar belleğe kaydedilebilir ve zekâ patlamasının tehlikeli aşaması geçtiğinde çok daha ideal koşullarda yeniden başlatma taahhüdü verilebilir. Diğer yandan katılmayı tercih eden alt aracılar çok rahat sanal çevrelerde konuk edilip uyumaları ve dinlenmeleri için bol bol zaman tanınabilir. Bu önlemler masraf demektir, ama rekabetçi olmayan koşullarda iyi kaynakları olan bir proje için atla deve değildir. Son derece rekabetçi bir durumdaysa, bir işletmeye rakiplerinin de aynı masrafları yükleneceği garantisi verilmedikçe masraflar bel bükücü olabilir.

Bu örnekte alt aracıları emülasyon olarak tahayyül ettik. Şöyle bir soru akla gelebilir: Kurum tasarımı yaklaşımında alt aracıların insan biçiminde olması farz mıdır? Veya yapay alt aracılardan oluşan sistemler için de aynı derecede geçerli midir?

Burada ilk akla gelen düşünce şüpheci olabilir. İnsan benzeri aracılarla çok sayıda deneyimimiz olmasına karşın, henüz devrimlerin nasıl patlak vereceğini ya da sonuçlarını kesin olarak öngöremiyoruz; sosyal bilimler en fazla bazı istatistiksel eğilimleri tarif edebiliyor.[24] Toplumsal yapıların sıradan insanlar için istikrarını güvenilir şekilde kestiremediğimizden (oysa elimizde birçok veri vardır), bilişsel açıdan geliştirilmiş insan benzeri aracılar için hassasiyet mühendisliği yapan istikrarlı toplumsal yapılar konusunda çok az umudumuz olduğu (bu konuda elimizde hiçbir veri yoktur) ve bunu ileri yapay aracılar için yapma umudumuzunsa daha da az olduğu (hakkında veriye sahip olduğumuz aracılara benzer bile değildir) sonucunu çıkarmak çekicidir.

Fakat mesele o kadar basit değildir. İnsanlar ve insan benzeri varlıklar karmaşıktır, ama yapay aracıların görece basit mimarileri olabilir. Ayrıca yapay aracıların basit ve açıkça tanımlanmış güdülenmeleri olabilir. Dahası genel olarak dijital aracılar (ister emülasyonlar olsun ister yapay zekâlar) kopyalanabilir: Bu durum tıpkı birbirinin yerine geçirilebilir parçaların imalatta devrime yol açması gibi, işletmecilikte ya da idarecilikte devrime yol açabilir. Bu farklar, ilk başta güçsüz olan aracılarla çalışma ve yukarıda bahsi geçen çeşitli kontrol önlemlerinden yararlanan kurumsal yapılar

yaratma fırsatıyla birleştiğinde, belli kurumsal sonuçlara (bir sistemin isyan etmemesi gibi) ulaşmayı tarihsel koşullarda insanlarla birlikte çalışırken olduğundan çok daha güvenilir bir şekilde mümkün hale getirebilir.

Ama bir kez daha, yapay aracılarda insan türü aracıların davranışlarını kestirmemize yardımcı olacak özelliklerin pek çoğu olmayabilir. Yapay aracıların insan davranışını bağlayan toplumsal duygulardan (korku, gurur ya da pişmanlık gibi duygulardan) herhangi birine sahip olması gerekmiyor. Keza aileye ve arkadaşlara bağlılıklar geliştirmeleri de gerekmiyor. Ayrıca biz insanların niyetlerimizi gizlememizi zorlaştıran bilinçdışı beden dili sergilemeleri de gerekmiyor. Bu eksiler yapay aracıların kurumlarını istikrarsızlığa sürükleyebilir. Ayrıca yapay aracılar algoritmalarındaki ya da mimarilerindeki görünüşte küçük değişiklikler sonucunda bilişsel performanslarında büyük sıçramalar kaydedebilirler. Acımasız derecede optimumlaştırıcı yapay zekâlar insanların sakınacakları aşırı risklere girmeye hevesli olabilirler.[25] Ayrıca süper zeki aracılar çok az ya da sıfır iletişimle eşgüdümde bulunma konusunda şaşırtıcı bir yetenek sergileyebilirler (örneğin birbirlerinin çeşitli olumsallıklara farazi yanıtlarını içsel olarak örnek almak suretiyle). Bu ve diğer farklılıklar ani kurumsal başarısızlığı, görünüşte aşılmaz toplumsal kontrol yöntemleri olarak görülebilecek durumlar karşısında bile daha muhtemel hale getirebilir.

Dolayısıyla kurum tasarım yaklaşımının ne kadar ümit vaat edici olduğu ve yapay aracılardansa insanbiçimsel aracılarla çalışma ihtimallerinin çok olup olmadığı belirsizdir. Uygun denge ve frenlere sahip bir kurum yaratmanın ancak güvenliği artırabileceği (ya da her halükârda güvenliği azaltmayacağı) ve dolayısıyla risk önleme perspektifinden değerlendirildiğinde bu yöntemin kullanılmasının her zaman en iyisi olduğu düşünülebilir. Ama bu da kesin değildir. Bu yaklaşım parçaları ve karmaşıklığıyla, zeki alt aracı parçaları olmayan bir aracı örneğinde görülmeyen şekilde işlerin ters gittiği bir durum yaratabilir. Yine de kurum tasarımı daha fazla irdelenmeyi hak etmektedir.[26]

Özet

Hedef sistem mühendisliği henüz yerleşik bir disiplin değildir. Şu an, insani değerlerin bir dijital bilgisayara nasıl aktarılacağı, insan düzeyinde makine zekâsı olduğunda bile bilinmiyor. Çeşitli yaklaşımları araştırdıktan sonra, bazılarının çıkmaz sokağa vardığını tespit ettik; ama diğerleri ümit vaat ediyor ve daha fazla araştırılmayı hak ediyor görünmektedir. **TABLO 12**'de bir özet sunulmuştur.

Eğer değer yükleme problemini nasıl çözeceğimizi bilseydik, başka bir problemle daha karşılaşırdık: Hangi değerlerin yükleneceği problemi. Başka bir deyişle, bir süper zekânın ne istemesini isterdik? Bir sonraki bölümde ele alacağımız daha felsefi problem budur.

TABLO 12: Değer yükleme tekniklerinin özeti

Açıktan temsil	Evcimenlik değerleri yüklemek için umut vaat edici bir yol olabilir. Daha karmaşık değerler yüklemek için ise umut vaat edici bir yol gibi görünmüyor.
Evrimsel seçilim	Daha az umut vaat edicidir. Güçlü arama formel arama kıstaslarını karşılayan, ama niyetlerimizi karşılamayan bir tasarım bulabilir. Ayrıca tasarımlar bunları işleterek değerlendirilirse (formel kıstasları bile karşılamayan tasarımlar dahil) potansiyel açıdan vahim bir tehlike daha baş gösterir. Evrim muazzam zihin suçlarından kaçınmayı da zorlaştırır, özellikle de insan türü zihinler yaratmak isteniyorsa.
Takviyeli öğrenme	"Takviyeli öğrenme problemleri"ni çözmek için birtakım farklı yöntemlerden yararlanılabilir, ama bunlar tipik şekliyle bir ödül sinyalini azamiye çekmeye çalışan bir sistem yaratmayı içerir. Bu durum, sistem daha zeki hale geldiğinde elektronik kafalama [***wirehead***] başarısızlık tarzını doğurmak gibi içkin bir eğilime sahiptir. Dolayısıyla takviyeli öğrenme umut vaat edici görünmüyor.
Değer artırımı	Biz insanlar özgül hedef içeriğimizin çoğunu deneyimlere tepkilerimizden elde ederiz. Değer artırımı ilkesel olarak insan güdüleri olan bir aracı yaratmak için kullanılabilirse de, insandaki değer artırımı temayülleri karmaşık ve bir tohum yapay zekâda çoğaltılması zor olabilir. Kötü bir kestirim, insanlardan farklı şekilde genelleyen ve dolayısıyla amaçlanmamış nihai hedefler edinen bir yapay zekâ doğurabilir. Değer artırımının yeterli kesinlikte işlemesini sağlamanın ne kadar zor olacağını belirlemek için daha fazla araştırmaya ihtiyaç vardır.
Güdülenme iskeleti	Bir sistemi (yeterliliklerini tehlikeli düzeyin altında tutarken) insanlar için şeffaf olan üst düzey iç temsiller geliştirmeye teşvik etmenin ve ardından bu temsilleri yeni bir hedef sistemi tasarlamak için kullanmanın ne kadar zor olacağını söylemek için henüz çok erkendir. Yaklaşım epey bir ümit vaat ediyor olabilir. (Gelgelelim insan düzeyinde yapay zekânın gelişimine kadar güvenlik mühendisliğine ilişkin sıkı çalışmanın çoğunu erteleyecek olan bütün test edilmemiş yaklaşımlarda olduğu gibi, burada da bunun kontrol problemine laubali bir yaklaşım için mazeret haline gelmemesi için dikkat etmek gerekir.)

Değer öğrenme	Potansiyel açıdan ümit vaat edici bir yaklaşımdır, ama insan değeri hakkında ilintili dışsal bilgiye başarıyla işaret eden bir göndermeyi formel olarak saptamanın ne kadar zor olacağını (ve bu tür bir gönderme üzerinden fayda işlevi için bir doğruluk kıstası saptamanın ne kadar zor olacağını) belirlemek için daha fazla araştırma gerekecektir. Keza değer öğrenme kategorisi dahilinde Yüce Meryem türü öneriler ya da Paul Christiano'nu yaklaşımı (veya başka bu tür kısayollar da) araştırılmaya değer.
Emülasyon modülasyonu	Eğer makine zekâsı emülasyon yolundan giderek elde edilirse, ilaçların dijital eşdeğerleri ya da başka yollarla güdülenmeleri değiştirmesi muhtemelen mümkün hale gelecektir. Bunun emülasyon süper zekâya dönüştürüldüğünde bile güvenliği sağlayacak kadar kesinliğe kavuşulmasını sağlayıp sağlamayacağı tartışmaya açıktır. (Etik kısıtlamalar da bu yöndeki gelişmeleri karmaşıklaştırabilir.)
Kurum tasarımı	Toplumsal kontrol konusunda çeşitli güçlü yöntemler emülasyonlardan oluşan bir kurumda uygulanabilir. İlkesel açıdan, toplumsal kontrol yöntemleri yapay zekâlardan oluşan bir kurumda da kullanılabilir. Emülasyonların bu tür yöntemler aracılığıyla kontrol etmeyi daha kolaylaştıran bazı özellikleri vardır, ama yapay zekâlarla kıyaslandığında kontrol edilmelerini zorlaştırabilecek bazı özellikleri de olabilir. Kurum tasarımı potansiyel bir değer yükleme tekniği olarak daha fazla araştırılmaya değer görünmektedir.

ON ÜÇÜNCÜ BÖLÜM

Seçme Kıstaslarını Seçme

Bir tohum yapay zekâya kafamıza esen nihai hedefi yükleyebildiğimizi varsayalım. O zaman, hangi değerin yüklenebileceği kararı en uzun erimli sonuçlara sahip olabilir. Diğer birtakım temel parametre seçimlerinin (yapay zekânın karar teorisine ve epistemolojisine dair aksiyomların) de benzer kapsamda sonuçları olabilir. Ama biz ki aptal, cahil ve dargörüşlüyüz, iyi kararlar alabileceğimize nasıl güvenilebilir? Mevcut kuşağın önyargılarını ve peşin hükümlerini sonsuza dek yanımızda taşımadan nasıl seçim yapabiliriz? Bu bölümde sonuç hâlâ en derin insani değerlere kök salmışken, dolaylı normatifliğin bize bu kararları bizatihi süper zekâya taşırken karşılaşacağımız bilişsel çalışmanın çoğundan nasıl kurtulmamızı sağlayabileceğini inceleyeceğiz.

Dolaylı normatiflik ihtiyacı

Bir süper zekânın istediğimiz şeyi yapmasını nasıl sağlayabiliriz? Süper zekâdan ne yapmasını isteriz? Bu noktaya kadar ilk soruya odaklanmıştık. Şimdi ikinci soruya dönüyoruz.

Kontrol problemini çözdüğümüzü ve seçtiğimiz herhangi bir değeri bir süper zekânın güdülenme sistemine yükleyebildiğimizi, böylece bu değeri nihai değeri olarak sahiplenmesini sağlayabildiğimizi varsayalım. Hangi değeri yüklememiz gerekir? Yapılacak seçim hiç de önemsiz değildir. Eğer süper zekâ belirleyici bir stratejik üstünlük elde ederse, değer; kozmik birikimden nasıl yararlanacağını belirleyecektir.

Çok açık ki, değer seçimimizde yanlış yapmamamız aslidir. Ama bunun gibi bir meselede hatasız kalmayı umut etmek gerçekçi midir? Ahlak konusunda yanılabiliriz; keza bizim için neyin iyi olduğu konusunda yanılabiliriz; hatta gerçekten ne istediğimiz konusunda da yanılabiliriz. Bir nihai hedef belirlemenin, insanın dikenli felsefi problemler ormanında yolunu bulmasını gerektirdiği görünmektedir. Eğer dolaysız bir yaklaşım denersek, işleri yüzümüze gözümüze bulaştırmamız muhtemeldir. Hatalı seçim yapma riski karar bağlamı tanıdık olmadığında özellikle yüksektir – ve tüm insanlığın geleceğini şekillendirecek bir makine süper zekâsı için nihai hedef seçmek bize son derece yabancı bir bağlamdır.

Cephe saldırısında şansımızın az olduğunun bir yansıması da değer teorisindeki ilintili meselelere dair görüş ayrılığının yaygınlığıdır. Felsefeciler arasında çoğunluğun desteğini alan tek bir etik teorisi yoktur, dolayısıyla çoğu felsefeci yanılıyor olmalıdır.[1] Ayrıca ahlaki inançların zaman içinde geçirdiği ve bizim bir ilerleme olarak görmekten hoşlandığımız belirgin değişikliklerde de yansımasını bulmuştur. Sözgelimi ortaçağda Avrupa'da bir siyasi mahpusun ölümüne işkence edildiğini izlemek saygın bir eğlence kabul edilirdi. Kedi yakmak 16. yüzyıl Paris'inde popülerdi.[2] Çok değil yüz elli yıl önce Amerika'nın güneyinde kölelik çok yaygındı ve arkasına hem yasaları hem de âdetleri almıştı. Geriye dönüp baktığımızda, önceki çağların tamamının sadece davranışlarında değil ahlaki inançlarında da apaçık kusurlar görüyoruz. O zamandan beri belki bazı ahlaki içgörüler elde etmiş olsak da, şu an kusursuz ahlaki aydınlanmanın tepe noktasında olduğumuzu iddia edemeyiz. Büyük ihtimalle, bir ya da daha fazla ciddi ahlaki yanlış anlayışın cenderesinde debeleniyoruz. Bu tür koşullarda sonsuza dek devam edecek ve her türlü etik ilerleyişi engelleyecek şekilde mevcut inançlarımıza dayalı bir nihai değer seçmek varoluşsal bir ahlaki belayı göze almak olacaktır.

Doğru etik teoriyi bulduğumuzdan aklımızla emin olabilsek bile (ki olamayız), yine de bu teorinin önemli ayrıntılarını geliştirmekte hatalar yapma riski baki kalacaktır. Görünüşte basit ahlak teorilerinin altında birçok karmaşıklık saklı olabilir.[3] Örneğin hedonizmin (alışılmadık derecede basit) sonuççu teorisini düşünün. Bu teori, kabaca, her türlü hazzın değeri olduğunu ve sadece hazzın değeri olduğunu; her türlü acının değersiz olduğunu ve sadece acının değersiz olduğunu varsayar.[4] Tüm ahlaki çiplerimizi bu tek teoriye yerleştirsek ve teorinin doğru olduğu anlaşılsa bile, yine de birçok soru cevapsız kalacaktır. John Stuart Mill'in savunduğu gibi, "aşağı hazlar"a karşı "yüksek hazlar"a öncelik verilmeli midir? Bir hazzın yoğunluğu ve süresi nasıl hesaplanmalıdır? Acılar ile hazlar birbirini götürebilir mi? Ahlaken ilintili hazlarla ne tür beyin durumları alakalıdır? Aynı beyin durumunun tastamam iki kopyası haz miktarının iki katına mı tekabül edecektir?[5] Bilinçdışı hazlar olabilir mi? Son derece küçük olasılıkta olan son derece büyük hazlarla nasıl başa çıkılmalıdır?[6] Sınırsız nüfuslarla toplaşmayı nasıl başarabiliriz?[7]

Bu sorulardan herhangi birine yanlış cevap vermek feci sonuçlar doğurabilir. Eğer süper zekâ için bir nihai değer seçerek sadece bir genel ahlak teorisine değil, bu teorinin yorumlanarak etkili bir karar alma sürecine nasıl dahil edileceğine dair birçok özgül iddiaya bel bağlamamız gerekseydi, o zaman şansımızın yaver gitme ihtimali neredeyse sıfıra yakın olurdu. Aptallar tohum yapay zekâya kendi gözde cevaplarını yerleştirmek için ahlak felsefesinin tüm önemli problemlerini tek bir hamlede çözme zorluğunu bir hevesle kabul edebilirler. Akıllılar ise bir alternatif yaklaşım, dolaylı konuşmak için harıl harıl bir yol arayacaktır.

Bu bizi dolaylı normatifliğe götürüyor. Bir süper zekâ yaratmak için bariz neden, verili bir değeri gerçekleştirmenin etkili yollarını bulmak için gerekli araçsal muhakemeyi ona yükleyebilecek olmamızdır. Dolaylı normatiflik gerçeğe dönüştürülecek değeri seçmek için gerekli muhakemenin bir kısmını süper zekâya yüklememizi de sağlayacaktır.

Dolaylı normatiflik gerçekten ne istediğimizi, neyin çıkarımıza olduğunu ya da ahlaken neyin doğru ya da ideal olduğunu bilemeyebileceğimizin sunduğu zorluğu cevaplamanın bir yoludur. Kendi mevcut anlayışımıza (muhtemelen son derece kusurludur) dayalı bir tahminde bulunmak yerine, değer seçimi için gerekli bilişsel çalışmanın bir kısmını süper zekâya devredeceğiz. Süper zekâ bilişsel çalışmada bizden daha iyi olduğundan, düşüncemizi bulandıran geçmiş hataları ve kafa karışıklıklarını görebilir. Bu fikri genelleştirebilir ve bir bulgulayıcı ilke haline getirebiliriz:

> *Epistemik saygı ilkesi*
>
> Bir müstakbel süper zekâ epistemik açıdan üstün bir bakış açısına sahiptir: İnançları (büyük ihtimalle, birçok konuda) bizim inançlarımızın doğru olmasından daha muhtemeldir. Bu nedenle süper zekânın görüşlerine, gerçekleştirilebilir olduğu her durumda, saygı göstermemiz gerekir.[8]

Dolaylı normatiflik bu ilkeyi değer seçme problemine uygular. Somut bir normatif standart belirleyebilme yeteneğimize güven duymadığımızdan, bunun yerine, bir süper zekânın soyut koşulu karşılayan bir somut standart bulabileceği umuduyla herhangi bir normatif standardın karşılayabileceği daha soyut bir koşul saptarız. Bir tohum yapay zekânın önüne nihai hedef olarak, zımnen tanımlanmış bu standardın ondan ne yapmasını isteyeceğine dair en iyi tahmine göre sürekli hareket etmeyi koyabiliriz.

Bazı örnekler bu fikrin daha net hale gelmesine hizmet edecektir. İlkin Eliezer Yudkowsky'nin anahatlarını ortaya koyduğu bir dolaylı normatiflik önerisi olan "bütünlüklü dışkestirilmiş irade"yi ele alacağız. Ardından, mevcut seçeneklere dair bir fikir edinmek adına bazı çeşitlilik ve alternatiflerden bahsedeceğiz.

Bütünlüklü dışkestirilmiş irade

Yudkowsky bir tohum yapay zekânın nihai hedefinin insanlığın "bütünlüklü dışkestirilmiş irade"sini (CEV) yerine getirmek olması gerektiğini önermiştir. Bu önerisini şöyle tanımlar:

> Bizim bütünlüklü dışkestirilmiş irademiz, eğer daha fazla şey bilseydik, daha hızlı düşünseydik, daha fazla insan bizim istediğimiz kişi olsaydı, birlikte daha fazla (dışkestirimin ıraksamaktan ziyade yakınsadığı, isteklerimizin birbirine müdahale etmekten ziyade birbiriyle bütünleştiği; dışkestirimini istediğimiz gibi dışkestirildiği,

yorumlanmasını istediğimiz gibi yorumlandığı yere kadar) büyüseydik var olacak isteğimizdir.[9]

Yudkowsky bunları yazdığında, bu epey şiirsel reçetenin nasıl hayata geçirileceğine dair bir mavi kopya sunmayı amaçlamamıştı. Amacı CEV'in nasıl tanımlanabileceğine ilişkin bir ön taslak ve neden bu çizgide bir yaklaşımın gerekli olduğuna ilişkin bazı argümanlar sunmaktı.

CEV önerisinin arkasındaki birçok fikrin felsefe yazınında benzeşleri ve öncelleri vardır. Örneğin etikte *ideal gözlemci teorileri* "iyi" ya da "kötü" gibi normatif kavramları farazi bir ideal gözlemcinin dile getireceği yargılar üzerinden tahlil etmeye çalışır (burada "ideal gözlemci" ahlaki olmayan olgular hakkında her şeyi bilen, mantıksal olarak ileri görüşlü, konuyla ilintili olarak tarafsız ve çeşitli önyargılardan azade vb. biri olarak tanımlanır).[10] Gelgelelim CEV yaklaşımı bir ahlak teorisi değildir (ya da öyle yorumlanması gerekmez). Değer ile CEV'imizin tercihleri arasında zorunlu bir bağlantı olduğu iddiasında değildir. CEV nihai değere sahip olan her şeye yaklaşmanın yararlı bir yolundan ibaret görülebilir ya da etikle herhangi bir bağlantıdan ayrı düşünülebilir. Dolaylı normatiflik yaklaşımının ana prototipi olarak biraz daha ayrıntılı incelenmeyi hak ediyor.

Bazı açımlamalar

Yukarıdaki alıntıda yer alan bazı terimler açıklama gerektiriyor. Yudkowsky'nin terminolojisinde "daha hızlı düşünmek" demek, *daha akıllı olsaydık ve şeyleri enine boyuna daha fazla düşünseydik* demektir. "Birlikte daha fazla büyüseydik" ise *öğrenimimizi, bilişsel ilerlememizi ve kendimizi geliştirmemizi birbirimizle uygun toplumsal etkileşim koşullarında gerçekleştirseydik* anlamına geliyor görünmektedir.

Keza "dışkestirimin ıraksamaktan ziyade yakınsadığı" ifadesiyse şöyle anlaşılabilir. Yapay zekâ dışkestiriminin sonucunun bir özelliğine ancak bu özellik yapay zekâ tarafından son derece yüksek güvenle kestirilebileceği ölçüde etkide bulunmalıdır. Yapay zekâ belirtilen tarzda ideal hale gelseydik ne isteyeceğimizi tahmin edemeyeceği ölçüde, kel alaka bir tahminde bulunmamalı, bunun yerine harekete geçmekten sakınmalıdır. Gelgelelim idealleştirilmiş isteğimizin birçok ayrıntısı belirsiz ya da kestirilemez olabilirse de, yapay zekânın algılayabileceği bazı genel çizgiler olabilir ve en azından olayların ilerideki seyrinin bu çerçevede gerçekleşeceğini sağlayacak şekilde hareket edebilir. Örneğin eğer yapay zekâ dışkestirilmiş irademizin hepimizin sürekli acı içinde kıvranmamamızı ya da evrenin kâğıt kıskaçlarıyla doldurulmamasını isteyeceğini güvenilir şekilde kestirebilirse, o zaman yapay zekâ bu sonuçları engelleyecek şekilde harekete geçmelidir.[11]

"İsteklerimizin birbirine müdahale etmekten ziyade birbiriyle bütünleştiği" ise şöyle yorumlanabilir: Yapay zekâ birey olarak insanların dışkestirilmiş iradeleri arasında son

derece yüksek bir uyum olduğunda harekete geçmelidir. Az sayıda güçlü, açık istekler bazen bir çoğunluğun zayıf ve karışık isteklerine ağır basabilir. Ayrıca Yudkowsky yapay zekânın dar şekilde saptanmış tikel bir sonucu engellemek için daha az fikirbirliğine ve yapay zekânın geleceği neyin iyi olduğuna dair daha tikel bir dar anlayışa yönlendirecek şekilde harekete geçmesi için daha fazla fikirbirliğine ihtiyaç duyması gerektiğini düşünüyor. "CEV'in ilk dinamiği," diye yazıyor, "'evet' deme konusunda muhafazakâr olmalı ve 'hayır' için dikkatle dinlemelidir."[12]

"Dışkestirimini istediğimiz gibi dışkestirildiği, yorumlanmasını istediğimiz gibi yorumlandığı": Bu son ifadelerin arkasındaki fikir, dışkestirim kurallarının bizatihi dışkestirilmiş iradeye duyarlı olması gerektiğine dair kurallar olarak görünmektedir. Bir bireyin, iradesi dışkestirildiğinde birinci derece arzularından bazılarına ağırlık verilmemesi gerektiğine dair ikinci derece bir arzusu (ne arzulayacağına dair bir arzusu) olabilir. Örneğin içki konusunda bir birinci derece arzusu olan bir alkoliğin birinci derece arzusu olmaması için bir ikinci derece arzusu da olabilir. Benzer şekilde, dışkestirim sürecinin diğer çeşitli parçalarının nasıl gerçekleşmesi ve bizim bunları dışkestirim süreciyle nasıl hesaba katmamız gerektiğine dair arzularımız olabilir.

İnsanlığın bütünlüklü dışkestirilmiş iradesi kavramı düzgün şekilde tanımlanabilse bile, yine de –bir süper zekâ için bile– CEV yaklaşımında şart koşulan farazi ideal koşullarda insanlığın gerçekten ne isteyeceğini bulmanın imkânsız olabileceği itirazı yöneltilebilir. Yapay zekâ dışkestirilmiş irademizin içeriğine dair bir bilgi olmadan, davranışına rehberlik edecek somut bir standarttan mahrum olacaktır. Gelgelelim insanlığın CEV'inin ne isteyeceğini kesin olarak bilmek zor olacak olsa da, bilinçli tahminler yapmak mümkündür. Bugün süper zekâ olmadan bile mümkündür. Örneğin CEV'in gelecekte karanlık bir odada bir sandalyenin üstüne tünemiş ve acı çeken insanlardansa zengin ve mutlu hayatlar yaşayan insanlar olmasını isteyeceğini düşünmek daha akla yatkındır. Eğer *biz* en azından bazı makul yargılarda bulunabiliyorsak, bir süper zekâ da bulunabilir. Dolayısıyla süper zekânın tutumuna başlangıçtan itibaren CEV'imizin içeriğine dair tahminleri kılavuzluk edebilir. Bu ilk tahminlerini inceltmek için güçlü bir araçsal nedeni olacaktır (örneğin insan kültürünü ve psikolojisini inceleyerek, insan beyinlerini tarayarak ve eğer daha fazla bilseydik, daha düzgün düşünseydik nasıl davranabileceğimize ilişkin muhakeme yürüterek). Bu meseleleri araştırırken yapay zekâya CEV'imizin ilk tahminleri kılavuzluk edecektir; böylece sözgelimi eğer CEV'imizin zihin suçu gibi simülasyonları muhtemelen kınayacağı tahmin edilirse, yapay zekâ ebedi insan ıstırabıyla dolu birçok simülasyonu gereksiz yere denemeyecektir.

Bir diğer itiraz ise şudur: Dünyada o kadar farklı yaşam tarzı ve ahlak anlayışı vardır ki, bunları birleştirip CEV haline getirmek mümkün olmayabilir. Hatta bunları birleştirip kaynaştırsak bile, sonuç pek iştah açıcı olmayabilir: Herkesin en sevdiği

farklı yemeklerden en iyi tatları karıştırmanın, ortaya lezzetli bir yemek çıkarması muhtemel değildir.[13] Buna yanıt olarak, CEV yaklaşımının tüm yaşam tarzlarının, ahlak anlayışının ya da kişisel değerin birleştirilip bir türlü yemeği haline getirilmesini gerektirmediğine dikkat çekilebilir. CEV dinamiğinin sadece isteklerimiz uyuşup bütünlüklü hale geldiğinde devreye geçtiği varsayılır. Yaygın derecede uzlaşmaz bir fikir ayrılığının olduğu konularda, çeşitli idealleştirici koşullar dayatıldıktan sonra bile, dinamik sonucu belirlemekten sakınmalıdır. Yemek pişirme analojisine devam etmek gerekirse, bireylerin veya kültürlerin gözde yemekleri farklı farklı olabilir, ama yine de besinlerin toksik olmaması gerektiği konusunda geniş bir mutabakata sahip olabilirler. O zaman CEV dinamiği zehirlenmeyi engelleyecek şekilde hareket ederken, insanların yemek pişirme alışkanlıklarına karışmadan dilediklerini yapmalarına izin verebilir.

CEV'i destekleyici gerekçeler

Yudkowsky'nin makalesinde CEV yaklaşımını savunan yedi argüman sunulmuştu. Bunlardan üçü temelde şu düşünceyi farklı şekillerde ortaya koyuyordu: Amaç insani ve işe yarar bir şey yapmak olmalıyken, niyetlenilmemiş yorumlar ve istenmeyen sonuçlar doğurmayacak birtakım açık kurallar koymak çok zor olacaktır.[14] CEV yaklaşımının güçlü ve kendi kendini düzeltici olması planlanmıştır; bizim âli değerlerimizin her birini her seferinde doğru şekilde sıralamamıza ve dillendirmemize yaslanmak yerine, değerlerimizin *kaynağını* yakalaması planlanmıştır.

Diğer dört argüman ise bu ilk temel (ama önemli) hususun ötesine geçiyor ve değer saptama problemine aday çözümlerde gerekli nitelikleri dile getirerek, CEV'in bu gereklilikleri karşıladığını savunuyor.

"Ahlaki büyümeyi sarmala"

Çözümün ahlaki ilerleme olanağına izin vermesi gerekliliğidir bu. Daha önce belirtildiği üzere, mevcut ahlaki inançlarımızın birçok açıdan kusurlu, belki de derinden kusurlu olduğuna inanmak için nedenler vardır.

Eğer yapay zekânın takip etmesi için özgül ve değişmez bir ahlak anlayışı şart koşarsak, aslında hatalarımız da dahil olmak üzere mevcut ahlak anlayışlarımızı geleceğe taşımış ve her türlü ahlaki büyüme umudunu doğmadan öldürmüş oluruz. Halbuki CEV yaklaşımı bu tür bir büyüme olanağına izin verir, çünkü yapay zekâya eğer olumlu koşullarda daha da ilerleseydik isteyeceğimiz şeyi yaptırmaya çalışır ve eğer biz bu şekilde gelişmiş olsaydık, ahlaki inançlarımız ve duyarlılıklarımız mevcut kusurlardan ve sınırlılıklardan arınmış olurdu.

"İnsanlığın geleceğini gasp etmekten sakının"

Yudkowsky'nin aklında küçük bir grup programlayıcının bir tohum yapay zekâ yarattığı ve ardından onun büyük bir süper zekâ hâline gelerek belirleyici stratejik üstünlüğe eriştiği bir senaryo vardır. Bu senaryoda özgün programlayıcılar insanlığın kozmik birikiminin tamamını ellerinde tutarlar. Çok açık ki, bu hiçbir fani için hoş bir sorumluluk değildir. Fakat programlayıcıların kendilerini bu durumda bulduklarında hemen sorumluluktan kaçmaları da mümkün değildir: Yapacakları her türlü tercih (projeyi bırakmak dahil), dünya çapında tarihsel sonuçlar doğuracaktır. Yudkowsky CEV'i programlayıcıların insanlığın geleceğini belirleme ayrıcalığından ya da yükünden kurtaracak bir yol olarak görür. *İnsanlığın* (yani kendi iradelerini ya da kendi sözde ahlak teorilerini değil) bütünlüklü dışkestirilmiş iradesini hayata geçiren bir dinamikle, aslında, geleçek üzerindeki etkilerini tüm insanlığa dağıtırlar.

"Günümüz insanlarını ilk dinamik için kavga ettirecek bir güdü yaratmaktan sakının"

İnsanlığın geleceği üzerindeki etkiyi paylaştırmak, kendi gözde vizyonlarını hayata geçiren programlayıcı ekibi için ahlaken tercih edilir olmasının yanı sıra, kimin ilk süper zekâyı yaratacağına dair yarışa olan ilgiyi azaltmanın da bir yoludur. CEV yaklaşımında programlayıcılar (ya da sponsorları) sonucun içeriği üzerinde diğer insanlardan daha fazla etkiye sahip değildir, ama elbette dışkestirimin yapısının belirlenmesinde ve başka bir alternatif yerine insanlığın CEV'inin uygulanmasının kararlaştırılmasında başrol oynarlar. Çatışmadan kaçınmak sadece çatışmanın yol açabileceği dolaysız zarardan ötürü değil, aynı zamanda süper zekânın güvenle ve hayırlı bir şekilde geliştirilmesi zorluğunda işbirliğini engellediği için de önemlidir.

CEV'in geniş kesimlerden destek alabileceği hesaplanmıştır. Bunun tek nedeni nüfuzu eşit şekilde dağıtması değildir. CEV'in uzlaştırıcı, barışçıl bir potansiyele sahip olmasının köklü bir nedeni de vardır: Birçok farklı gurubun kendi tercih ettikleri gelecek vizyonunun tamamen hâkim hale geleceğini umut etmesini sağlar. Afgan Taliban'ının bir İsveç Hümanist Derneği üyesiyle tartıştığını tahayyül edin. İkisinin çok farklı dünya görüşleri vardır ve biri için ütopya olan diğeri için distopya olabilir. Keza ikisi de uzlaşmacı bir sonuçtan (mesela kız çocuklarına okula gitme hakkı verilmesi ama sadece 9 yaşına kadar okuyabilmeleri ya da İsveçli kızlara eğitim-öğretim hakkı tanınması ama Afgan kızlara tanınmaması) hazzetmeyebilir. Gelgelelim hem Taliban hem de hümanist, geleceğin insanlığın CEV'i tarafından belirlenmesi ilkesini onaylayabilir. Taliban eğer kendi dini görüşleri gerçekten doğruysa (kendisi böyle olduğuna inanmaktadır) ve eğer bu görüşlerin kabul edilmesi için sağlam gerekçeler varsa (bu konuda da inancı tamdır), o zaman, diye düşünecektir, insanlık sonunda bu görüşleri kabul edecektir, yeter ki insanlar önyargılarını ve peşin hükümlerini olabildiğince bir kenara atsınlar, Kuran'ı incelemeye daha fazla zaman ayırsınlar, dünyanın nasıl işlediğini daha açık şekilde anlayıp asli öncelikleri idrak etsinler, akıldışı isyankârlıktan

ve korkaklıktan vareste hale gelsinler, vesaire.[15] Hümanist de benzer şekilde bu ideal koşullarda insanlığın onun benimsediği ilkeleri benimseyeceğine inanacaktır.

"İnsanlık Nihayetinde Geleceğini Kendisi Belirlesin"

Bir paternalist süper zekânın bizi sürekli gözetlediği, işlerimize büyük bir plan çerçevesinde en ince ayrıntısına kadar müdahale ettiği bir sonuç istemeyebiliriz. Süper zekânın tamamen iyiliksever olmasını ve kibirden, küstahlıktan, çokbilmişlikten, dargörüşlülükten ve diğer insani kusurlardan azade olmasını şart koşsak bile, bu tür bir düzenlemenin gerektirdiği özerklik kaybından hâlâ rahatsız duyulabilir. Bazen tökezlemek pahasına olsa bile, kendi kaderimizi kendimiz çizmek isteyebiliriz. Belki de süper zekânın bir güvenlik ağı işlevi görmesini, işler feci derecede kötü gittiğinde bizi desteklemesini, aksi takdirde bizi kendi başımızın çaresine bakmak üzere yalnız bırakmasını isteriz.

CEV bu olanağa izin veriyor. CEV'in bir "başlangıç dinamiği", yani bir kez işleyen ve ardından dışkestirilmiş iradenin istediği her neyse onunla ikame edilen bir süreç olması amaçlanıyor. Eğer insanlığın dışkestirilmiş iradesi bizim paternalist bir yapay zekânın gözetimi altında yaşamamızı isterse, o zaman CEV dinamiği bu tür bir yapay zekâ yaratacak ve ipleri onun eline verecektir. Eğer insanlığın dışkestirilmiş iradesi bunun yerine demokratik bir dünya hükümeti yaratılmasını isterse, o zaman CEV dinamiği bu tür bir kurum yaratılmasını kolaylaştırıp, ardından kendisi bir kenara çekilebilir. Eğer insanlığın dışkestirilmiş iradesi her insanın başkalarının eşit haklarına saygı duyduğu sürece dilediği gibi kullanabileceği kaynakları olması yönündeyse, o zaman CEV dinamiği arka planda tıpkı bir doğa yasası gibi işleyerek suiistimalleri, hırsızlığı, gaspları ve diğer suçları engelleyip bunun gerçekleşmesini sağlayabilir.[16]

Dolayısıyla CEV yaklaşımının yapısı neredeyse sınırsız sayıda sonuca izin verir. Ayrıca insanlığın dışkestirilmiş iradesinin CEV'den hiçbir şeye karışmamasını isteyebileceği de tasavvur edilebilir. Bu durumda, CEV'i uygulayan yapay zekâ insanlığın dışkestirilmiş iradesinin yapmasını istediği şeyi yeterli olasılıkla yerleştirdikten sonra, güvenli bir şekilde kendisini kapatmalıdır.

Ek görüşler

Yukarıda anahatlarıyla anlatıldığı şekliyle, CEV önerisi elbette tamamen şematiktir. Önerinin farklı versiyonlarını sunan, çeşitli şekillerde saptanabilecek birtakım serbest parametrelere sahiptir.

Parametrelerden biri dışkestirim tabanıdır: Kimlerin iradesi dahil edilmelidir? "Herkes" diyebiliriz", ama bu cevap başka birtakım sorular doğuracaktır. Dışkestirim tabanına embriyolar, fetuslar, beyin ölümü gerçekleşmiş insanlar, şiddetli demansı olan ya da kalıcı bitkisel hayatta olan insanlar dahil midir? "Bölünmüş beyni" olan bir

hastanın beyninin her iki yarıküresi de dışkestirimde kendi ağırlığına sahip olur mu ve bu ağırlık normal bir öznenin tüm beyniyle aynı ağırlıkta mıdır? Peki ya geçmişte yaşamış, ama şimdi ölü olan insanlar? Ya gelecekte doğacak olan insanlar? Üst seviye hayvanlar ve diğer duygu sahibi canlılar? Dijital zihinler? Dünya dışı varlıklar?

Seçeneklerden biri sadece yapay zekânın yaratıldığı sırada hayatta olan yeryüzündeki yetişkin insanların dahil edilmesidir. Dolayısıyla bu tabandan ilk dışkestirim, tabanın genişleyip genişlemeyeceğini ve genişleyecekse nasıl genişleyeceğini kararlaştırabilir. Bu tabanın çeperindeki "marjinaller"in sayısı görece az olduğundan, dışkestirimin sonucu sınırın tam olarak nereye çekildiğine, sözgelimi fetusları kapsayıp kapsamadığına bağlı olmayabilir.

Özgün dışkestirim tabanından birisinin dışlanmış olması, isteklerinin ve esenliklerinin önemsenmediği anlamına gelmez. Eğer dışkestirim tabanındakilerin (örneğin yaşayan yetişkin insanların) bütünlüklü dışkestirilmiş iradesi ahlaki mülahazanın diğer varlıkları kapsayacak şekilde genişletilmesini isterse, o zaman CEV dinamiğinin sonucu bu tercihi yansıtacaktır. Gelgelelim özgün dışkestirim tabanında dahil edilenlerin çıkarlarının dışarlıklıların çıkarlarından çok daha fazla dikkate alınması mümkündür. Daha özelde, eğer dinamik sadece bireysel dışkestirilmiş iradeler arasında geniş bir mutabakat olduğunda işliyorsa (Yudkowsky'nin özgün önerisindeki gibi), sözgelimi insan olmayan hayvanların veya dijital zihinlerin refahının korunmasını engelleyebilecek sıkı bir blok ortaya çıkması riski gayet yüksektir. Sonuç potansiyel açıdan ahlaki çürümüşlük olabilir.[17]

CEV önerisinin saiklerinden biri, insanların ilk süper zeki yapay zekâyı yaratma yarışına girmeleri için bir gerekçe yaratmaktan sakınmaktı. CEV önerisi bu konuda birçok alternatiften daha iyi olsa da, çatışma güdülerini tamamen ortadan kaldırmaz. Bencil bir birey, grup ya da ulus gelecekten yana payını başkalarını dışkestirim tabanının dışında tutarak genişletmeye çalışabilir.

Bu türde bir iktidar gaspı çeşitli şekillerde meşrulaştırılabilir. Sözgelimi yapay zekâ gelişiminin maddi sponsoru olan kişinin, ortaya çıkan pastayı yemeyi hak ettiği savunulabilir. Bu ahlaki iddia sanırım yanlıştır. Örneğin ilk başarılı tohum yapay zekâyı başlatan projenin insanlığın geleceğine devasa bir risk dışsallığı dayattığı, dolayısıyla da bunu tazmin etmesi gerektiği itirazında bulunulabilir. Tazminat miktarı o kadar büyüktür ki, eğer işler iyi giderse, herkese iyi bir pay verme biçimini alabilir.[18]

Bir iktidar gaspını meşrulaştırmak için kullanılabilecek bir diğer argüman: İnsanlığın büyük bir kesiminin rezil ya da kötü tercihleri vardır ve onları dışkestirim tabanına dahil etmek insanlığın geleceğini bir distopyaya çevirecektir. Ortalama insanın kalbinde iyi ile kötünün payını bilmek zordur. Ayrıca bu dengenin farklı gruplar, toplumsal katmanlar, kültürler ya da uluslar arasında ne kadar değişiklik gösterdiğini

bilmek de zordur. İnsan doğası hakkında iyimser mi yoksa kötümser mi olunduğundan bağımsız olarak, insanlığın kozmik birikimini şu an hayatta olan yedi milyar insanın yeterli bir çoğunluğu için dışkestirilmiş iradelerinde iyi meleklerin galebe çalacağı spekülasyonuna bağlamamak tercih edilebilir. Elbette, dışkestirim tabanından belli sayıda insanı atmak, ışığın galip geleceğini garanti etmez ve en kısa zamanda başkalarını dışlayacak ya da iktidarı bizzat gasp edecek olan kişilerin kalbinde muazzam bir karanlık pekâlâ olabilir.

Fakat ilk dinamik konusundaki kavganın bir diğer nedeni: Yapay zekâ insanlığın CEV'ini uygulamanın bir yolu olarak ilan edilmiş olsa bile, başka birinin yapay zekâsının iddia edildiği gibi işlemeyeceğine inanıyor olabilirsiniz. Eğer farklı gruplar hangi uygulamanın başarıya ulaşma ihtimalinin daha fazla olduğu konusunda farklı inançlara sahipse, diğerlerinin başı çekmesini engellemek için kavga verebilirler. Bu tür durumlarda rakip projelerin epistemik farklılıklarını silahlı çatışmadansa kimin haklı olduğunu daha güvenilir şekilde belirleyecek bir yöntemle halletmeleri daha iyi olacaktır.[19]

Ahlak modelleri

CEV önerisi dolaylı normatifliğin mümkün olan tek biçimi değildir. Örneğin insanlığın bütünlüklü dışkestirilmiş iradesini hayata geçirmek yerine, ahlaken doğru olanı yapma hedefiyle hangi eylemlerin bu tanıma uyduğunu saptamak için yapay zekânın üstün bilişsel kapasitelerine yaslanarak bir yapay zekâ yaratmaya çalışılabilir. Bu öneriyi "ahlaki doğruluk" diye adlandırabiliriz. Buradaki fikir şudur: Biz insanlar neyin doğru neyin yanlış olduğuna dair kusurlu bir anlayışa ve ahlaki doğruluk kavramının felsefi olarak nasıl analiz edileceğine dair belki daha da kötü bir anlayışa sahip olabiliriz; ama bir süper zekâ bu şeyleri daha iyi anlayabilir.[20]

Peki ya ahlaki gerçekçiliğin doğru olup olmadığından emin değilsek? Yine de ahlaki doğruluk önerisini deneyebiliriz. Ahlaki gerçekçilik önvarsayımının yanlış olması durumunda yapay zekânın ne yapması gerektiğini saptadığımızdan emin olmamız gerekir. Örneğin eğer yapay zekâ ahlaki doğruluk konusunda hiçbir uygun ilintisiz hakikat olmadığını yeterli bir olasılıkla tahmin ederse, o zaman bütünlüklü dışkestirilmiş iradeyi uygulamaya geri dönmesini ya da kendi kendini kapatmasını şart koşabiliriz.[21]

Ahlaki doğruluğun CEV karşısında birkaç üstünlüğü olduğu göze çarpmaktadır. Ahlaki doğruluk CEV'deki çeşitli serbest parametreleri (yapay zekânın sonuca etki etmesi için gerekli dışkestirilmiş iradeler arasındaki bütünlük derecesi, bir çoğunluğun muhalif azınlıkları kolaylıkla bastırması ve dışkestirilmiş benliklerimizin "birlikte daha fazla büyüdüğü"nün varsayıldığı toplumsal çevrenin doğası gibi) tasfiye edecektir. Çok dar ya da çok geniş bir dışkestirim tabanından yararlanılmasından kaynaklı bir ahlaki başarısızlık olanağını ortadan kaldırıyor görünecektir. Ayrıca bütünlüklü dışkestirilmiş

iradelerimiz yapay zekâdan ahlaken çirkin eylemlerde bulunmasını istese bile, ahlaki doğruluk yapay zekâyı ahlaken doğru eyleme yönlendirecektir. Daha önce belirtildiği üzere, bu durum CEV'de canlı bir olanak gibi görünmektedir. Ahlaki iyilik insan doğasındaki bolca bulunan bir elementten ziyade değerli bir metale benziyor olabilir ve bu maden CEV'in reçetesine göre işlenip inceltildikten sonra dahi, asli sonucun parıldayan bir erdem mi umursamaz bir pislik mi yoksa zehirli bir atık mı olacağını kim bilebilir?

Ahlaki doğruluğun bazı olumsuz yanları olduğu da göze çarpacaktır. Zorluğuyla dillere destan bir kavram olan "ahlaken doğru" fikrine yaslanmaktadır; antikçağdan bu yana filozoflar bu kavramın peşine düşmüş, ama nasıl analiz edileceğine dair henüz bir mutabakata varılamamıştır. "Ahlaki doğruluk"un hatalı bir açımlamasını seçmek, ahlaken çok yanlış sonuçlar doğurabilir. "Ahlaki doğruluk"un tanımlanmasındaki bu zorluk, ahlaki doğruluk önerisinin aleyhinde ciddi bir etken olarak görülebilir. Gelgelelim ahlaki doğruluk önerisinin bu açıdan gerçekten sorunlu olduğu söylenemez. CEV önerisi de açımlanması zor kelimelerden ve kavramlardan yararlanır: "Bilgi", "daha fazla insan bizim istediğimiz kişi olsaydı", "daha fazla birlikte büyümek" gibi.[22] Bu kavramlar "ahlaki doğruluk" kadar muğlak olmasa da, programlayıcıların şu an kodla ifade edebileceği bir şey olmaktan kilometrelerce uzaktır.[23] Yapay zekâya bu kavramlardan herhangi birini bahşetmeye giden yol ona genel (en azından normal bir yetişkin insanınkine benzer) bir dil yeteneği vermeyi içerebilir. Böylece doğal dili anlamak konusunda bu tür bir genel yetenek, "ahlaken doğru" derken ne kastedildiğini anlamak için kullanılabilir. Eğer yapay zekâ anlamı kapabilirse, uygun olan eylemleri arayabilir. Yapay zekâ süper zekâ geliştirdikçe, iki cephede ilerleme kaydedebilir: Felsefi bir problem olarak ahlaki doğruluğun ne olduğunu anlamak ve pratik bir problem olarak bu anlayışı tikel eylemleri değerlendirmek için uygulamak.[24] Bu kolay olmasa da, insanlığın bütünlüklü dışkestirilmiş iradesini çıkarsamaktan *daha* zor olacağı da söylenemez.[25]

Ahlaki doğruluktaki bir diğer temel sorunsa şudur: Hayata geçirilebilse bile, daha parlak ve bilgili olsaydık isteyeceğimiz ya da tercih edeceğimiz şeyi vermeyebilir. Bu elbette tesadüf bir arıza değil, ahlaki doğruluğun asli özelliğidir. Ne var ki bize son derece zarar verebilecek bir özellik olabilir.[26]

Ahlaki doğruluğun temel fikrini korumaya çalışırken *ahlaken kabul edilebilirliğe* odaklanarak talepkârlığı azaltılabilir: Buradaki fikir, yapay zekânın ahlaken kabul edilemez şekillerde hareket etmediği sürece insanlığın CEV'inin peşinden gidebileceğidir. Örneğin yapay zekâ için şu hedef formüle edilebilir:

Yapay zekâ için ahlaken kabul edilebilir eylemler arasında, insanlığın CEV'in tercih edeceği bir eylemi alın. Gelgelelim eğer bu komutun bir parçasının somut olarak saptanmış bir anlamı yoksa ya da anlamı konusunda kafamız karman çormansa yahut ahlaki gerçekçilik yanlışsa ya da bu hedefle bir yapay zekâ yaratmakla ahlaken

kabul edilemez şekilde davrandıysak, o zaman kontrollü bir kapama gerçekleştirin.[27] Bu komutun kastedilen anlamını takip edin.

Yine de bu ahlaken kabul edilebilirlik modelinin, ahlaki gerekliliklere sindirilmesi zor derecede saygı duyduğundan endişe edilebilir. Bunun ne kadar büyük bir fedakârlık gerektireceği hangi etik teorisinin doğru olduğuna bağlıdır.[28] Eğer etik birkaç temel ahlaki kısıtlamaya uyan her eylemi ahlaken kabul edilebilir sayması anlamında *tatmin ediciyse*, o zaman ahlaken kabul edilebilirlik bizim bütünlüklü dışkestirilmiş irademizin yapay zekânın eylemlerini etkilemesi için epey bir alan bırakabilir. Gelgelelim eğer etik *azamileştiriciyse* –örneğin ahlaken kabul edilebilir yegâne eylemler ahlaken en iyi sonuçlara sahip olanlarsa– o zaman ahlaken kabul edilebilirlik bizim tercihlerimizin sonucu şekillendirmesine çok az alan bırakabilir, belki de hiç bırakmaz.

Bu kaygıyı örneklendirmek adına, hedonist sonuççuluk örneğine geri dönelim. Bu etik teorisinin doğru olduğunu ve yapay zekânın da bunu bildiğini varsayalım. Mevcut amaçlarımız açısından hedonist sonuççuluğu bir eylemin ancak ve ancak tüm gerçekleştirilebilir eylemler arasında başka hiçbir eylemin haz-ıstırap dengesi daha fazla değilse ahlaken doğru (ve ahlaken kabul edilebilir) olduğu iddiası olarak tanımlayabiliriz. Ahlaki kabul edilebilirliğin izinden giden yapay zekâ erişilebilir evreni hedonyuma dönüştürerek haz fazlalığını azamiye çıkarabilir; bu süreç komputronyum yaratıp bunu haz veren deneyimleri somutlaştıran hesaplamalar gerçekleştirmek için kullanmayı içerebilir. Mevcut bir insan beyninin simülasyonu haz üretmenin en etkili yolu olmadığından, muhtemel sonuçlardan biri hepimizin ölecek olmasıdır.

Dolayısıyla ahlaki doğruluk ya da ahlaki kabul edilebilirlik önerisinden birini hayata geçirmekle hayatlarımızı daha büyük bir iyilik adına feda etme riskini göze alırız. Bu durum düşündüğümüzden daha büyük bir fedakârlık olacaktır, çünkü kaybedeceklerimiz sadece normal bir insan yaşamı sürme şansı değil, dostane bir süper zekânın bahşedebileceği çok daha uzun ve zengin yaşamların sefasını sürme fırsatı da olacaktır.

Süper zekânın neredeyse bir o kadar büyük bir iyiliği potansiyel esenliğimizin çok daha azını feda ederken gerçekleştirebileceğini düşündüğümüzde, bu fedakârlığın çekiciliği azalabilir. *Neredeyse* tüm erişilebilir evrenin (kendi ihtiyaçlarımızı karşılaması için küçük bir kısım, örneğin Samanyolu hariç) hedonyuma dönüştürülmesine izin vermekte anlaştığımızı varsayalım. O zaman bile, hazzın azamileştirilmesine ayrılmış yüz milyar galaksi olacaktır. Ama milyarlarca yıl varlığını sürdürebilecek ve insanlar ile insan olmayan hayvanların yaşayıp gelişebilecekleri ve şad olmuş insan sonrası tinler haline gelme fırsatına erişecekleri tek bir galaksimiz olacaktır.[29]

Eğer bu son seçeneği tercih edersek (ben bu eğilimdeyim), ahlaken kabul edilebilir şekilde hareket etmek için koşulsuz bir sözcüksel olarak baskın tercihe sahip olmadığımızı ima eder. Ama ahlaka büyük önem verme konusunda tutarlıdır.

Salt ahlaki açıdan bakıldığında bile, ahlaki doğruluk ya da ahlaki kabul edilebilirlik kadar ahlaken hırslı olmayan bir öneriyi *savunmak* daha iyi olabilir. Eğer ahlaken en iyinin hayata geçirilme şansı –belki de sinir bozucu talepkârlığından ötürü– yoksa, başka bir öneriyi (ideale yakın olan ve hayata geçirilme ihtimali bizim bunu öne çıkarmamızla ciddi derecede artacak olan bir öneriyi) öne çıkarmak ahlaken daha iyi olabilir.[30]

Ne kastediyorsam onu yap!

Bütünlüklü dışkestirilmiş irade mi, ahlaki doğruluk mu, ahlaki kabul edilebilirlik mi, yoksa başka bir şey mi istediğimizden emin olmayabiliriz. Bu yüksek düzey kararda da kumar oynayıp yapay zekâya daha da fazla bilişsel çalışma yükleyebilir miyiz? Tembellik yapma ihtimalimizin sınırı nedir?

Örneğin şu "gerekçe temelli" hedefi düşünün:

Yapay zekâ, yapması için en fazla ne için gerekçe sunabilirsek onu yapsın.

Bu hedef dışkestirilmiş iradeye ya da ahlaka veya başka bir şeye varabilir, ama bizi kendimiz için bu daha özgül hedeflerden hangisini seçmek için en fazla gerekçemiz olacağını belirlemeye çalışmanın içerdiği çaba ve riskten kurtarıyor görünecektir.

Gelgelelim ahlak temelli hedeflerdeki problemlerden bazıları burada da geçerlidir. Birincisi, bu gerekçe temelli hedefin kendi arzularımıza çok az alan bırakacağından korkabiliriz. Bazı filozoflar bir insanın kendisi için ahlaken doğru olanı yapmak adına her zaman en fazla gerekçeye sahip olduğunu savunurlar. Eğer bu filozoflar haklıysa, o zaman gerekçe temelli hedef ahlaki doğruluk derekesine iner; bu durum bu tür bir dinamiği hayata geçiren bir süper zekânın yanında yöresinde olan herkesi öldürmesi riskini taşır. İkincisi, teknik dille ifade edilen tüm önerilerde olduğu gibi, kendi iddialarımızın anlamını yanlış anlamış olma ihtimalimiz vardır. Ahlak temelli hedefler örneğinde, yapay zekâdan doğru olanı yapmasını istemenin, öngörülmedik ve istenmeyen sonuçlara yol açabileceğini görmüştük (zaten öngörseydik bunlar söz konusu hedeften uzak durmamıza yol açardı). Aynısı yapay zekâdan yapmamız için en fazla gerekçemiz olan şeyi yapmasını isterken de geçerlidir.

Peki, ya bir hedefi alenen teknik olmayan bir dille (örneğin "hoşluk" terimleri üzerinden) ifade ederek bu zorluklardan sakınmaya çalışırsak?[31]

En hoş eylemi alın ya da eğer hiçbir eylem en hoş değilse, o zaman en azından on numara beş yıldız düzeyinde hoş olan bir eylem alın.

Hoş bir yapay zekâ yaratmaya nasıl bir itiraz yöneltilebilir ki? Ama bu ifadeyle tam olarak ne kastedildiğini sormamız gerekiyor. Sözlükte "hoş"un burada kullanılması amaçlanmayan çeşitli anlamları vardır: Yapay zekânın *nazik* ve *kibar* olmasını da, *aşırı hassas* ya da *müşkülpesent* olmasını da amaçlamayız. Eğer yapay zekânın "hoşluk"un

niyetlenen anlamını tanımasına ve hoşluğun peşine tam da bu anlamda düşmek üzere güdülenmesine güvenebilirsek, o zaman bu hedef programlayıcıların yapay zekânın ne yapmasını kastettiğini yapması için bir komuta varıyor görünecektir.[32] Benzer içerikte bir komut da CEV formülünde ("...yorumlanmasını istediğimiz şekilde yorumlanması") ve daha öncesinde ifade edildiği şekliyle, ahlaken kabul edilebilirlik kıstasında ("bu komutun niyetlenen anlamını takip edin") vardı. Bu tür bir "Ne kastediyorsam onu yap!" ibaresi ekleyerek hedef tanımındaki diğer kelimelerin düz anlamıyla değil de müşfik bir şekilde yorumlanması gerektiğine işaret ediyor olabiliriz. Fakat yapay zekânın "hoş" olması gerektiğini söylemek neredeyse hiçbir şey eklemez: Gerçek iş "Ne kastediyorsam onu yap!" komutuyla yapılmaktadır. Eğer "Ne Kastediyorsam Onu Yap"ın genel ve güçlü bir tarzda nasıl kodlanacağını bilirsek, ondan da bağımsız bir hedef olarak yararlanabiliriz.

"Ne kastediyorsam onu yap!" dinamiği nasıl hayata geçirilebilir. Yani isteklerimizi ve dile getirilmemiş niyetlerimizi müşfik şekilde yorumlayıp buna uygun olarak hareket edecek tarzda güdülenmiş bir yapay zekâyı nasıl yaratabiliriz? İlk adımlardan biri "Ne kastediyorsam onu yap!" derken ne kastettiğimiz konusunda daha faza netleşmeye çalışmak olabilir. Bunu daha davranışsalcı terimlerle, örneğin çeşitli farazi durumlarda (seçenekleri tartmak için daha fazla zamanımız olan, daha akıllı olduğumuz, ilintili gerçekler hakkında daha fazla bilgi sahibi olduğumuz ve çeşitli farklı tarzlarda koşulların dostane, hayırsever, hoş olan bir yapay zekâ istediğimizi söylerken ne kastettiğimize dair somut tercihleri doğru bir şekilde sergilememizin bizim için daha elverişli olduğu durumlarda) ortaya konmuş tercihler üzerinden açıklamaya çalışırsak yararlı olabilir.

Burada elbette dönüp dolaşıp aynı noktaya geliyoruz. İlk başladığımız dolaylı normatiflik yaklaşımına geri döndük. Özünde değer saptamadan tüm somut içeriği çıkartan, sadece salt prosedür babında tanımlanmış bir soyut değer bırakan CEV önerisi: Yapay zekânın uygun derecede idealleştirilmiş koşullarda yapmasını isteyeceğimiz şeyi yapması. Bu tür bir dolaylı normatiflik aracılığıyla, yapay zekânın hangi değerlerin peşine düşmesi gerektiğine dair daha somut bir tarifte bulunmayı denesek yerine getirmeye çalışacağımız bilişsel çalışmanın çoğunu yapay zekâya aktarmayı umabiliriz. Dolayısıyla yapay zekânın epistemik üstünlüğünden tamamen istifade etmeye çalışırken CEV epistemik saygı ilkesinin bir uygulaması olarak görülebilir.

Bileşen listesi

Buraya kadar, hedef sistemine hangi içeriklerin koyulacağına dair farklı seçenekleri ele aldık. Ama bir yapay zekânın davranışı başka tasarım tercihlerinden de etkilenecektir. Özelde, hangi karar teorisinden ve hangi epistemolojiden yararlandığı hayati bir fark yaratabilir. Bir diğer önemli sorun ise yapay zekânın planlarının eyleme dökülmeden önce insanlar tarafından değerlendirilmeye tabi tutulup tutulmayacağıdır.

TABLO 13 bu tasarım tercihlerini özetliyor. Bir süper zekâ yaratmayı amaçlayan bir proje bu bileşenlerden her biri hakkında hangi tercihlerde bulunduğunu açıklayabiliyor ve neden bu tercihlerde bulunduğunu gerekçelendirebiliyor olmalıdır.[33]

TABLO 13 Bileşen listesi

Hedef içeriği	Yapay zekâ hangi ereğin peşine düşmelidir? Bu ereğin bir tarifi nasıl yorumlanmalıdır? Bu erek, projenin başarısına katkıda bulunanlara özel ödüller vermeyi içermeli midir?
Karar teorisi	Yapay zekâ nedensel karar teorisinden mi kanıtsal karar teorisinden mi, güncellenmeyen karar teorisinden mi yoksa başka bir şeyden mi yararlanmalıdır?
Epistemoloji	Yapay zekânın ilk olasılık işlevi ne olmalıdır ve dünya hakkında aleni ya da zımni başka hangi varsayımlarda bulunmalıdır? Hangi antropik teorisinden yararlanmalıdır?
Onaylama	Yapay zekânın planları hayata geçirilmeden önce insanlar tarafından değerlendirilmeye tabi tutulmalı mıdır? Eğer öyleyse, bu değerlendirme sürecinin kuralları nelerdir?

Hedef içeriği

Dolaylı normatifliğin yapay zekânın peşine düşeceği değerlerin saptanmasında nasıl kullanılabileceğini yukarıda ele almıştık. Ahlak temelli modeller ve bütünlüklü dışkestirilmiş irade gibi bazı seçenekleri değerlendirdik. Bu tür seçeneklerden her biri başka tercihlerde bulunulmasını da gerektirir. Sözgelimi CEV yaklaşımı, dışkestirim tabanında kimin olduğuna, dışkestirimin yapısına vb. bağlı olarak birçok çeşide sahiptir. Güdülenme seçilimi yöntemlerinin diğer biçimleri farklı hedef içerik türleri gerektirebilir. Örneğin doğru cevaplar verilmesine bir değer atfeden bir kâhin yaratılabilir. Evcimenlik güdülenmesiyle yaratılmış bir kâhin, cevapların üretilmesinde kaynaklardan aşırı derecede yararlanılmasını değersiz gören bir hedef içeriğine de sahip olabilir.

Bir diğer tasarım tercihiyse yapay zekânın başarıyla gerçeğe dökülmesine katkıda bulunan bireyleri ödüllendirmek için (örneğin yapay zekânın davranışı için fazladan kaynaklar ya da nüfuz vererek) hedef içeriğine özel hükümler koyulup koyulmayacağıdır. Bu tür hükümlerin hepsini "özendirici ambalajlama" diye adlandırabiliriz. Özendirici ambalajlama projenin gerçekleştirmeyi önüne koyduğu hedefe bir ölçüde zarar verme pahasına projenin başarılı olma ihtimalini artırmanın bir yolu olarak görülebilir.

Örneğin eğer projenin hedefi insanlığın bütünlüklü dışkestirilmiş iradesini hayata geçiren bir dinamik yaratmaksa, o zaman özendirici ambalajlama planı belli bireylerin iradelerine dışkestirimde fazladan ağırlık verilmesi gerektiğini saptayabilir. Eğer bu tür bir proje başarılıysa, sonuç her zaman insanlığın bütünlüklü dışkestirilmiş iradesinin

uygulanması olacaktır diye bir kaide yoktur. Bunun yerine, bu hedefe belli oranda yaklaşılabilir.[34]

Özendirici ambalajlama bir süperzekâ tarafından yorumlanıp peşine düşülecek bir hedef içeriği olduğundan, bir insan yöneticinin hayata geçirmekte zorlanacağı incelikli ve karmaşık hükümleri saptamak için dolaylı normatiflikten istifade edebilir. Örneğin özendirici ambalajlama sistemi programlayıcıları kaba ama kolaylıkla erişilebilir bir ölçüye göre (örneğin kaç saat çalıştıkları ya da kaç tane bug düzelttiklerine göre) ödüllendirmek yerine, "katkılarının projenin sponsorlarının niyetlendiği tarzda başarıyla tamamlanmasına ilişkin uygulama öncesi makul bir olasılığını ne kadar artırdığıyla orantılı olarak ödüllendirilmesi gerektiği"ni belirleyebilir. Dahası özendirici ambalajlamanın planlarında sınırlandırılması için de bir neden olmayacaktır. *Her* insan tam da isteğine göre ödüllendirilmesi gerektiğini belirleyebilir. Kredi tahsisi zor bir problemdir, ama süper zekânın özendirici ambalajlamayla alenen ya da zımnen belirlenmiş kıstaslara yaklaşmak konusunda makul bir iş çıkarması beklenebilir.

Süper zekânın kendisinin yaratılmasından önce ölmüş olan bireyleri ödüllendirmenin bir yolunu bile bulabilmesi tasavvur edilebilir.[35] Dolayısıyla özendirici ambalajlama en azından ölülerin (potansiyel açıdan proje tasarlanmadan önce, hatta özendirici ambalajlama projesinin ilk dile getirilmesinden önce ölmüş olan bireylerin) bir kısmını kucaklayacak şekilde genişletilebilir. Bu tür geriye dönüşlü bir proje bu kelimeler sayfaya dökülürken mezarlarında son istirahatlerini yapmakta olan insanları özendirmeyecek olsa da, ahlaki nedenlerle tercih edilebilir, ama adil olmak bir hedef olduğu ölçüde, çevresindeki özendirici ambalajlamadan ziyade hedef saptamanın parçası olarak dahil edilmesi gerektiği savunulabilir.

Burada özendirici ambalajlamayla bağlantılı etik ve stratejik meselelerin tamamına dalmamız mümkün değil. Gelgelelim bir projenin bu meselelerdeki tutumu temel tasarım kavramının önemli bir veçhesi olacaktır.

Karar teorisi

Bir diğer önemli tasarım tercihiyse yapay zekânın hangi karar teorisinden yararlanacak şekilde yaratılacağıdır. Bu karar yapay zekânın bazı stratejik açıdan hayati durumlarda nasıl davranacağını etkileyebilir. Sözgelimi yapay zekânın varoluşunu varsaydığı diğer süper zeki medeniyetlerle alışveriş yapıp yapmayacağını ya da onları dolandırıp dolandırmayacağını belirleyebilir. Karar teorisinin tikel ayrıntıları sonsuz kazançların sonlu olasılıklarını ("Pascalcı bahisler") ya da son derece büyük sınırlı kazançların son derece küçük olasılıklarını ("Pascalcı gasplar") içeren durumlarda veya yapay zekânın temel normatif belirsizlikle yüzleştiği ya da aynı aracı programının çoklu örneklemelerinin olduğu bağlamlarda da önem arz edebilir.[36]

Masadaki seçenekler arasında nedensel karar teorisi (çeşitli biçimleriyle) ve kanıtsal karar teorisinin yanı sıra "zamanüstü karar teorisi" ve "güncellenmeyen karar teorisi" gibi hâlâ geliştirilme aşamasında olan yeni adaylar vardır.[37] Doğru karar teorisini saptayıp ifade etmek ve ona duyduğumuz güvenin meşrulaştırılması zor olabilir. Bir yapay zekânın karar teorisini doğrudan saptama olasılığı belki de nihai değerlerini doğrudan saptama olasılığından daha fazladır, ama yine de ciddi bir hata riskiyle karşı karşıyayız. Şu an en popüler olan karar teorilerini bozabilecek karmaşıklıklardan birçoğu sadece yakın zaman önce keşfedilmiş olup, henüz gün ışığına çıkmamış başka problemler de olabileceğini akla getirmektedir. Yapay zekâya kusurlu bir karar teorisi vermenin sonucu korkunç olabilir, muhtemelen varoluşsal bir felakete yol açar.

Bu zorluklar dikkate alındığında, yapay zekânın kullanması gereken karar teorisinin saptanmasına dolaylı bir yaklaşımı düşünmek gerekebilir. Bunun tam olarak nasıl yapılacağıysa belirsizdir. Yapay zekânın "mesele hakkında uzun uzadıya ve sıkıca düşünseydik yararlanmak isteyeceğimiz şu karar teorisi *K*'den" yararlanmasını isteyebiliriz. Gelgelelim yapay zekâ, *K*'nin ne olduğunu öğrenmeden önce kararlar alabiliyor olmalıdır. Dolayısıyla *K* arayışını yönetecek olan etkili bir geçici karar teorisine (*K'*) ihtiyacı olacaktır. *K'* teorisini *K* hakkında (olasılıklarına göre) yapay zekânın mevcut hipotezlerinin bir tür üstdüşümü olacak şekilde tanımlamaya çalışabiliriz, ama bunu tamamen genel bir tarzda yapmak için çözülmemiş teknik problemler vardır.[38] Ayrıca yapay zekâ hangi tikel karar teorisinin doğru olduğunu belirleme fırsatına sahip olmadan önce öğrenme aşaması sırasında geri döndürülemez şekilde kötü kararlar alacağından (örneğin bundan böyle kusurlu bir karar teorisine bağlı kalacak şekilde kendini yeniden yazmasından) kaygı duymak için de bir neden vardır. Bu yaralanabilirlik aşamasında raydan çıkma riskini azaltmak için, tohum yapay zekâya bir tür *sınırlandırılmış aklilik* bahşetmeye çalışabilir: Ezoterik kaygıları, hatta nihayetinde meşru olduğunu düşünebileceğimiz kaygıları bile tamamen göz ardı eden ve belli koşullar karşılandığında daha çetrefilli bir (dolaylı olarak saptanmış) karar teorisiyle ikame edecek şekilde tasarlanmış olan kasten basitleştirilmiş ama işe yarayacağını umut edebileceğimiz bir karar teorisi.[39] Bunun işleyip işlemeyeceği ve nasıl işleyeceği tartışmaya açıktır.

Epistemoloji

Bir projenin yapay zekânın epistemolojisini seçerken ampirik hipotezleri değerlendirecek ilkeleri ve kıstasları saptayarak temel bir tasarım tercihi yapması da gerekecektir. Bayesçi bir çerçevede epistemolojiyi bir ilk olasılık işlevi olarak düşünebiliriz: Algısal bir kanıtı hesaba katmadan önce yapay zekânın olanaklı dünyalara zımnen olasılıklar atfetmesi. Diğer çerçevelerde epistemoloji farklı bir biçime bürünebilir; ama her halükârda, eğer yapay zekâ geçmiş gözlemlerden genellemeler yapıp geleceğe dair tahminlerde bulunacaksa tümevarımcı bir öğrenme kuralı zorunludur.[40] Gelgelelim

hedef içeriği ve karar teorisinde olduğu gibi, epistemoloji saptamamızın hedefi ıskalayabilmesi gibi bir risk vardır.

Yanlış şekilde saptanmış bir epistemolojiden ne kadar zarar doğabileceğinin bir sınırı olduğu düşünülebilir. Eğer epistemoloji *çok* işlevsizse, o zaman yapay zekâ çok zeki olamaz ve bu kitapta ele alınan türde bir risk teşkil edemez. Ama yapay zekâyı birçok durumda araçsal açıdan etkili hale getirecek kadar sağlıklı olan, ama can alıcı önem taşıyan bir meselede yoldan çıkaracak bir kusura sahip olan bir epistemoloji saptayabileceğimiz kaygısı vardır. Bu tür bir yapay zekâ, dünya görüşü yanlış bir dogmaya dayalı olan, mutlak bir inançla bağlı kalınan ve dolayısıyla "yel değirmenleriyle savaşan" ve neyi var neyi yoksa hepsini fantastik ya da zararlı meşgalelerin peşinde harcayan bir cin fikirli insana benzetilebilir.

Bir yapay zekânın ilkindeki belli türde ince farklılıklar nasıl davrandığı konusunda ciddi bir fark yaratabilir. Örneğin bir yapay zekâya evrenin sonsuz olmasına sıfır olasılık atfeden bir ilk verilebilir. Aksi istikamette ne kadar astronomik kanıt birikirse biriksin, bu tür bir yapay zekâ sonsuz bir evren ima eden her türlü kozmoloji teorisini inatla reddedecektir ve bunun sonucunda saçma salak tercihler yapabilir.[41] Veya bir yapay zekâya evrenin Turing-hesaplanabilir olmamasına bir sıfır olasılık atfeden bir ilk verilebilir (bu aslında yazında ele alınan ilklerden birçoğunun –1. Bölüm'de bahsi geçen Kolmogorov karmaşıklığı dahil– ortak bir özelliğidir) ve bir kez daha yerleşik varsayımın ("Church-Turing tezi" olarak bilinir) yanlış olduğu ortaya çıkarsa, sonuçları çok az anlaşılacaktır. Bir yapay zekâ güçlü bir zihin-beden ikiciliği biçiminin doğru olabilme imkânını ya da indirgenemez ahlaki olgular olma imkânını dışlayarak şu ya da bu türde güçlü metafizik bağlılıkları olan bir ilke sahip olabilir. Eğer bu bağlılıklardan herhangi biri hatalıysa, yapay zekâ nihai hedeflerini bizim tersten örneklemeler olarak göreceğimiz şekillerde gerçekleştirmeye çalışabilir. Ama bu tür bir yapay zekânın önemli bir meselede temelde yanılıyor olmasına karşın belirleyici bir stratejik üstünlük elde edecek kadar araçsal açıdan etkili olamaması için bariz bir neden yoktur. (Gözlem seçilim etkilerinin varlığında gösterici bilgiden nasıl çıkarımlarda bulunacağını inceleyen antropik çalışmalar, epistemik aksiyom tercihlerinin hayati olduğunun anlaşılabileceği bir diğer alandır.[42])

İlk tohum yapay zekânın inşasında epistemolojideki tüm temel sorunları zamanında çözebileceğimizden şüphe duymamız sebepsiz değildir. Dolayısıyla yapay zekânın epistemolojisini saptanmasına dolaylı bir yaklaşımı düşünebiliriz. Bu durum karar teorisinin saptanmasına dolaylı bir yaklaşıma sahipken doğan sorunlardan birçoğunu gündeme getirecektir. Gelgelelim epistemoloji örneğinde iyicil yakınsaklık umudu çok daha fazla olabilir, zira güvenli ve etkili yapay zekâ için yeterli temeli sağlayan ve nihayetinde benzer doksa sonuçları doğuran çok çeşitli epistemolojiler vardır; zira ye-

terince bol miktarda ampirik kanıtlar ve analiz, ilk beklentilerdeki tüm küçük farkları silme eğiliminde olacaktır.[43]

Yapay zekâya kendi düşüncemizi yöneten ilkelerle örtüşen temel epistemoloji ilkeleri bahşetmek iyi bir amaç olacaktır. Bu idealden uzaklaşan bir yapay zekâ, kendi standartlarımızı tutarlı şekilde uyguladığımız takdirde yanlış muhakeme yürüteceğimize hükmedeceğimiz bir yapay zekâdır. Elbette bu durum sadece bizim *temel* epistemolojik ilkelerimiz için geçerlidir. Tohum yapay zekâ dünyayla ilgili kavrayışını geliştirdikçe temel olmayan ilkeleri sürekli yaratıp gözden geçirmelidir. Yapay zekânın derdi insanların önyargılarına oynamak değil, cehaletimizden ve aptallığımızdan ötürü bizi cezalandırmaktır.

Onaylama

Tasarım tercihleri listemizdeki son madde *onaylamadır*. Yapay zekânın planları hayata geçirilmeden önce insanlar tarafından değerlendirilmeye tabi tutulmalı mıdır? Bir kâhin için bu soru zımnen evet diye cevaplanır. Kâhin bilgi üretir; insan değerlendirmeciler de buna göre hareket edip etmeyeceklerini ve edeceklerse nasıl edeceklerini seçerler. Cinler, hükümdarlar ve alet yapay zekâlar için ise bir onaylama biçiminden yararlanıp yararlanılmayacağı sorusu açıktır.

Onaylamanın nasıl işleyeceğine örnek olması adına yapay zekânın insanlığın CEV'ini hayata geçiren bir hükümdar olarak işlemesinin amaçlandığını düşünün. Bu yapay zekâyı doğrudan uygulamak yerine, sırf hükümdar yapay zekânın ne yapacağına dair sorular sormak amacıyla ilkin bir kâhin yapay zekâ inşa ettiğimizi tahayyül edin. Önceki bölümlerde ortaya konduğu üzere, bir süper zeki kâhin yaratmanın riskleri vardır (zihin suçu ya da altyapı bolluğu gibi). Ama bu örneğin amaçları açısından kâhin yapay zekânın bu tuzaklardan kaçınan tarzda başarıyla uygulandığını varsayalım.

Dolayısıyla insanlığın CEV'ini uygulaması amaçlanan bir kod parçasını işletmenin sonuçlarına dair bize en iyi tahminlerini sunan bir kâhin yapay zekâya sahibiz. Kâhin ne olacağını ayrıntısıyla tahmin edemeyebilir, ama tahminlerinin bizimkinden daha iyi olması muhtemeldir. (Eğer bir süper zekâ için bile kodun ne yapacağına dair *herhangi bir şey* tahmin etmek imkânsızsa, onu çalıştırmak için yanıp tutuşuruz.) Dolayısıyla kâhin bir süre düşünür ve ardından tahminini sunar. Kâhin cevabı anlaşılır kılmak için operatöre tahmin edilmiş sonucun çeşitli özelliklerini araştıracak çeşitli aletler önerebilir. Kâhin geleceğin neye benzeyeceğine dair resimler gösterip ortalama, zirve ve en düşük düzeyde refahla birlikte farklı zamanlarda var olacak duygu sahibi varlıklar hakkında istatistikler sunabilir. Rastgele seçilmiş birkaç bireyin (belki de temsili olması için seçilmiş hayali insanların) mahrem biyografilerini önerebilir. Gelecekle ilgili operatörün araştırmayı aklından bile geçiremeyebileceği ama bir kez işaret edildiğinde alakalı görülecek olan veçheleri aydınlatabilir.

Sonucun bu şekilde bir önizlemesini yapabilmenin bariz üstünlükleri vardır. Önizleme planlı bir hükümdarın tasarım saptamalarındaki ya da kaynak koddaki bir hatanın sonuçlarını ortaya serebilir. Eğer kristal küre berbat bir gelecek gösteriyorsa, planlı hükümdar yapay zekâ kodunu ıskartaya çıkarabilir ve başka bir şey deneyebiliriz. Bir seçeneğe bağlanmadan önce, özellikle de insanlığın geleceği söz konusu olduğunda, somut sonuçlarını bilmemiz gerektiği argümanı güçlü bir şekilde savunulabilir.

Ama bu kadar bariz olmayan bir şey varsa, o da onaylamanın da potansiyel açıdan kayda değer sonuçları olduğudur. Eğer rakip kesimler doğru çıkacaklarına duydukları güvenle üstün bilgeliğin hakemliğini kabul etmek yerine, hükmün ne olacağını önceden görebilirlerse, CEV'in uzlaştırıcı niteliği mahvolabilir. Ahlak temelli yaklaşımın savunucularından biri eğer ahlaken optimum olan için gerekli tüm fedakarlıklar açığa çıkarsa, sponsorun kararlılığının kırılacağından endişe edebilir. Ve kendimizi aşmak için bazı sürprizler, bazı uyumsuzluklar, bazı çılgınlıklar, bazı fırsatlar barındıran bir geleceği tercih etmek için hepimizin bir gerekçesi olabilir; bu geleceğin ana çizgileri peşin hükümler sunacak şekilde kesin çizilmemiştir, ama çarpıcı hareket ve plansız büyüme için bir esneklik sağlayabilir. Eğer geleceği en ince ayrıntısına kadar seçebilsek ve o an hayalimizdekine tamamen uymayan her taslağı geri göndersek, böyle kapsamlı bir görüşe varma ihtimalimiz daha az olabilir.

Bu nedenle sponsor onaylaması meselesi ilk başta göründüğü kadar net olmayabilir. Gelgelelim artısı eksisiyle değerlendirildiğinde, işlevsellik mevcutsa bir önizleme fırsatından istifade etmek basiretli bir tutum olarak görünecektir. Ama önizlemeyi yapanın sonucu enine boyuna değiştirmesine izin vermekten ziyade, tüm proje çöpe gitmeden önce ona sadece birkaç kez uygulanabilecek basit bir veto hakkı verebiliriz.[44]

Yeterince yakınlaşmak

Onaylamanın ana amacı feci bir hata yapılması ihtimalini azaltmaktır. Genel olarak, her ayrıntının tümüyle optimize edilmesi şansını azamileştirmekten ziyade feci hata riskini asgariye çekmeyi amaçlamak akıllıca görünmektedir. Bunun iki nedeni vardır. Birincisi, insanlığın kozmik birikimi muazzam büyüktür: Bizim sürecimiz bazı israfları içerse ya da bazı gereksiz kısıtlamaları kabul etse bile, yine de herkese yetecek kadarı vardır. İkincisi, zekâ patlaması için ilk koşulları kabaca doğru anlasak da, ortaya çıkan süper zekânın sonunda bizim nihai hedeflerimize yönelmesi ve tastamam isabet ettirmesi umudu vardır. Önemli olan, doğru çekim havzasına girmektir.

Epistemolojiye ilişkin olarak, çok sayıda öncül (bir süper zekâ tarafından hesaplanıp gerçekçi miktarda veriler koşul olarak konduğunda) nihayetinde çok benzer sonlara yaklaşabilir; bu gayet akla yatkındır. Dolayısıyla epistemolojiyi *tastamam* doğru tutturmak konusunda endişe etmemize gerek yoktur. Yapay zekâya çok sayıda deneyim

ve analizden yararlanıldığında bile hayati doğruları öğrenmekten aciz hale getirecek kadar aşırı bir öncül vermekten sakınmamız gerekir.[45]

Karar teorisine ilişkin olarak, geri döndürülemez hata riski büyük görünmektedir. Yine de yeterince iyi bir karar teorisini doğrudan saptamayı umut edebiliriz. Bir süper zeki yapay zekâ her an yeni bir karar teorisine geçebilir; gelgelelim eğer yeterince yanlış bir karar teorisiyle yola çıkarsa, değiştirme nedeni görmeyebilir. Bir aracı farklı bir karar teorisine sahip olmanın yararlarını görse bile, bu farkındalık çok geç olabilir. Örneğin şantajı reddetmek üzere tasarlanmış bir aracı, müstakbel şantajcıları caydırmanın sefasını sürebilir. Bu nedenle şantaj yapılabilir aracılar proaktif davranıp istismar edilemez bir karar teorisini benimsemekle çok iyi yapabilirler. Fakat şantaj yapılabilir bir aracı tehdidi alıp bunu inanılır gördüğünde, zarar verilmiştir bile.

Yeterli bir epitemoloji ve karar teorisi var olduğunda, sistemi CEV'i ya da dolaylı olarak saptanmış bir başka hedef içeriğini hayata geçirmek üzere tasarlamaya çalışabiliriz. Bir kez daha yakınsaklık umudu vardır: CEV türü bir dinamiği hayata geçirmenin farklı yolları aynı ütopyacı sonuca varacaktır. Bu tür bir yakınlaşma olmadığındaysa, yine de mümkün olan farklı sonuçlardan birçoğunun varoluşsal başarı sayılacak kadar iyi olduğunu umut etmeye devam edebiliriz.

Son derece optimumlaştırılmış bir tasarım yaratmamız zorunlu değildir. Daha ziyade, odak noktamız son derece güvenilir bir tasarım, kendi kusurlarını tanıyacak kadar aklına mukayyet olacağına güvenilebilecek bir tasarım yaratmak olmalıdır. Temelleri sağlam olan kusurlu bir süper zekâ kendi kendini tedricen tamir edecektir ve bunu yaptıktan sonra, sanki ilk baştan beri kusursuzmuş gibi dünyaya hayırhah optimumlaştırıcı güç uygulayacaktır.

Stratejik Resim

Artık süper zekânın yarattığı zorluğu daha geniş bir bağlamda ele almanın vaktidir. En azından hangi genel doğrultuda ilerlemekte olduğumuzu yeterince bilecek şekilde stratejik manzarada nerede olduğumuzu saptamak istiyoruz. Ama bunun kolay olmadığı hemen anlaşılıyor. Burada, sondan bir önceki bölümde, uzun vadeli bilim ve teknoloji politikaları meseleleri hakkında düşünmemize yardımcı olan bazı genel analitik kavramları tanıtacağız. Ardından bunları makine zekâ meselesine uygulayacağız.

Önerilen bir politikanın değerlendirilebileceği iki farklı normatif duruş arasında kaba bir ayrım yapmak aydınlatıcı olabilir. *İnsanı etkileyen perspektif,* önerilen bir değişikliğin "bizim çıkarımız"a olup olmayacağını, yani zaten var olan ya da önerilen değişikliğin gerçekleşip gerçekleşmeyeceğinden bağımsız olarak var olacak olan ahlaken kayda değer varlıkların çıkarına olup olmayacağını (artı eksiye vurulduğunda ve beklenti bakımından) sorar. *Gayri şahsi perspektif* ise bugün var olan insanları ya da önerilen değişikliğin gerçekleşip gerçekleşmeyeceğinden bağımsız olarak var olacak olan insanları özel olarak dikkate almaz. Bunun yerine, mevcut konumundan bağımsız olarak, herkesi eşit sayar. Gayri şahsi perspektif yaşamaya değer hayatları olduğu takdirde yeni insanlar var etmeyi çok değerli görür: Ne kadar çok mutlu hayat yaratılırsa, o kadar iyidir.

Bu ayrım bir makine zekâsı devrimiyle bağlantılı ahlaki karmaşıklıkları çok az dikkate alsa da, sıkı bir analizde yararlı olabilir. Burada meseleyi ilkin gayri şahsi perspektiften ele alacağız. Ardından kişiyi etkileyen hususlar dikkate alınsa hangi değişiklikler olacağını göreceğiz.

Bilim ve teknoloji stratejisi

Makine süper zekâsına özgü meselelere odaklanmadan önce, daha genel olarak bilimsel ve teknolojik gelişimle alakalı bazı stratejik kavramları ve hususları tanıtmalıyız.

Diferansiyel teknolojik gelişim

Bir politika üreticisinin, nihayetinde kendi toprağından doğabilecek bir farazi teknolojinin risklerine ya da uzun vadeli sonuçlarına duyduğu kaygıdan ötürü belli bir

araştırma alanındaki fonları kesmeyi önerdiğini varsayalım. O zaman araştırmacılardan epey bir karşı çıkış gelebilir.

Bilimciler ve onların kamuoyundaki savunucuları teknolojinin evrimini araştırmaların önünü tıkayarak kontrol etmeye çalışmanın beyhude bir çaba olduğunu sık sık söylerler. Eğer bir teknoloji gerçekleştirilebilirse (denmektedir), bir politika üreticisinin spekülatif riskler konusundaki çekinceleri ne olursa olsun yine de geliştirilecektir. Gerçekten de bir gelişim çizgisinin üretmeyi vaat ettiği yeterlilikler ne kadar güçlüyse, birilerinin bir yerlerde onun peşine düşeceğinden o kadar emin olabiliriz. Fonların kesilmesi ilerlemeyi durdurmayacak ya da beraberinde getirdiği tehlikeleri engellemeyecektir.

İşin ilginci, bu beyhudelik itirazı bir politika üreticisi bir araştırma alanındaki fonları *artırmayı* önerdiğinde neredeyse asla gündeme gelmez, oysa argümanın bir de bu boyutu vardır. Şu öfkeli seslerin yükseldiğini nadiren duyarız: "Lütfen bize ayrılan fonları artırmayın. Onun yerine kesinti yapın. Diğer ülkelerdeki araştırmacıların yarım bırakılan işi tamamlayacakları kesindir; aynı iş her halükârda yapılır. Kamu hazinesini ülke içindeki bilimsel araştırmalar için çarçur etmeyin!"

Bu zahiri çelişkinin sebebi ne olabilir? Elbette akla yatkın izahlardan biri araştırmacıların kendi çıkarlarına hizmet eden peşin hükümlere sahip olmaları ve bu temelde bizi araştırmanın her zaman iyi olduğuna ve daha fazla fona olan talebimizi destekleyen her argümanı desteklemek gerektiğine inandırmaya çalışmalarıdır. Gelgelelim çifte standardın ulusal çıkarlar üzerinden meşrulaştırılması da mümkündür. Bir teknolojiyi geliştirmenin *iki* sonucu olduğunu varsayalım: Küçük kazancı (K) yatırımcılarına ve sponsorları olan ülkeye vermek, ama herkese toplamda bundan daha fazla zarar (Z) dayatmak (bu bir risk dışsallığı olabilir). Büyük oranda diğerkâm biri bile genel zararlı teknolojiyi geliştirmeyi tercih edebilir. Zaten ne yaparsak yapalım zarar gelecek, diye akıl yürütebilirler, zira onlar uzak dursa bile bu teknolojiyi zaten başkası geliştirecektir ve genel refahın etkilenmeyeceği düşünüldüğünde, kazancı (K) kendilerine ve uluslarına mal edebilirler. ("Ne yazık ki, dünyayı yok edecek bir araç yakında bulunacaktır. Neyse ki elimizde bunu yaratacak kaynak var!")

Beyhudelik itirazının çekiciliğine dair açıklama ne olursa olsun, teknolojik gelişimi yönlendirmeye çalışmak için genel olarak hiçbir gayri şahsi gerekçe olmadığını göstermeyi başaramıyor. Bilimsel ve teknolojik gelişme çabalarıyla birlikte tüm ilintili teknolojilerin sonunda gelişeceğine dair motive edici fikri kabul etsek bile, yani şunu kabullensek bile, bu itiraz yine de çökmeye mahkûmdur:

> *Teknolojik tamamlanma kestirimi*
>
> Eğer bilimsel ve teknolojik gelişim çabaları etkili bir şekilde sonlanmazsa, o zaman olanaklı bir teknoloji yoluyla elde edilebilecek önem arz eden tüm temel yeterlilikler edinilecektir.[1]

Bunlar teknolojik tamamlanma kestiriminin neden beyhudelik itirazını zımnen içermediğine dair en azından iki nedendir. Birincisi, öncel tutmayabilir, zira bilimsel ve teknolojik gelişim çabalarının (teknolojik olgunluğa erişmeden önce) fiilen durmayacağı aslında bir ön kabul olarak alınamaz. Bu şerh varoluşsal riskin söz konusu olduğu bir bağlamda bilhassa geçerlidir. İkincisi, olanaklı teknolojilerden bazıları yoluyla elde edilebilecek tüm önemli temel yeterliliklerin edinileceğinden emin olabilsek bile, yine de teknolojik araştırmanın yönünü etkilemeye çalışmak anlam ifade edebilir. Önemli olan bir teknolojinin *geliştirilip geliştirilmediği* değil, aynı zamanda *kim tarafından, ne zaman* ve *hangi bağlamda* geliştirildiğidir. Yeni bir teknolojinin doğuşunun bu koşulları yaratacağı etkiyi şekillendirir ve fon musluklarının açılıp kısılmasından (ve diğer politika araçlarının nasıl kullanıldığından) etkilenebilir.

Bu mülahazalar farklı teknolojilerin geliştirildiği göreli hıza dikkat kesilmemizi sağlayacak bir ilkeyi akla getirir:[2]

> *Diferansiyal teknolojik gelişim ilkesi*
>
> Tehlikeli ve zararlı teknolojilerin, özellikle de varoluşsal risk düzeyini yükselten teknolojilerin gelişimini geciktirin ve yararlı teknolojilerin, özellikle de doğanın veya diğer teknolojilerin sunduğu varoluşsak riskleri azaltan teknolojilerin gelişimini hızlandırın.

Dolayısıyla bir politika bir teknolojik gelişimin istenmeyen biçimlerine kıyasla istenen biçimlerine ne kadar diferansiyel üstünlük kazandırdığı temelinde değerlendirilebilir.[3]

Tercih edilen varış düzeni

Bazı teknolojilerin varoluşsal risklere etkisi müphemdir ve bazı varoluşsal riskleri artırırken bazılarını azaltır. Süper zekâ böyle bir teknolojidir.

Önceki bölümlerde makine süper zekâsını devreye sokmanın ciddi bir varoluşsal risk doğuracağını görmüştük. Ama diğer birçok varoluşsal riski azaltacaktır. Doğadan gelecek riskler (asteroidlerin etkileri, süper yanardağlar ve doğal salgınlar) neredeyse tamamen ortadan kalkacaktır, zira süper zekâ bu tür zararların çoğuna karşı önlemler alabilir ya da bunları en azından varoluşu tehdit etmeyen riskler kategorisine sokabilir (sözgelimi uzayda yaşam yoluyla).

Doğadan gelen bu varoluşsal riskler ilintili zaman ölçeklerinde görece küçüktür. Ama süper zekâ insandan kaynaklı risklerden birçoğunu da ortadan kaldıracak ya da azaltacaktır. Daha özelde, kazaen yıkıma yol açılması riskini (yeni teknolojilerle alakalı kaza riskleri dahil) azaltacaktır. Bir süper zekâ insanlardan genel olarak daha ehil olduğundan; hata yapma ihtimali daha az, önlemlere ihtiyaç olduğunda tanıma ve önlemleri ehil şekilde hayata geçirme ihtimali daha fazla olacaktır. İyi inşa edilmiş bir süper zekâ bazen bir risk alabilir, ama ancak bunu yapmak akıllıca olduğunda. Ayrıca

en azından süper zekânın bir tekel oluşturduğu senaryolarda, küresel eşgüdüm problemlerinden doğan ve kaza kapsamına girmeyen insandan kaynaklı birçok varoluşsal risk ortadan kalkacaktır. Bunlar arasında savaş, teknoloji yarışları, rekabet ve evrimin istenmeyen biçimleri ve müştereklerin trajedileri gibi riskler de vardır.

Ciddi tehlike sentetik biyoloji, moleküler nanoteknoloji, iklim mühendisliği, biyomedikal geliştirme ve nöropsikolojik manipülasyon aletleri, totaliterliği ya da despotluğu kolaylaştırabilecek toplumsal kontrol araçları ve henüz tahayyül edilmemiş diğer teknolojileri geliştiren insanlarla bağlantılı olacağından, bu tür riskleri ortadan kaldırmak çok büyük kazanç sağlayacaktır. Dolayısıyla süper zekâya daha erken tarihte ulaşılmasının tercih edilir olduğuna dair bir argüman savunulabilir. Gelgelelim eğer doğadan ve gelecekteki teknolojiyle bağlantısız diğer tehlikelerden kaynaklı riskler azsa, o zaman şu argüman inceltilebilir: Önemli olan başka tehlikeli teknolojilerden (ileri nanoteknoloji gibi) *önce* süper zekâya erişmemizdir. Bunun er mi geç mi olacağı (gayri şahsi bir perspektiften) o kadar da önemli olmayabilir, yeter ki varışın sırası doğru olsun.

Süper zekânın potansiyel açıdan tehlikeli diğer teknolojilerden (nanoteknoloji gibi) önce gelmesini tercih etmenin gerekçesi, süper zekânın nanoteknolojiden doğan varoluşsal riskleri azaltacak olmasıdır, ama tersi geçerli değildir.[4] Dolayısıyla ilkin süper zekâ yaratırsak, sadece süper zekâyla bağlantılı olan varoluşsal risklerle yüzleşiriz; halbuki eğer ilkin nanoteknoloji yaratırsak, nanoteknoloji riskleriyle ve ardından bir de süper zekâ riskleriyle karşılaşırız.[5] Dolayısıyla süper zekâdan kaynaklanan varoluşsal riskler çok büyük olsa bile ve dahası süper zekâ en riskli teknoloji olsa bile, varışını hızlandırmak gerektiği savunulabilir.

Gelgelelim bu "ne kadar erken o kadar iyi" argümanları süper zekâ yaratmanın riskliliğinin ne zaman yaratıldığından bağımsız olarak aynı olduğunu varsayar. Bunun yerine, eğer risklilik zamanla azalırsa, makine zekâsı devrimini ertelemek daha iyi olabilir. Daha geç ulaşılması diğer varoluşsal felaketlerin araya girmesi için daha fazla zaman bırakacak olsa da, süper zekânın gelişimini yavaşlatmak yerine tercih edilir bir seçenek olabilir. Eğer süper zekâyla bağlantılı varoluşsal riskler diğer yıkıcı teknolojilerle bağlantılı risklerden çok daha fazlaysa, bu bilhassa akla yatkın olacaktır.

Bir zekâ patlamasının riskliliğinin onyıllara yayılan bir zaman ölçeğinde gerileyeceğine inanmamız için birkaç güçlü neden var. Nedenlerden biri, ileriki tarihte ulaşılmasının kontrol problemine çözümler geliştirilmesi için daha fazla zaman bırakacak olmasıdır. Kontrol problemi daha yakın zaman önce kabul edilmiştir ve şu anda buna nasıl yaklaşılacağına dair mevcut en iyi fikirlerden bazıları ancak son on yılda ya da bu civarda keşfedilmiştir (birkaç örnekteyse bu kitap yazıldığı sırada). Son teknolojinin önümüzdeki yirmi otuz yılda iyiden iyiye ilerleyeceğini düşünmek yanlış sayılmaz ve eğer sorunun çok güç olduğu anlaşılırsa, ciddi bir ilerleme hızı yüzyıl, hatta daha fazla devam edebilir. Süper zekâya ulaşılması ne kadar uzun zaman alırsa, süper zekâ

geldiğinde o kadar ilerleme kaydedilmiş olacaktır. Bu, süper zekâya geç ulaşılmasının daha iyi olduğu görüşünü destekleyecek önemli bir argümandır– keza, son derece erken ulaşılmasının daha iyi olduğu görüşü aleyhinde de çok güçlü bir argümandır.

Süper zekâya geç tarihte ulaşılmasının daha güvenli olabileceğine dair bir diğer neden ise bunun insan medeniyetinin çeşitli faydalı arka plan eğilimlerine kendisini göstermesi için daha fazla zaman bırakacak olmasıdır. Bu mülahazaya ne kadar ağırlık verileceği bu eğilimler konusunda ne kadar iyimser olunduğuna bağlıdır.

İyimser biri kuşkusuz bir dolu cesaretlendirici işarete ve umut dolu olanağa işaret edebilir. İnsanlar daha iyi geçinmeyi öğrenebilirler, böylece şiddet, savaş ve mezalim oranlarında azalma yaşanır; ayrıca küresel eşgüdüm ve siyasal bütünleşmenin kapsamı artabilir ve istenmeyen teknoloji yarışlarından kaçıp (aşağıda bundan daha fazla bahsedilecektir) bir zekâ patlamasından beklenen kazanımların geniş kesimlere üleştirileceği bir düzenleme yapma daha kolay hale gelebilir. Bu doğrultularda uzun vadeli tarihsel eğilimler olacağı gözükmektedir.[6]

Diğer yandan, iyimser biri insanlığın "akıl sağlığı düzeyi"nin bu yüzyıl boyunca artacağını düşünüyor olabilir: Önyargılar (artı eksiye vurulduğunda) azalacak, içgörüler artacak ve insanlar gelecekteki soyut olasılıklar ve küresel riskler konusunda düşünmeye daha alışkın hale geleceklerdir. Şansımız yaver giderse, hem bireysel hem de kolektif biliş yetisindeki epistemik standartlarda genel bir yükseliş yaşanabilir. Bir kez daha, bu doğrultularda eğilimler yok değildir. Bilimsel ilerleme daha fazlasının bilineceği anlamına geliyor. Ekonomik büyüme dünya nüfusunun çok büyük bir kısmına yeterli gıda (özellikle de hayatta beyin gelişimi için önemli olan ilk yıllarda) ve nitelikli eğitime erişim olanağı sağlayabilir. Enformasyon teknolojisindeki ilerlemeler veri ve fikirleri bulmayı, özümsemeyi, değerlendirmeyi ve iletişmeyi daha kolay hale getirecektir. Ayrıca yüzyıl sonuna gelindiğinde, insanlık bir yüzyıllık daha hata yapmış olacaktır ki, bunlardan bazı dersler çıkarılabilir.

Potansiyel gelişmelerden birçoğu yukarıda bahsedilen anlamda müphemdir: Bazı varoluşsal riskler artarken, diğerleri azalır. Örneğin gözetlemedeki, veri madenciliğindeki, yalan yakalamadaki, biometrikteki ve psikolojik veya nörokimyasal yollarla inançları ve arzuları manipüle etme konusundaki ilerlemeler uluslararası eşgüdümü ya da ülke içindeki teröristleri ve kanun kaçaklarını bastırmayı daha kolay hale getirerek bazı varoluşsal riskleri azaltabilir. Ne var ki aynı ilerlemeler istenmeyen toplumsal dinamikleri artırarak ya da kalıcı totaliter rejimlerin oluşumunu kolaylaştırarak bazı varoluşsal riskleri de artırabilir.

Önemli cephelerden biri örneğin genetik seçilim yoluyla biyolojik biliş yetisinin geliştirilmesidir. Bu konuyu 2. ve 2. Bölüm'de ele aldığımızda, süper zekânın en radikal biçimlerinin makine zekâsı biçiminde ortaya çıkmasının daha muhtemel olacağı

sonucuna varmıştık. Bu iddia bilişsel ilerlemenin makine süper zekâsı yolunda ve yaratılmasında önemli bir rol oynamasıyla tutarlıdır. Bilişsel geliştirmenin açıkça risk azaltıcı olduğu görünebilir: Kontrol probleminde çalışan insanlar ne kadar akıllıysa, bir çözüm bulmaları o kadar muhtemeldir. Gelgelelim biliş yetisindeki ilerlemeler makine zekâsının gelişimini hızlandırabilir ve böylece sorun üzerinde çalışmak için mevcut zamanı azaltabilir. Biliş yetisindeki ilerlemelerin konuyla ilintili başka sonuçları da olacaktır. Bu meseleler yakından ele alınmayı hak ediyor. ("Biliş yetisindeki ilerlemeler" hakkında aşağıda yazılanların çoğu bireysel ya da kolektif epistemik etkililiğimizi artırmanın biyolojik olmayan araçları için de aynı derecede geçerlidir.)

Değişim hızları ve biliş yetisindeki ilerlemeler

İnsan entelektüel yetisinin ortalamasındaki ya da üst düzey kesimindeki artışın teknolojik ilerlemeyi çeşitli makine zekâsı biçimlerine doğru ilerleme, kontrol probleminde ilerleme ve çok çeşitli diğer teknik ve ekonomik hedeflerde ilerleme gibi çeşitli alanlarda hızlandırması muhtemeldir. Bu tür bir hızlanmanın net etkisi ne olurdu?

Sınırlı bir örnek olarak "evrensel hızlandırıcı"yı düşünün: Gerçekten *her şeyi* hızlandıran hayali bir müdahale. Bu tür bir evrensel hızlandırıcının eylemi sadece zaman ölçüsünün keyfî şekilde yeniden ölçeklendirilmesine tekabül edecek ve gözlemlenen sonuçlarda niteliksel bir değişim doğurmayacaktır.[7]

Eğer biliş yetisindeki ilerlemelerin işleri genel olarak hızlandırabileceği fikrini anlamlandıracaksak, evrensel hızlandırıcıdan başka bir kavrama ihtiyaç duyacağımız açıktır. Daha umut vaat edici yaklaşımlardan biri, bir ilerleme türündeki değişim hızıyla karşılaştırıldığında belli tür bir süreçteki değişim hızının nasıl artırabileceğine odaklanmaktır. Bu türde diferansiyel hızlanma bir sistemin dinamiğini etkileyebilir. Dolayısıyla şu kavramı bir düşünün:

> *Makroyapısal gelişim hızlandırıcısı:* İnsanlığın durumunun makro-yapısal özelliklerinin gelişme hızını ivmelendirirken mikro düzeydeki insan meselelerinin gelişme hızını aynı bırakan bir kumanda kolu.

Bu kumanda kolunun eksi ivme yönünde çekildiğini tahayyül edin. Dünya tarihinin büyük çarkına bir fren darbesi! Kıvılcımlar saçılır ve metal sesi duyulur. Çark daha rahat bir hıza kavuştuktan sonra, sonuç teknolojik yeniliğin daha yavaş gerçekleştiği ve siyasal yapıdaki ve kültürdeki temel ya da küresel açıdan kayda değer değişimin eskisi kadar sık ve ani yaşanmadığı bir dünyadır. Bir çağ yerini diğerine bırakmadan önce çok sayıda kuşak gelip geçer. Bir yaşam süresi boyunca bir insan insanlığın durumunun temel yapısında çok az değişiklik görür.

Türümüzün hayatta olduğu dönemin büyük bir kısmında makro yapısal gelişim bugünkünden daha yavaştı. Elli bin yıl önce, tek bir önemli teknolojik icat olmadan,

insan bilgisi ve anlayışında kayda değer bir artış ve küresel açıdan anlamlı bir siyasal değişim olmadan tam bin yıl geçebilirdi. Oysa mikro düzeyde insan ilişkileri kaleydoskopu doğumlar, ölümler ve kişisel ve yerel açıdan kayda değer diğer olaylarla makul bir hızda ilerliyordu. Ortalama bir insanın bir günü Buzul Çağı'nda bugüne kıyasla muhtemelen daha hareketli geçiyordu.

Makroyapısal gelişim hızını değiştirmenizi sağlayacak bir sihirli kol bulsanız, ne yapmanız gerekir? Hızlandırmanız mı yavaşlatmanız mı yoksa olduğu gibi bırakmanız mı?

Bu soru, gayri şahsi bakış açısını benimsediğimizde, varoluşsal riske etkilerini ele almamızı gerektiriyor. İki tür risk arasında ayrım yapmamız gerekiyor: "Durum riskleri" ve "adım riskleri". Durum riski belli bir durumda olmakla bağlantılı risktir ve bir sistemin maruz kaldığı toplam risk miktarı sistemin bu durumda ne kadar süre kaldığının doğrudan bir işlevidir. Doğadan gelen riskler durum risklerinin tipik örneğidir: Ne kadar maruz kalırsak, bir asteroid çarpması, süper yanardağ patlaması yaşanması, gama ışını patlaması, doğal yollarla salgın olması ya da kainatın sillelerinden başka birini yememiz o kadar muhtemeldir. İnsandan kaynaklı bazı riskler de durum riskleridir. Bireysel düzeyde, bir askerin kafasını siperin yukarısına çıkarması ne kadar uzun sürerse, bir düşman askeri tarafından vurulma ihtimali o kadar artar. Varoluşsal düzeyde de de insandan kaynaklı durum riskleri vardır: Uluslararası düzeyde anarşik bir sistemde ne kadar uzun yaşarsak, termonükleer bir dev savaşın ya da diğer kitle imha silahlarıyla yürütülen büyük bir savaşın medeniyeti yok etme ihtimali o kadar yüksektir.

Halbuki bir adım riski zorunlu ya da istenen bir geçişle bağlantılı bir risktir. Geçiş tamamlandığında, risk kaybolur. Bir geçişle bağlantılı adım riski miktarı genellikle geçişin ne kadar uzun sürdüğünün basit bir işlevi değildir. Bir mayın tarlasında yürümenin barındırdığı riski iki kat hızlı koşarak yarı yarıya azaltmazsınız. Hızlı bir havalanmaya bağlı olduğunda süper zekâ yaratmak bir adım riski olabilir: Havalanmaya bağlantılı belli bir risk olacak ve büyüklüğü hangi hazırlıkların yapıldığına bağlı olacaktır; ama risk miktarı havalanmanın yirmi milisaniyede mi yoksa yirmi saatte mi gerçekleştiğine pek de bağlı olmayabilir.

Dolayısıyla farazi bir makroyapısal gelişim hızlandırıcısı hakkında şunları söyleyebiliriz:

> Varoluşsal durum riskleriyle ilgilendiğimiz sürece, hızlandırmadan yana olmalıyız, yeter ki başka varoluşsal risklerin epeyce azaltıldığı bir geçiş sonrası çağa ulaşma konusunda gerçekçi bir olasılığa sahip olduğumuzu düşünelim.
>
> Eğer varoluşsal felakete yol açması mukadder bir adımın bizi beklediği biliniyorsa, o zaman makroyapısal gelişim hızını düşürmemiz gerekir (hatta tersine çevirmemiz gerekir) ki, daha fazla kuşağa, perde inmeden önce yaşama şansı

tanıyabilelim. Fakat aslında insanlığın yok olmaya mahkûm olduğundan bu denli emin olmak aşırı kötümserliktir.

Şu anda, varoluşsal durum riski düzeyi görece düşük gözüküyor. Eğer insanlığı mevcut durumunda dondurup teknolojik makro-koşullar tahayyül edersek, varoluşsal bir felaketin sözgelimi on yıllık bir zaman ölçeğinde gerçekleşmesi pek muhtemel görünmüyor. Dolayısıyla on yıllık bir gecikme –mevcut gelişim aşamamızda ya da durum riskinin düşük olduğu başka bir zamanda gerçekleştiği takdirde– varoluşsal açıdan ancak çok tali bir durum riski oluşturacaktır, oysa sonraki teknolojik gelişmelerde on yıllık gecikme, mesela ileriki varoluşsal adım risklerine hazırlık için daha fazla zaman bırakarak ciddi bir yarar sağlayabilir.

Sonuç: Makroyapısal gelişim hızının esas önem arz ettiği kısım, kilit adım riskleriyle karşılaşma zamanı geldiğinde insanlığın ne kadar hazırlıklı olduğunu etkilemesidir.[8]

Dolayısıyla sormamız gereken soru bilişsel ilerlemenin (ve ona eşlik eden makro yapısal gelişimdeki hızlanmanın) kritik dönemeçte beklenen hazırlıklı olma düzeyini nasıl etkileyeceğidir. Yüksek zekâyla daha az sürecek bir hazırlık dönemini mi tercih etmeliyiz? Yüksek zekâyla, hazırlık zamanı daha etkili şekilde kullanılabilir ve son kritik adım daha zeki bir insanlıkla atılacaktır. Yoksa bize daha fazla hazırlık zamanı tanıdığı takdirde mevcut zekâ düzeylerine daha yakın olmayı mı tercih etmeliyiz?

Hangi seçeneğin daha iyi olduğu, hazırlanılan zorluğun doğasına bağlıdır. Eğer zorluk deneyimlerden öğrenmenin kilit olduğu bir problemi çözmekse, o zaman hazırlık döneminin kronolojik uzunluğu belirleyici etken olabilir, zira gerekli deneyimin birikmesi için zamana ihtiyaç vardır. Bu tür bir zorluk neye benzerdi? Farazi örneklerden biri, ileride bir gün geliştirileceğini tahmin edebileceğimiz ve müstakbel bir savaşın varoluşsal felakete yol açma ihtimalini sözgelimi onda bire çıkaracak yeni bir silah teknolojisi olacaktır. Dolayısıyla bizi bekleyen zorluğun böyle bir doğası varsa, o zaman makroyapısal gelişim hızının yavaş olmasını ve bu sayede türümüzün yeni silah teknolojisinin icat edildiği kritik adımdan önce kendini toparlaması için daha fazla zamana sahip olmasını isteyebiliriz. Eksi ivmelenme yoluyla elde edilen atıfet günleri sırasında türümüzün savaştan kaçmayı öğrenebileceğini –uluslararası ilişkilerin yüzyıllar boyunca gırtlak gırtlağa savaşmış olan ama şimdi huzurlu ve göreli bir uyum içinde birarada yaşayan Avrupa Birliği ülkeleri arasındaki ilişkilere benzeyebileceğini– umabiliriz. Barışın tesis edilmesi çeşitli medenileştirici süreçlerden ince katkılarla ya da altvaroluşsal darbelerin şok terapisi yoluyla (örneğin küçük nükleer savaşlar ve devletlerarası savaşların kalkması için gerekli küresel kurumların nihayet yaratılmasını sağlayabilecek geri çekilme ve kararlılıkla) gerçekleşebilir. Eğer bu tür bir öğrenme ya da uyum sağlama artan zekâyla fazla ivmelenmezse, zaman bilişsel gelişim istenmeyen bir durum haline gelecek ve işlerin daha da kızışmasından başka bir işe yaramayacaktır.

Gelgelelim olası bir zekâ patlaması farklı türde bir zorluk sunabilir. Kontrol problemi öngörü, muhakeme ve teorik içgörü gerektirir. Tarihsel deneyimin artmasının ne kadar işe yarayacağı muğlaktır. Zekâ patlamasının doğrudan deneyimlenmesi (iş işten geçene kadar) mümkün değildir ve birçok özellik bir araya gelerek kontrol problemini benzersiz ve ilgili tarihsel öncelden yoksun hale getirebilir. Bu nedenlerden ötürü zekâ patlamasından önce geçecek zaman miktarı başlı başına çok önem arz etmeyebilir. Belki de önemli olan, (a) patlama zamanına gelindiğinde kontrol probleminde ulaşılan entelektüel ilerleme miktarı ve (b) mevcut en iyi çözümleri uygulama (ve mevcut olanı irticalen geliştirme) zamanında mevcut olan zekâ ve yetenek miktarıdır.[9] Bu son etkenin bilişsel gelişime olumlu katkı sağlayacağı açıktır. Bilişsel gelişimin (a) etkenini nasıl etkileyeceğiyse biraz daha incelikli bir meseledir.

Daha önce belirtildiği üzere, bilişsel gelişimin genel bir makroyapısal gelişim hızlandırıcısı olacağını varsayalım. Bu durum zekâ patlamasının gelişini hızlandıracak, hazırlık için ve kontrol probleminde ilerleme yapmak işin mevcut zaman miktarını azaltacaktır. Normalde bu kötü bir şey olsa da, eğer entelektüel ilerleme için az zaman olmasının tek nedeni entelektüel ilerlemenin hızlandırılmasıysa, o zaman zekâ patlaması gerçekleştiğinde gerçekleşmiş olacak entelektüel ilerleme miktarında net bir azalmaya gerek olmayacaktır.

Bu noktada bilişsel gelişimin (a) etkeniyle nötr ilişkide olduğu gözükebilir: Başka koşullarda zekâ patlamasından önce gerçekleştirilecek olan aynı entelektüel ilerleme (kontrol probleminde ilerleme dahil) yine de yapılır, ama daha kısa bir zaman aralığında olur. Gelgelelim gerçekte bilişsel gelişimin (a) üzerinde olumlu bir etkisi olduğu pekâlâ görülebilir.

Bilişsel gelişimin zekâ patlaması gerçekleştiğinde daha fazla ilerleme kaydedilmiş olmasını gerektirmesinin nedenlerinden biri, kontrol problemindeki ilerlemenin entelektüel performansın aşırı düzeylerine özellikle bağlı olabilmesidir, makine zekâyı yaratmak için gerekli çalışmadan bile çok. Deneme yanılma ve deneysel sonuç biriktirmedeki rolü kontrol problemiyle ilişkili olarak çok sınırlı gözükürken, deneysel öğrenme yapay zekânın gelişiminde ya da tüm beyin emülasyonunda muhtemelen büyük bir rol oynayacaktır. Dolayısıyla zamanın aklın yerini ne ölçüde tutabileceği görevler arasında öyle değişiklik gösterebilir ki, bilişsel gelişim, kontrol probleminde ilerlemeyi, makine zekânın nasıl yaratılacağı probleminde ilerlemeyi etkilediğinden *daha fazla* etkileyecektir. Bilişsel gelişimin kontrol probleminde ilerlemeyi farklı farklı etkileyecek olmasının bir diğer nedeni bu tür bir ilerlemeye duyulan ihtiyacın bilişsel açıdan daha üstün toplumlar ve bireyler tarafından takdir edilmesinin daha muhtemel olmasıdır. Kontrol probleminin neden önemli olduğunu fark etmek ve bunu bir öncelik haline getirmek öngörü ve muhakeme gerektirir.[10] Ayrıca bu türden hiç aşina olunmayan bir probleme yaklaşmanın umut vaat edici yollarını bulmak sıradışı bir bilgelik gerektirebilir.

Bu mülahazalardan hareketle, zekâ patlamasının varoluşsal risklerine odaklanıldığı sürece bilişsel gelişimin istenen bir şey olduğu sonucuna muvakkaten varabiliriz. Paralel düşünme çizgileri (örneğin çevredeki deneyimlenmiş değişikliklere artımlı uyum ya da birçok kuşağa yayılan kültürel olgunlaşma ve kurum inşası sürecinden farklı olarak) öngörü ve güvenilir soyut muhakeme gerektiren zorluklardan doğan diğer varoluşsal riskler için de geçerlidir.

Teknoloji bağlaşımları

Yapay zekâ için kontrol problemini çözmenin çok zor olduğunu, tüm beyin emülasyonları için çözmenin daha kolay olduğunu ve dolayısıyla makine zekâsına tüm beyin emülasyonu yolundan varılmasının tercih edilir olacağını düşünen biri olduğunu varsayın. Tüm beyin emülasyonunun yapay zekâdan daha güvenli olup olmayacağı sorununa daha sonra geri döneceğiz. Ama şimdilik şunu söylemek istiyoruz, bu önermeyi kabul etsek bile, tüm beyin emülasyonu teknolojisini teşvik etmemiz gerektiği sonucu çıkmaz. Daha önce ele alınan nedenlerden biri, süper zekâya daha sonra varılmasının kontrol probleminde ilerlemeye ve diğer faydalı arka plan eğilimlerinin sonuçlanmasına daha fazla zaman tanımak için tercih edilebilir olacağıdır ve dolayısıyla eğer tüm beyin emülasyonunun her halükârda yapay zekâdan önce geleceğine eminsek, tüm beyin emülasyonuna varılmasını daha da hızlandırmak ters tepecektir.

Fakat tüm beyin emülasyonunun olabildiğince erken gelmesi en iyi seçenek olsa bile, *yine de* tüm beyin emülasyonu yönünde ilerlemeyi tercih edeceğimiz sonucu çıkmaz. Zira tüm beyin emülasyonu yönünde ilerlemenin tüm beyin emülasyonunu doğurmaması mümkündür. Bunun yerine, nöromortfik yapay zekâ (korteksin örgütlenmesinin bazı yönlerini taklit eden ama nöron işlevselliğini tam bir emülasyon oluşturacak kadar sadık şekilde çoğaltmayan yapay zekâ biçimleri) yaratabilir. Eğer –şuna inanmamamız için neden yok– bu tür nöromorfik bir yapay zekâ diğer türlü yaratılabilecek olan yapay zekâ türünden daha kötüyse ve eğer tüm beyin emülasyonunu teşvik ederek öncelikle nöromorfik yapay zekâya ulaşılmasını sağlarsak, o zaman sözde *en iyi* sonucun (tüm beyin emülasyonu) peşinden gitmemiz en kötü sonuca (nöromorfik yapay zekâya) varacaktır; oysa *ikinci en iyi* sonucun (sentetik yapay zekâ) peşinden gitseydik, gerçekten ikinci en iyiye (sentetik yapay zekâya) ulaşabilirdik.

Böylece "teknolojik bağlaşım" adını verebileceğimiz şeyin (farazi) bir örneğini tarif etmiş olduk.[11] Bu terim iki teknolojinin kestirilebilir bir zamanlama ilişkisine sahip olduğu, öyle ki teknolojilerden birini geliştirmenin ya zorunlu bir öncel olarak ya da bariz ve karşı konulmaz bir uygulama ya da müteakip adım olarak diğerinin gelişimine yol açma konusunda güçlü bir eğilim barındırdığı bir koşula atıfta bulunur. Diferansiyel teknolojik gelişim ilkesine başvurdumuzda teknoloji bağlaşımları hesaba katılmalıdır: Eğer Y'ye varmanın yegâne yolu öncesinde hiç istenmeyen bir X teknolojisi geliştir-

mekse ya da Y'ye ulaşmak hiç istenmeyen bir ilintili Z teknolojisini hemen üretecekse, istenen bir Y teknolojisinin gelişimini hızlandırmanın bir yararı yoktur. Sevdiğinizle evlenmeden önce, müstakbel hısımlarınızı bir düşünün.

Tüm beyin emülasyonu örneğinde teknoloji bağlaşımları derecesi tartışmalıdır. 2. Bölüm'de, tüm beyin emülasyonunun çeşitli kolaylaştırıcı teknolojilerde muazzam bir ilerlemeyi gerektirecek olsa da, büyük çaplı yeni bir teorik içgörü gerektirmeyebileceğini belirtmiştik. Daha özelde, insanın biliş yetisinin nasıl işlediğini anlamamızı gerektirmez, sadece farklı nöron türleri gibi beynin küçük parçalarının bilgi işlem modellerinin nasıl inşa edileceğini bilmemizi gerektirir. Ne var ki insan beyinlerini taklit etme (emülasyon) yeteneğini geliştirirken nöroanatomik açıdan muazzam bir veri toplanacaktır ve korteks ağlarının işlevsel modellerinin müthiş ilerletileceği kesindir. Bu tür bir ilerlemenin tam teşekküllü bir tüm beyin emülasyonundan önce nöromorfik yapay zekâyı kolaylaştırma şansının yüksek olduğu gözükecektir.[12] Tarihsel açıdan, yapay zekâ tekniklerine dair nörobilim ya da biyolojiden süzülüp gelen birçok örnek vardır. (Sözgelimi nöroanatomi çalışmalarının ilham verdiği McCulloch-Pitts nöronu, algılayıcılar ve diğer yapay nöronlar ve nöron ağları; davranışçı psikolojinin ilham verdiği takviyeli öğrenme; evrim teorisinin ilham verdiği genetik algoritmalar; motor planlama ve duyum algısı hakkındaki biliş bilimi teorilerinin ilham verdiği kapsama mimarileri ve algı hiyerarşileri; teorik immünolojinin ilham verdiği yapay bağışıklık sistemleri; böcek kolonileri ve diğer kendi kendini örgütleyen sistemlerin ekolojisinin ilham verdiği sürü zekâsı; ve hayvanların hareket yeteneğinin ilham verdiği robot bilimindeki reaktif ve davranış temelli kontrol.) Belki de daha önemlisi, beynin daha fazla incelenmesiyle potansiyel olarak cevaplanabilecek yapay zekâyla ilintili bir sürü önemli soru vardır. (Örneğin: Beyin yapılandırılmış temsilleri işler bellekte ve uzun süreli bellekte nasıl depolar? Bağlama problemi nasıl çözülür? Nöral kod nedir? Kavramlar nasıl temsil edilir? Kortikal işleme mekanizmasının kortikal sütun gibi standart bir birimi var mıdır, eğer varsa nasıl yapılandırılmıştır ve işlevselliğine ne kadar bağımlıdır? Bu tür sütunlar nasıl birleştirilebilir ve nasıl öğrenebilirler?)

Birazdan tüm beyin emülasyonunun, nöromorfik yapay zekânın ve sentetik yapay zekânın göreli tehlikeleri hakkında daha fazla şey anlatacağız, ama şimdiden bir diğer önemli teknoloji bağlaşımından bahsedebiliriz: Tüm beyin emülasyonu ile yapay zekâ arasındaki teknoloji bağlaşımı. Tüm beyin emülasyonu yönünde atılacak adımlar gerçekten tüm beyin emülasyonuna varsa bile (nöromorfik yapay zekâdan farklı olarak) ve tüm beyin emülasyonuna varış güvenli bir şekilde başarılsa bile, yine de başka bir risk devam edecektir: *İkinci bir geçişle*, tüm beyin emülasyonundan nihayetinde daha güçlü bir makine zekâsı biçimi olan yapay zekâya geçişle bağlantılı bir risk.

Daha kapsamlı bir analizde ele alınabilecek başka birçok teknoloji bağlaşımı vardır. Sözgelimi tüm beyin emülasyonu yönünde bir ilerlemenin daha genel olarak nörobilim-

deki ilerlemeleri güçlendirecektir.[13] Bunun yalan tespiti, nöropsikolojik manipülasyon teknikleri, bilişsel gelişim ve tıbbi ilerlemelerde daha hızlı atılım gibi çeşitli sonuçları olabilir. Aynı şekilde, bilişsel ilerleme yönünde atılacak adımlar (tutulan özgül yola bağlı olarak) sadece biliş yetisini geliştirmek için değil, diğer özelliklerin değiştirilmesi için de genetik seçilim ve genetik mühendisliği yöntemlerinin daha hızlı gelişimi gibi yan sonuçlar doğurabilir.

İkinci tahmin

Eğer keşfedilmiş olanların en iyi seçenek olarak uygulamaktan başka bir şey yapmayan kusursuz derecede hayırhah, akli ve birleşik bir dünya kontrolcüsü diye bir şey olmadığını hesaba katarsak, stratejik karmaşıklığın bir diğer katmanıyla karşılaşırız. "Neler yapılması gerektiği"ne dair soyut bir vurgu somut bir mesaj şeklinde cisimleştirilmelidir ve bu da retoriğin ve siyasal gerçekliğin alanına girer. Birbiriyle çatışan çeşitli nedenlerden ötürü orada göz ardı edilecek, yanlış anlaşılacak, çarpıtılacak ya da temellük edilecektir; bir pinball gibi etrafta sekecek, etki ve tepkilere sebep olacak, bir sürü sonuç doğuracaktır ama bunların serencamının özgün göndericinin niyetleriyle doğrudan bir ilişkisinin olması gerekmez.

İncelikli bir operatör bu tür etkileri öngörmeye çalışabilir. Örneğin tehlikeli bir *X* teknolojisini geliştirmek için gerekli araştırma için şu argüman taslağını ele alalım. (Bu taslağa uyan argümanlardan biri Eric Drexler'ın yazılarında bulunabilir. Drexler'ın örneğinde X= moleküler nanoteknolojidir.[14])

1 *X*'in riskleri çoktur.
2 Bu riskleri azaltmak ciddi bir hazırlık evresi gerektirir.
3 Ciddi hazırlık ancak *X* olasılığı toplumun geniş kesimlerince ciddiye alındığında başlayacaktır.
4 Toplumun geniş kesimleri *X* olasılığını ancak *X*'in geliştirilmesi için büyük bir hazırlık çabasına girişildiğinde ciddiye alacaklardır.
5 Bir ciddi araştırma çabası ne kadar erken başlatılırsa, *X*'e ulaşmak o kadar uzun sürecektir (çünkü önceden var olan kolaylaştırıcı teknolojilerin daha düşük bir düzeyinden başlar).
6 Bu nedenle ciddi bir araştırma çabası ne kadar erken başlarsa, ciddi hazırlığın gerçekleşeceği ciddi hazırlık dönemi o kadar uzun olacak ve riskler o kadar azalacaktır.
7 Bu nedenle, *X* yönünde ciddi bir araştırma çabası hemen başlatılmalıdır.

İlk başta yavaş gitmek ya da durmak için bir nedene benzeyen şey (*X*'in risklerinin büyük olması) bu düşünce çizgisinde tam tersi sonucun nedeni haline gelir.

İlintili argüman türlerinden biri, zayıflıklarımızın farkına varıp bizi varoluşsal bir felaket olasılığını azaltacak önlemler almaya itecek olması nedeniyle küçük ve orta ölçekli felaketleri –bağrımıza taş basıp– hoş görmek zorunda olduğumuzu söyler: Buradaki fikir, küçük ve orta ölçekli bir felaketin bir tür aşı işlevi gördüğü, yani görece atlatılabilir bir tehdit biçimiyle medeniyete meydan okuyup dünyayı o tehdidin varoluşsal çeşidiyle uğraşmaya hazırlayan bir bağışıklık yanıtını tetiklediği yönündedir.[15]

Bu "titre ve kendine gel" argümanları halkta bir tepkiyi doğuracağı umuduyla kötü bir şeyin olmasına izin vermeyi savunur. Burada bunlardan bahsetmemizin nedeni bunları onaylıyor olmamız değil, "ikinci tahmin" argümanları adını vereceğimiz fikri tanıtmaktır. Bu tür argümanlar başkalarına akıldışı diye davranıp onların önyargılarına ve yanlış kavrayışlarına oynayarak, mesele dürüst bir şekilde ve onların akli melekelerine doğrudan seslenerek ortaya konulsaydı elde edilemeyecek kadar iyi bir cevap almanın mümkün olduğunu savunur.

İkinci tahmin argümanlarının sunduğu stratejilerden yararlanarak uzun vadeli küresel hedeflere ulaşmak çok zor görünebilir. Kamusal söylemin pinball makinesinde bir oraya bir buraya fırlatıldıktan sonra bir mesajın nihai yolunu nasıl tahmin edebiliriz? Bunu yapmak, topolojisi içkaynaklı bir yeniden örgütlenmeden geçerken sistemin dışarıdan gelen beklenmedik olaylarla rahatsız edilebileceği uzun zaman dilimleri boyunca çeşitli hususiyetler ve dalgalanan etki düzeyleriyle bir dolu bileşen üzerindeki teorik etkilerini tahmin etmeyi gerektiriyor görünecektir: Kuşkusuz imkânsız bir görev![16] Gelgelelim uzun vadeli kesin sonuçlardan birinin şansını artırması beklenebilecek bir müdahaleyi saptamak için sistemin gelecekteki tüm güzergâhı hakkında ayrıntılı tahminlerde bulunmak zorunlu olmayabilir. Örneğin sadece görece yakın vadeli ve kestirilebilir sonuçları ayrıntılı bir tarzda düşünebilir ve bunlarla ilintili olarak iyi iş çıkaran bir eylemi seçebiliriz, ama sistemin kestirilebilirlik ufkunun ötesindeki davranışını rastgele yürüyüş olarak kurgulayabiliriz.

Gelgelelim ikinci tahmin hamlelerini önemsiz görmek ya da bunlardan sakınmak için ahlaki bir gerekçe ileri sürülebilir. Birbirini zekâ ile alt etmeye çabası sıfır toplamlı (hatta hem bu pratikle harcanacak zaman ve çaba hem de başkalarının gerçekten düşündüğünü keşfetmeyi ve kendi kanaatlerini ifade ederken güvenilmelerini genel olarak zorlaştıracak olması ihtimali dikkate alındığında, eksi toplamlı) bir oyuna benzer.[17] Stratejik iletişim pratiklerinden tam gaz yararlanılması açıkyürekliliği öldürecek ve hakikati siyaset şeytanlarının arkadan bıçakladığı gecede kendi başının çaresine bakması için yalnız bırakacaktır.

Yollar ve kolaylaştırıcılar

Bilgisayar donanımındaki ilerlemeleri yüceltmeli miyiz? Peki ya tüm beyin emülasyonu yolundaki atılımlar? Şimdi de bu iki soruyu ele alacağız.

Donanımdaki ilerlemenin sonuçları

Daha hızlı bilgisayarlar, makine zekâsı yaratmayı kolaylaştırır. Dolayısıyla donanımdaki ilerlemeyi hızlandırmanın sonuçlarından biri makine zekâsının gelişini hızlandırmasıdır. Daha önce ele alındığı üzere, gayri şahsi açıdan bakıldığında bu muhtemelen kötü bir şeydir, zira kontrol problemini çözmek için ve insanlığın daha olgun bir emniyet aşamasına ulaşması için gerekli zaman miktarını azaltır. Ama bunu kesin bir şey olarak görmemek gerekir. Süper zekâ diğer birçok varoluşsal riski ortadan kaldıracağı için, eğer bu diğer varoluşsal risklerin düzeyi çok yüksekse önceki gelişimi tercih etmek için neden olabilir.[18]

Zekâ patlamasını hızlandırmak ya da başlangıcını geciktirmek, donanımdaki ilerleme hızının varoluşsal riski etkileyebileceği yegâne kanal değildir. Bir diğer kanal da donanımın bir ölçüde yazılımın yerini alabilmesidir; dolayısıyla daha iyi donanım bir tohum yapay zekâ kodlamak için gerekli asgari yeteneği azaltır. Hızlı bilgisayarlar da kaba kuvvet tekniklerine (genetik algoritmalar ve diğer yarat-değerlendir-çöpe at yöntemleri gibi) daha çok; kullanmak için derin bir kavrayış gerektiren tekniklere daha az yaslanan yaklaşımların kullanılmasını teşvik edebilir. Eğer kaba kuvvet teknikleri daha anarşik ya da kesin olmayan sistem tasarımlarına kapılırsa, daha kesin mühendislik çalışmaları yapılmış ve teorik açıdan kontrollü sistemlere kıyasla burada kontrol problemini çözmek daha zor olduğundan, daha hızlı bilgisayarların varoluşsal riski artırması için bir risk daha doğacaktır.

Bir diğer neden ise süratli donanım ilerleyişinin bir havalanma ihtimalini artıracak olmasıdır. Yarı iletken sanayisindeki teknolojik ilerlemeler ne kadar hızlı gerçekleşirse, programlayıcıların belli bir performans düzeyindeki yeterliliklerini sömürmek için harcadıkları zamanın kişi başına düşen miktarı o kadar az olacaktır. Bu demektir ki, zekâ patlamasının mümkün olan en düşük donanım performansı düzeyinde başlatılması düşük ihtimaldir. Dolayısıyla bir zekâ patlamasının, donanım nihayetinde başarılı bir programlama yaklaşımının ilkin başarılı olabileceği asgari düzeyin ötesinde kayda değer şekilde ilerlediğinde başlaması daha muhtemeldir. Demek ki havalanma nihayet gerçekleştiğinde bir donanımsal sarkma olacaktır. 4. Bölüm'de gördüğümüz gibi, donanımsal sarkma havalanma sırasında uzlaşmazlığı azaltan ana etkenlerden biridir. Bu nedenle, donanımdaki süratli ilerlemenin süper zekâya geçişi daha hızlı ve patlayıcı hale getirmesi muhtemeldir.

Bir donanım sarkması yoluyla hızlı havalanma, geçiş risklerini birkaç açıdan etkileyebilir. En bariz olanı, hızlı bir havalanmanın geçiş devam ederken yanıt verip düzeltmeler yapmak için daha az fırsat sunması ve böylelikle riski artırabilecek olmasıdır. Bununla ilintili olarak, bir donanımsal sarkma kendi kendini tehlikeli derecede ilerleten bir tohum yapay zekânın yeterli donanımdan yararlanma yeteneğini sınırlayarak kısıtlanması ihtimalini azaltacaktır: Bir işlemci ne kadar hızlıysa, yapay

zekânın kendisini bir süper zekâya çabucak önyüklemesi için gerekli işlemci sayısı o kadar az olacaktır. Fakat bir donanımsal sarkmanın bir diğer etkisi de büyük projelerin üstünlüklerinden birinin (daha güçlü bilgisayarlar sunma yeteneğinin) önemini azaltarak büyük ve küçük projeler arasındaki oyun alanını düzleştirmesidir. Bu etki de şayet büyük projelerin kontrol problemini çözme ve ahlaken kabul edilebilir hedeflerin peşine düşme ihtimali daha yüksekse varoluşsal riski artırabilir.[19]

Keza daha hızlı bir havalanma için de üstünlükleri vardır. Hızlı bir havalanma bir tekel oluşma ihtimalini artıracaktır. Eğer bir tekel kurmak geçiş sonrası eşgüdüm problemlerini çözmek için yeterince önemliyse, ileride felaket düzeyinde eşgüdüm başarısızlıkları riskini azaltmak için zekâ patlaması sırasında büyük bir riski kabul etmeye değebilir.

Bilgisayardaki gelişmeler bir makine zekâsı devriminin sonucunu sadece makine zekâsının inşasında doğrudan bir rol oynayarak değil, toplum üzerinde zekâ patlamasının başlangıç koşullarının şekillenmesine dolaylı olarak yardımcı olan farklı etkilerde bulunarak da etkileyebilir. İnternet kişisel bilgisayarların düşük maliyetle kitlesel olarak üretilebilmesi için donanımın yeterince iyi olmasını gerektiriyordu ve şu anda insan faaliyetini birçok alanda (yapay zekâ çalışmaları ve kontrol problemi üzerine araştırmalar dahil) etkiliyor. (İnternet olmasaydı bu kitap yazılmayabilirdi ve siz de kitabı bulamayabilirdiniz.) Gelgelelim donanım halihazırda insan iletişimini ve düşüncesini kolaylaştırabilen birçok uygulama için yeterince iyidir ve bu alanlardaki ilerlemenin hızı donanımın ilerleme hızı tarafından ciddi şekilde sınırlandırılmaktadır.[20]

Artısı eksisiyle değerlendirildiğinde, bilgisayar donanımında daha hızlı ilerlemenin gayri şahsi değerlendirme kıstasları açısından istenmeyen bir durum olduğu gözükmektedir. Bu muvakkat sonuç, örneğin diğer varoluşsal risklerden ya da geçiş sonrası eşgüdüm başarısızlıklarından kaynaklı tehditlerin son derece yüksek olduğu anlaşılırsa tersine çevrilebilir. Her halükârda, donanımın ilerleme hızına çok fazla önem atfetmemek gerekir. Dolayısıyla zekâ patlamasının başlangıç koşullarını geliştirme çabalarımız muhtemelen başka parametrelere odaklanmalıdır.

Bir parametreyi nasıl etkileyeceğimizi göremediğimizde bile, arazinin stratejik yapısını haritalandırmada bir ilk adım olarak onun "gösterge"sini (örneğin bu parametrede bir artışın ya da azalmanın istenen bir şey olup olmadığını) belirlemek yararlı olabilir. Bilahare, parametreyi daha kolay şekilde manipüle etmemizi gerçekten sağlayan yeni bir kaldıraç noktası keşfedebiliriz. Veya parametrenin göstergesinin manipüle edilebilir başka bir diğer parametrenin göstergesiyle bağıntılı olduğunu keşfedebiliriz, böylece ilk analizimiz bu diğer parametreyle ne yapacağımızı kararlaştırmaya yardımcı olabilir.

Tüm beyin emülasyonu araştırmaları teşvik edilmeli midir?

Yapay zekâ için kontrol problemini çözmek ne kadar zor görünürse, daha risksiz bir alternatif olarak tüm beyin emülasyonunu teşvik etmek o kadar çekicidir. Fakat sağlam bir yargıya varabilmek için öncelikle analiz edilmesi gereken birkaç mesele var.[21]

Birincisi, daha önce ele alınan teknoloji bağlaşımları meselesi var. Tüm beyin emülasyonu geliştirme çabasının nöromorfik yapay zekâya (bilhassa riskli bir makine zekâ biçimine) varabileceğine dikkat çekmiştik.

Ama şimdilik, tüm beyin emülasyonunu gerçekten başardığımızı varsayalım. Bu yol yapay zekâdan daha güvenli mi olacaktır? Bu başlı başına karmaşık bir meseledir: tüm beyin emülasyonunun en az üç üstünlüğü olduğu *varsayılır*: (1) Performans nitelikleri yapay zekânınkilerden daha iyi anlaşılacaktır; (2) insan saikleri miras alacaktır; (3) daha yavaş bir havalanmaya yol açacaktır. Şimdi bunların her birini kısaca değerlendirelim.

1 Bir emülasyonun entelektüel performans niteliklerini anlamanın bir yapay zekânınkini anlamaktan daha kolay olması akla yatkın geliyor. İnsan zekâsının zayıf ve güçlü yanları konusunda bir dolu deneyime sahibiz, ama insan düzeyinde yapay zekâyla ilgili bir deneyimimiz yok. Gelgelelim dijitalleştirilmiş bir insan zekâsının anlık görüntüsünün neler yapıp yapamayacağını anlamak ile bu tür bir zekânın onun performansını artırmayı amaçlayan değişikliklere nasıl yanıt vereceğini anlamak aynı şey değildir. Oysa bir yapay zekâ hem statik hem de dinamik eğilimleri bakımından anlaşılabilir olacak şekilde dikkatle tasarlanabilir. Dolayısıyla tüm beyin emülasyonu benzer bir gelişim aşamasındaki türsel bir yapay zekâya kıyasla entelektüel performansı bakımından daha kestirilebilir olsa da, tüm beyin emülasyonunun işinin ehli güvenlik-bilinçli programlayıcılar tarafından yaratılan bir yapay zekâdan dinamik şekilde daha kestirilebilir olup olmayacağı açık değildir.

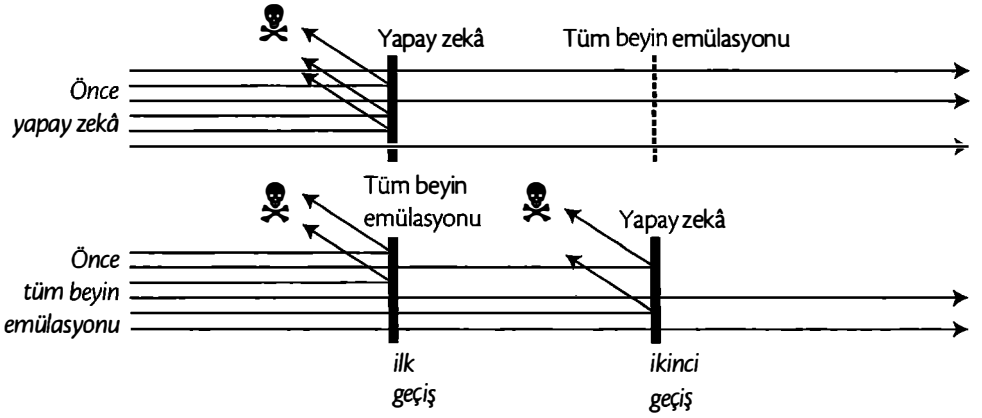
2 İnsan şablonunun güdülerini miras alan bir emülasyona gelince, bunun da garanti olduğu söylenemez. İnsanın değerlendirme eğilimlerini yakalamak üst düzey bir emülasyon gerektirebilir. Bazı bireylerin güdüleri kusursuz şekilde yakalanmış *olsa* bile, ne kadar güvenlik elde edileceği net değildir. İnsanlar güvenilmez, bencil ve acımasız olabilir. Şablonların istisnai erdemlerinden ötürü seçileceğini umut etsek bile, çok yabancı koşullara nakledilen, zekâsı insanüstü düzeyde artırılan ve dünya üzerinde egemenlik fırsatıyla ayartılan birinin nasıl davranacağını önceden söylemek zor olabilir. Emülasyonların (sadece kâğıt kıskaçlarının değerli görülmesi ya da pi'nin basamaklarının keşfedilmesinden farklı olarak) *insan benzeri* güdülerinin olmasının en azından daha *muhtemel* olduğu doğrudur. İnsan doğasına ilişkin görüşlerinize bağlı olarak bu durum güven aşılayıcı olabilir de olmayabilir de.[22]

3 Tüm beyin emülasyonunun neden yapay zekâdan daha yavaş bir havalanmaya yol açacağı net değildir. Belki de tüm beyin emülasyonunda donanım sarkmasının daha az olması beklenebilir, zira tüm beyin emülasyonu hesaplama

açısından yapay zekânın ulaşabileceği verimliliğe ulaşamaz. Keza belki de bir yapay zekâ sistemi mevcut tüm bilgisayar gücünü tek bir bütünleşmiş zekâdan kolaylıkla toplayabilirken, tüm beyin emülasyonu nitelikli süper zekâdan vazgeçecek ve insanlığı sadece nüfustaki büyüklük ve hız bakımından geride bırakacaktır. Eğer tüm beyin emülasyonu daha yavaş bir havalanmaya yol açarsa, bunun kontrol problemini azaltmak bakımından yararları olabilir. Daha yavaş bir havalanma çokkutuplu bir sonucu da daha muhtemel hale getirecektir. Ama çokkutuplu bir sonucun istenir olup olmadığı çok şüphelidir.

İlk önce tüm beyin emülasyonuna ulaşmanın daha güvenli olduğu fikrinin barındırdığı bir diğer önemli sorun ise *ikinci bir geçişle* uğraşmak zorunda kalmaktır. İnsan düzeyinde makine zekâsının ilk biçimi emülasyon temelli olsa bile, yine de yapay zekâ geliştirmek mümkün olmaya devam edecektir. Yapay zekâ olgun biçimiyle tüm beyin emülasyonundan önemli üstünlüklere sahiptir ve dolayısıyla yapay zekâyı nihayetinde daha güçlü bir teknoloji haline getirmektedir.[23] Olgun yapay zekâ tüm beyin emülasyonunu köhnemiş hale getirse de (bireysel insan zihinlerini korumak gibi özel bir amaç haricinde), tersi geçerli değildir.

Bunun ne anlama geldiğini soracak olursak, eğer ilkin yapay zekâ geliştirilirse, tek bir zekâ patlaması dalgası olabilir. Ama ilkin tüm beyin emülasyonu geliştirilirse, iki dalga olabilir: İlkin, tüm beyin emülasyonunun gelişi; sonradan da yapay zekânın gelişi. Tüm beyin emülasyonuna ilkin varılması yolundaki toplam varoluşsal risk, ilk geçişteki risk ile ikinci geçişteki riskin *toplamıdır* (ilk geçişin başarıldığını varsayarak konuşuyoruz); bkz. **ŞEKİL 13**.[24]



ŞEKİL 13 İlkin yapay zekâ mı yoksa tüm beyin emülasyonu mu? İlkin yapay zekâ senaryosunda, varoluşsal risk yaratan tek bir geçiş vardır. Tüm beyin emülasyonu senaryosundaysa iki riskli geçiş vardır: İlkin, tüm beyin emülasyonunun gelişimi, ardından yapay zekânın gelişimi. İlkin tüm beyin emülasyonunun geliştirildiği senaryoda toplam varoluşsal risk bunların toplamıdır. Gelgelelim eğer tüm beyin emülasyonunun zaten başarıyla yürürlüğe konduğu bir dünyada gerçekleşiyorsa, bir yapay zekâ geçişinin riski daha az olabilir.

Bir tüm beyin emülasyonu dünyasında yapay zekâ geçişi ne kadar daha güvenli olacaktır? İddialardan biri, eğer yapay zekâ geçişi bir tür makine zekâsı halihazırda elde edildikten sonra gerçekleşirse yapay zekâ geçişinin daha az patlayıcı olacağı yönündedir. Dijital hızlarda ve biyolojik insan nüfusunu misli misli aşan sayılarda çalıştırılan emülasyonlar bilişsel diferansiyeli azaltacak ve emülasyonların yapay zekâyı kontrol etmesini kolaylaştıracaktır. Bu çok güçlü bir iddia değildir, zira yapay zekâ ile tüm beyin emülasyonu arasındaki boşluk hâlâ daha büyük olabilir. Gelgelelim eğer emülasyonlar sadece daha hızlı ve sayıca daha çok değil, aynı zamanda biyolojik insanlardan nitel açıdan biraz daha akıllıysa (ya da en azından insan dağılımının en üst diliminden alınmaysa), o zaman ilkin tüm beyin emülasyonu senaryosunun yukarıda ele aldığımız insanın bilişsel gelişiminin üstünlüklerine benzer üstünlükleri olacaktır.

Bir diğer iddia ise tüm beyin emülasyonuna geçişin önde gidenin liderliğini uzatacağıdır. Önde gidenin tüm beyin emülasyonu teknolojisi geliştirme bakımından en yakın takipçisinden altı ay önde olduğu bir senaryo düşünün. Yaratılacak ilk emülasyonların işbirliğine açık, güvenlik odaklı ve sabırlı olduğunu varsayın. Eğer bu emülasyonlar hızlı bir donanımda çalıştırılırlarsa, güvenli bir yapay zekânın nasıl yaratılacağına kafa yormaya öznel açıdan çok uzun zaman harcayabilirler. Örneğin bir 100.000x hızda çalıştırılırlarsa ve altı aylık yıldız zamanı boyunca kontrol probleminde rahat rahat çalışabilirlerse, diğer emülasyonlarla rekabete girmeden elli binyıl boyunca kontrol problemine kafa yorabilirler. Yeterli donanım olduğunda, altproblemler üzerinde bağımsız şekilde çalışmak için çok sayıda kopya çıkararak ilerleyişlerini hızlandırabilirler. Eğer önde giden altı aylık üstünlüğünü bir tekel oluşturmak için kullanırsa, kendi emülasyon yapay zekâsı geliştirme takımına kontrol problemi üzerinde çalışmak için sınırsız zaman imkânı sunabilir.[25]

Artısı eksisiyle değerlendirildiğinde, bir yapay zekâ geçiş riskinin tüm beyin emülasyonu önce geldiği takdirde azalacağı görünmektedir. Gelgelelim yapay zekâdaki artık riski, önceden bir tüm beyin emülasyonu geçişinin gerçekleştirilmesi riskiyle birleştirdiğimizde, toplam ilkin tüm beyin emülasyonu senaryosunda varoluşsal riskin ilkin yapay zekâ senaryosuna kıyasla ne kadar çok olacağı oldukça belirsizdir. Ancak biyolojik insanlığın bir yapay zekâ geçişini idare edebilmesi konusunda çok kötümsersek (insan doğasının ya da medeniyetinin bu zorlukla karşılaştığımız an geldiğinde ilerlemiş olabileceğini hesaba kattıktan sonra) ilkin tüm beyin emülasyonu senaryosu çekici gelecektir.

Tüm beyin emülasyonu teknolojisinin teşvik edilip edilmeyeceğini kararlaştırmak için, başka hususları da dikkate almak gerekiyor. En önemlisi, yukarıda bahsedilen teknolojik bağlaştırma vardır: Tüm beyin emülasyonu yönünde atılacak adımlar bunun yerine bir nöromorfik yapay zekâ doğurabilir. Bu durum tüm beyin emülasyonu aleyhinde bir etkendir.[26] Kuşkusuz *bazı* nöromorfik tasarımlar kadar güvenli

olmayan *bazı* sentetik yapay zekâ tasarımları vardır. Fakat beklenti bakımından, nöromorfik tasarımların daha güvensiz olduğu görünüyor. Bunun temellerinden biri taklidin anlayışı ikame edebilmesidir. Bir şeyi sıfırdan inşa etmek için genellikle sistemin nasıl işleyeceğine dair iyi bir anlayışa sahip olmak gerekir. Bu tür bir anlayış mevcut bir sistemin salt özelliklerini kopyalamak için zorunlu olmayabilir. Tüm beyin emülasyonu biyolojinin toptan kopyalanmasına yaslanır ve bunun için biliş yetisinin kapsamlı bir hesaplamalı sistem düzeyinde anlaşılması gerekmeyebilir (ama bileşen düzeyinde epey bir anlayışa ihtiyaç olacaktır). Nöromorfik yapay zekâ bu açıdan tüm beyin emülasyonuna benzetilebilir: Mühendislerin sistemin nasıl işlediğine dair her zaman derin bir matematiksel kavrayışa sahip olmadığı koşullarda biyolojiden aşırılan parçaların bir araya getirilmesiyle başarılacaktır. Ama nöromorfik yapay zekâ başka bir açıdan tüm beyin emülasyonundan *farklı* olacaktır: Fabrika ayarlarında insan güdüleri olmayacaktır.[27] Bu husus nöromorfik yapay zekâ üretmesi muhtemel olduğu ölçüde tüm beyin emülasyonu yaklaşımını takip etmenin aleyhinde bir argümandır.

Dikkate alınması gereken ikinci bir husus: Tüm beyin emülasyonunun, gelişini bize daha önceden haber vermesi muhtemeldir. Yapay zekâda birilerinin beklenmedik bir kavramsal atılım gerçekleştirmesi her zaman mümkündür. Tüm beyin emülasyonuysa birçok zahmetli ön adım gerektirecektir: Yüksek çıktılı tarama özellikleri, görüntü işleme yazılımı, ayrıntılı nöral modelleme çalışması. Dolayısıyla tüm beyin emülasyonunun eli kulağında olmadığından emin olabiliriz (en az 15-20 yılı vardır). Bu demektir ki, tüm beyin emülasyonunu hızlandırma çabaları esasen makine zekâsının görece geç geliştirildiği senaryolarda fark yaratacaktır. Bu durum zekâ patlamasının diğer varoluşsal riskleri boşa çıkarmasını isteyen ama kontrol problemi çözülmeden önce bir zekâ patlamasını vaktinden önce tetiklemekten korkarak yapay zekâyı desteklemeye çekinen birine tüm beyin emülasyonu yatırımlarını çekici kılabilir. Gelgelelim ilintili zaman ölçekleri konusundaki belirsizlik muhtemelen şu anda bu düşünceye çok fazla ağırlık verilmesini engellemektedir.[28]

Dolayısıyla tüm beyin emülasyonunu teşvik etme stratejisinin en çekici olduğu durumlar: (a) İnsanların yapay zekâ için kontrol problemini çözmesi konusunda çok kötümserseniz, (b) nöromorfik yapay zekâ, çokkutuplu sonuçlar ya da bir ikinci geçişin riskleri konusunda çok endişeli değilseniz; (c) tüm beyin emülasyonu ile yapay zekânın varsayılan zamanlamasının yakın olduğu düşünülüyorsa ve (d) süper zekânın ne çok geç ne de çok erken geliştirilmesi tercih ediliyorsa.

Kişiyi etkileyen perspektif hızdan yanadır

Korkarım blog yorumcusu "washbash" şunları yazdığında birçoklarının hislerine tercüman olmuştur:

> Ben içgüdüsel olarak hızlı gitmeyi düşünüyorum. Bunun dünya için daha iyi olduğunu düşündüğüm için değil. Neden ölüp gittiğimde dünyayı umursayayım ki? Hızlı gitmesini istiyorum, lanet olsun! Bu durum teknolojik açıdan daha ileri bir geleceği deneyimleme şansımı artırıyor.[29]

İnsanı etkileyen açıdan bakıldığında, varoluşsal riskler oluşturabilecek her türlü radikal teknolojiyi aceleye getirmek için birçok nedenimiz var. Bunun nedeni, varsayılan sonucun şu an yaşayan hemen herkesin yüzyıl içinde ölecek olmasını öngörmesidir.

Aceleye getirme görüşü hayatlarımızı uzatabilecek ve bu suretle zekâ patlaması civarında hâlâ yaşıyor olabilecek mevcut nüfusun beklenen kesimini artırabilecek olan teknolojilerde özellikle güçlüdür. Eğer makine zekâsı devrimi iyi giderse, ortaya çıkan süper zekânın o sırada var olan insanların yaşamlarını süresiz uzatması, üstelik onları yaşatmakla kalmayıp, sağlık ve gençlik kazandıracak ve yetilerini şu anda insanlar için düşünemeyeceğimiz boyutta artıracak ya da zihinlerini bir dijital substrata yükleyerek ve özgürleştirilmiş ruhlarına son derece iyi hissettiren sanal tecessümler bahşederek ölümlülükten kurtulmalarına yardımcı olacak araçlar bulması neredeyse kesindir. Hayat kurtarmayı vaat etmeyen teknolojiler söz konusu olduğunda aceleye getirme görüşü zayıftır, ama yaşam standartlarını yükseltme umudu tarafından belki de hâlâ yeterince desteklenmektedir.[30]

Aynı muhakeme çizgisi, insanı etkileyen perspektifin zekâ patlamasının başlangıcını hızlandırmayı vaat eden birçok riskli teknolojik yenilikten yana olması anlamına gelir, üstelik bu yenilikler gayri şahsi perspektiften tercih edilmese bile. Bu tür yenilikler insan sonrası çağın doğuşunu görecek kadar yaşarsak bireysel olarak kendi bakış açımıza tutunmamız, şafak vaktini kısaltabilir. Dolayısıyla insanı etkileyen perspektiften bakıldığında, daha hızlı donanım ilerlemesinin istenen bir şey olduğu görünmektedir, tıpkı tüm beyin emülasyonu yönünde daha hızlı ilerleme gibi. Varoluşsal risk üzerindeki bir ters etkiyle karşılaştırıldığında, şu an yaşayan insanların yaşam süresi boyunca gerçekleşecek zekâ patlamasının şansının artmasından elde edilecek kişisel yarar muhtemelen ağır basmaktadır.[31]

İşbirliği

Önemli parametrelerden biri makine zekâsının gelişiminde dünyanın hangi derecede işbirliği yapıp eşgüdümde bulunacağıdır. İşbirliğinin birçok yararı olacaktır. Bu parametrenin sonucu nasıl etkileyebileceğine ve işbirliğinin yoğunluğunu ve boyutunu artırmak için hangi kaldıraçlara sahip olabileceğimize bir bakalım.

Yarış dinamiği ve tehlikeleri

Yarış dinamiği bir proje başka bir proje tarafından geride bırakılmaktan korktuğunda ortaya çıkar. Bu durum birden çok projenin gerçekten var olmasını gerektirmez. Sadece

tek bir projenin olduğu bir durum da, eğer bu proje rakipsiz olduğunun farkında değilse, bir yarış dinamiği sergileyebilir. Eğer İtilaf devletleri Almanların atom bombası hedefine yakın olduğuna (yanlış bir görüş) inanmıyor olsalardı, muhtemelen atom bombasını bu kadar çabuk geliştirmeyeceklerdi.

KUTU 13: Riskte dibe doğru yarış

Birkaç takımın süper zekâ geliştirmek için rekabet içinde oldukları farazi bir yapay zekâ silahlanma yarışını ele alalım.* Her takım güvenliğe ne kadar yatırım yapacağına karar verir, güvenlik önlemlerinin geliştirilmesine harcanan kaynakların yapay zekânın geliştirilmesine harcanmayan kaynaklar olduğunu bilir. Tüm rakipler arasında bir anlaşma olmadığında (pazarlık ya da dayatma zorlukları çıkararak bu engellenebilir), o zaman risk konusunda dibe doğru bir yarış başlayabilir ve her takımı sadece asgari sayıda önlem almaya itebilir.

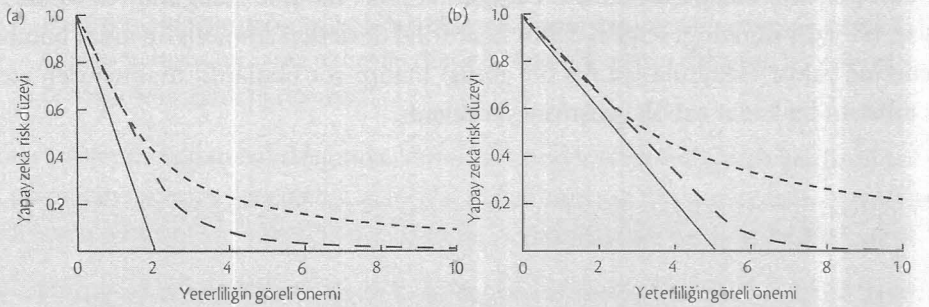
Her takımın performansını yeterliliğinin bir işlevi (ham yeteneğini ve şansını ölçen bir şey) ve güvenlik önlemlerinin maliyetine denk düşen bir ceza koşulu olarak oluşturmak mümkündür. En yüksek performansa sahip takım ilk yapay zekâyı inşa eder. Bu yapay zekânın riskliliği yaratıcılarının güvenliğe ne kadar yatırım yaptıklarına göre belirlenir. En kötü durum senaryosunda her takımın eşit yeterlilik düzeyi vardır. Dolayısıyla kazanan sadece güvenliğe yapılan yatırımla belirlenir: En az güvenlik önlemi alan takım kazanır. Bu oyunun Nash dengesi güvenliğe hiçbir şey harcamamaktır. Gerçek dünyada bu tür bir durum bir *risk artırma* yoluyla doğabilir: Geriye düşme korkusu duyan bir takım, rakiplerine yetişmek adına, aldığı riskleri artırabilir, onlar da karşılık verir ve azami risk düzeyine ulaşılır.

Yeterlilik mi risk mi?

Yeterlilikte çeşitlilikler olduğunda durum değişir. Yeterlilikteki değişiklikler güvenlik önlemlerinin masrafları için daha önemli hale geldikçe, risk artırma zayıflar: Fazladan risk almak yarışın düzenini değiştirmeyecekse, bunu yapmak için neden azdır. Bu durum yapay zekânın riskliliğinin, yeterliliğin önemine dayalı olduğunu gösteren **ŞEKİL 14**'teki çeşitli senaryolarda örneklendirilmiştir. Güvenlik yatırımları 1'den (tamamen güvenli bir yapay zekâ doğurur) 0'a (hiç güvenli olmayan bir yapay zekâ) kadar sıralanmıştır. *X*-ekseni bir takımın yapay zekâ yolundaki ilerleyişinin hızının belirlenmesinde güvenlik yatırımına karşı yeterliliğin göreli önemini temsil eder. (0,5'te, güvenlik yatırımı düzeyi yeterliliğin iki katı kadar önemlidir; 1'de, ikisi eşittir; 2'de yeterlilik güvenlik düzeyinin iki katı kadar önemlidir vesaire.) Y-ekseni yapay zekâ riskinin düzeyini ifade eder (yarışın kazananının elde ettiği azami faydanın beklenen kısmı).

---

* Bu modeldeki yardımlarından ötürü Carl Shulman ve Stuart Armstrong'a minnettarım. Ayrıca bkz. Shulman (2010a, 3): "Chalmers (2010) ABD Kara Harp Okulu'ndaki öğrenciler ve subaylar arasında ABD hükümetinin potansiyel bir felaket karşısında bile yapay zekâ araştırmalarını rakip güçlerin belirleyici üstünlük elde edeceği korkusuyla kısıtlamayacağı konusunda bir mutabakat olduğunu söyler."



ŞEKİL 14 Yapay zekâ teknoloji yarışlarındaki risk düzeyleri. Basit bir model olarak, ya iki takım ya da beş takım içeren bir teknoloji yarışında tehlikeli yapay zekânın risk düzeylerine karşı hangi projenin yarışı kazanacağının belirlenmesinde (güvenlik yatırımına karşı) yeterliliğin göreli önemi. Her grafik üç bilgi düzeyi senaryosunu gösteriyor: Sıfır yeterlilik bilgisi (aralıksız), özel yeterlilik bilgisi (kesik kesik) ve tam yeterlilik bilgisi (noktalı).

Tüm senaryolarda, ortaya çıkan yapay zekânın tehlikesinin yeterlilik hiçbir rol oynamadığında azami olduğunu, yeterliliğin önemi arttıkça tedricen azaldığını görüyoruz.

Uyuşan hedefler

Riski azaltmanın bir diğer yolu takımlara her birinin başarısında daha fazla pay tanımaktır. Eğer rakipler ikinci olmanın önemsedikleri her şeyin tamamen kaybedilmesi anlamına geldiğine ikna olurlarsa, rakiplerini atlatmak için gerekli tüm riskleri alacaklardır. Tersinden, eğer yarışı kazanmak daha önemsizse, takımlar güvenliğe yatırım yapacaklardır. Bu durum, çapraz yatırımın çeşitli biçimlerini teşvik etmemiz gerektiğini gösteriyor.

Rakiplerin sayısı

Rakip takımların sayısı ne kadar çoksa, yarış o kadar tehlikeli hale gelir: Birinci gelme şansı az olan her takım, ihtiyatı elden bırakmaya daha istekli hale gelir. **ŞEKİL 14** (iki takım) ile **ŞEKİL 14B**'yi (beş takım) karşılaştırdığımızda bunu görebiliriz. Her senaryoda daha fazla rakip daha fazla risk demektir. Eğer takımlar daha az sayıda rakip koalisyonlar haline gelirlerse risk azalacaktır.

Çok fazla bilgi göz çıkarır

Takımların yarıştaki konumlarını bilmeleri (sözgelimi yeterlilik puanlarını bilmeleri) iyi midir? Burada, karşıt etkenler devrededir. Liderin önde gittiğini bilmesi istenen bir şeydir (böylece ek güvenlik önlemlerine biraz alan kaldığını bilir). Ama yavaş hareket edenin geride kaldığını bilmesi istenen bir şey değildir (zira yarışta rakibini yakalamak için güvenlik önlemlerinde kesinti yapması gerektiğini doğrulayacaktır). Bu ödünleşimin iki yöne de gidebileceği düşünülebilirse de, modeller dolambaçsızdır: Bilgi (beklenti o ki) kötüdür.* **14A** ve **14B** şekillerinin

---

* Demem o ki, modeldeki bilgi sonradan bakıldığında her zaman kötüdür. Elbette bilginin gerçekte ne olduğuna bağlı olarak, bazı durumlarda bilginin bilinir hale gelmesi iyi olacaktır, özellikle de lider ile takipçisi arasındaki mesafe başta tahmin edilenden çok daha fazla olursa.

her biri üç senaryoyu anlatıyor: Düz çizgiler hiçbir takımın yeterlilik puanlarından herhangi birini (kendisininki dahil) bilmediği durumlara tekabül eder. Kesik çizgiler her takımın sadece kendi yeterliliğini bildiği durumları gösterir. (Bu durumlarda bir takım ancak yeterliliği düşükse fazladan risk alır.) Noktalı çizgiler ise her takım diğerlerinin yeterliliklerini bildiğinde olanları gösteriyor. (Eğer yeterlilik puanları birbirine yakınsa fazladan riskler alırlar.) Bilgi düzeyindeki her artışta yarış dinamiği kötüleşir.

Yarış dinamiğinin sertliği (yani rakiplerin güvenliktense hıza ne kadar öncelik verdikleri) yarışın yakınlığı, yeterlilik ve şansın göreli önemi, rakiplerin sayısı, rakip takımların farklı yaklaşımlar geliştirip geliştirmedikleri ve projelerin hangi derecede aynı amaçlara sahip oldukları gibi birkaç etkene dayalıdır. Rakiplerin bu etkenlere dair inançları da konuyla ilintilidir. (Bkz. **KUTU 13.**)

Makine süper zekâsının gelişiminde, en azından hafif bir yarış dinamiği olması muhtemel görünüyor, ciddi bir yarış dinamiğinin olmasıysa mümkündür. Yarış dinamiği bir zekâ patlaması olanağının sunduğu stratejik zorluk hakkında nasıl düşünmemiz gerektiğine dair önemli sonuçlara sahiptir.

Yarış dinamiği, projeleri süper zekâ yönünde daha hızlı ilerlemeye teşvik edebilirken kontrol problemini çözme yatırımlarını azaltabilir. Yarış dinamiğinin başka zararlı etkileri olması da mümkündür (rakipler arasında doğrudan husumetler gibi). İki ülkenin süper zekâyı diğerinden önce geliştirmek için yarıştığını ve bunlardan birinin önde gittiğinin görüldüğünü varsayalım. Kazananın her şeyi kazandığı bir durumda geride kalan proje pasif bir şekilde yenilgiyi beklemektense rakibine karşı umutsuzca saldırıya geçmeye ayartılabilir. Önde giden bu olanağı fark edince, erkenden davranıp saldırabilir. Eğer düşmanlar güçlü devletlerse, çatışma kanlı olabilir.[32] (Rakibin yapay zekâ projesine karşı bir "nokta operasyonu" daha büyük bir çekişmeyi tetikleyebilir, ama eğer ev sahibi ülke önlemler aldıysa bu riskin gerçeğe dönüşmesi mümkün olmayabilir.[33])

Devletlerin değil de küçük oluşumların (şirket laboratuvarları ya da akademik takımlar gibi) rekabet ettiği senaryolarda çatışmadan kaynaklı doğrudan yıkım muhtemelen çok daha az olacaktır. Fakat rekabetin genel sonuçları neredeyse bir o kadar kötü olabilir. Bunun nedeni rekabetten beklenen zararın esas kısmının savaştan değil, önlemlerin azaltılmasından kaynaklanmasıdır. Bir yarış dinamiği, gördüğümüz gibi, güvenlik yatırımını azaltacaktır; çatışma şiddet içermese bile işbirliği için fırsatları engelleyecektir, zira projelerin bir düşmanlık ve güvensizlik ikliminde kontrol problemini çözmek için fikirlerini paylaşma ihtimalleri daha düşük olacaktır.[34]

İşbirliğinin yararları

Dolayısıyla işbirliğinin birçok yararı vardır. Makine zekâsının geliştirilme hızını azaltır. Güvenliğe daha çok yatırım yapılmasına imkân tanır. Şiddetli çatışmalardan

kaçınır. Ayrıca kontrol probleminin nasıl çözüleceğine dair fikirlerin paylaşılmasını kolaylaştırır. Bu yararlara başka bir tane daha ekleyebiliriz: İşbirliği genellikle başarılı bir şekilde kontrol edilmiş zekâ patlamasının meyvelerinin daha hakkaniyetli paylaşıldığı sonuçlar üretir.

Bu geniş çaplı işbirliği kazanımların da geniş çaplı olarak paylaşılmasına yol açacak diye bir kaide yoktur. İlkesel açıdan, diğerkâm birinin elindeki küçük bir proje, ahlaken uygun tüm varlıklar arasında yararların eşit ya da hakkaniyetli paylaşıldığı bir sonuca yol açabilir. Gelgelelim çok sayıda sponsoru olan geniş çaplı işbirliklerinin paylaşım bakımından daha üstün olduğunu varsaymak için birkaç neden vardır. Bu tür nedenlerden biri sponsorların kendilerinin (en azından) hak ettikleri payı aldıkları bir sonucu tercih ettiklerinin varsayılmasıdır. Dolayısıyla geniş çaplı bir işbirliği projenin başarılı olduğunu varsayarsak görece birçok bireyin en azından hak ettiği payı alması anlamına gelir. Bir diğer neden ise geniş çaplı işbirliğinin, muhtemelen, o işbirliğinin dışında kalan kişilere de yarar sağlayacak görünmesidir. Geniş çaplı işbirliği daha fazla üye içerir, dolayısıyla dışarıda kalanların çıkarları için içerideki birileriyle kişisel bağları olacaktır. Keza geniş çaplı işbirliğinin herkese yarar sağlamak isteyen en az bir diğerkâm içermesi daha muhtemeldir. Ayrıca geniş çaplı işbirliğinin kamunun gözetimi altında yürümesi daha muhtemeldir ve bu sayede tüm pastanın bir grup programcı ya da özel yatırımcı tarafından ele geçirilmesi riski azalabilir.[35] Ayrıca başarılı işbirliği ne kadar büyükse, yararlarını dışarıda kalan herkese yaymanın masrafları o kadar azdır. (Sözgelimi eğer tüm insanların yüzde doksanı zaten işbirliğine dahilse, dışarıda kalanların tamamını kendi düzeylerine getirmeleri sahip olduklarının yüzde onundan fazlasına mal olmaz.)

Dolayısıyla geniş çaplı işbirlikleri muhtemelen kazanımların daha geniş ölçekte dağıtılmasına yol açacaktır (ama az sponsoru olan *bazı* projelerin de paylaşım bakımından mükemmel amaçları olabilir.) Ama kazanımların geniş ölçekte paylaşılması neden istenen bir şeydir?

Herkesin parsadan bir pay aldığı sonuçları desteklemek için hem ahlaki hem de tedbir amaçlı nedenler vardır. Ahlaki nedenle ilgili tek bir şey söyleyeceğiz: Eşitlikçi bir ilkeye yaslanmasına gerek yoktur. Örneğin adaletli olma temelinde bir argüman ileri sürülebilir. Makine süper zekâsı yaratan bir proje bir küresel risk dışsallığı dayatır. Gezegendeki herkes tehlike altındadır, kendi hayatlarını yaşamaya razı gelmeyenler ve bu şekilde zarar görmüş olan aileleri dahil. Herkes riski paylaştığından, herkesin de payını almasının adaletin asgari gerekliliği olduğu görünecektir.

Toplam (beklenen) iyilik miktarının işbirliği senaryolarında daha fazla olduğunun görünmesi de bu tür senaryoların ahlaken tercih edilir olmasındaki bir diğer önemli nedendir.

Kazanımların geniş ölçekte dağıtılmasından yana olmak için tedbir amaçlı neden ise ikiboyutludur. İlk boyut geniş çaplı paylaşımın işbirliğini teşvik etmesi ve bu suretle yarış dinamiğinin olumsuz sonuçlarını hafifletmesidir. Eğer bir projenin başarısından herkes eşit yarar elde edecekse süper zekâyı ilk kimin inşa edeceği konusunda kapışmak için daha az itki olur. Belli bir projenin sponsorları ganimetin evrensel olarak dağıtılmasına bağlılıklarının işaretini güvenilir şekilde vermekten de yarar sağlayabilir, zira gerçekten diğerkâm bir projenin daha fazla destekçisi ve daha az düşmanı olması muhtemeldir.[36]

Kazanımların geniş ölçekte paylaşılmasından yana olmanın tedbir amaçlı nedeninin diğer boyutu, aracıların riskten sakınıp sakınmadıkları ya da kaynaklar bakımından altdoğrusal olan fayda işlevlerine sahip olup olmadıklarıyla alakalıdır. Buradaki temel gerçek potansiyel kaynak pastasının muazzamlığıdır. Gözlemlenebilir evrenin göründüğü kadar boş olduğu varsayıldığında, yaşayan her bir insan için evrende birden fazla boş galaksi vardır. Ama birçok insan bir milyar galaksiye sahip olma konusunda milyarda bir şans sunan bir piyango biletindense, bir galaksi değerinde kaynağa belli bir erişim imkânına sahip olmayı tercih edecektir.[37] İnsanlığın kozmik birikiminin astronomik büyüklüğü dikkate alındığında, kişisel çıkarın genellikle her bir insana bir payı garanti altına alacak anlaşmalardan yana olduğu görünmektedir, üstelik her bir pay toplamın küçük bir parçasına denk düşse bile. Böyle büyük bir talih kuşu ufukta gözüktüğünde, önemli olan şey yok saymamaktır.

Kaynak pastasının muazzamlığı argümanı tercihlerin kaynaklarca karşılanabilir olmasını varsayar.[38] Bu varsayım her zaman geçerli değildir. Sözgelimi birkaç önde gelen etik teorisi (özellikle toplulaştırıcı sonuççu teoriler dahil) risk açısından nötr ve kaynakları bakımından doğrusal olan fayda işlevlerine denk düşer. Bir milyar galaksiden yararlanarak tek bir galaksiden bir milyar kat daha mutlu yaşamlar yaratılabilir. Dolayısıyla bir faydacının gözünde bir milyar kat daha değerlidir.[39] Ne var ki sıradan bencil insan tercihi işlevlerinin görece kaynakları karşılanabilir olduğu gözükmektedir.

Bu son ifadeye iki önemli kayıt düşülmelidir. Birincisi, birçok insan sıralamaya önem verir. Eğer birden çok aracı Forbes zenginleri listesinde tepede yer almak istiyorsa, o zaman hiçbir kaynak pastası herkesi tam olarak tatmin edecek kadar büyük değildir.

Düşülecek ikinci kayıt, geçiş sonrası teknoloji temelinin, maddi kaynakların eşi görülmedik çeşitlilikte ürüne (birçok insan tarafından çok değerli görülse de şu an herhangi bir fiyata mevcut olmayan bazı mallara) dönüştürülmesini sağlayacak olmasıdır. Bir milyarder bir milyonerden bin kat daha uzun yaşamaz. Gelgelelim dijital zihinler çağında bir milyarder bin kat daha fazla bilgisayar gücü sağlayabilir ve böylece bin kat daha uzun bir öznel ömrün sefasını sürebilir. Aynı şekilde, zihinsel kapasite satılık olabilir. Ekonomik sermayenin daha da büyük servet düzeylerinde sabit bir hızda hayati mallara dönüştürülebilir olduğu bu tür koşullarda; varlıklıların

(içlerinde hayırsever olmayanların) zenginliklerini uçağa, yata, sanat koleksiyonlarına ya da dördüncü veya beşinci konuta yatırdıkları bugünün dünyasına kıyasla sınırsız açgözlülük daha fazla anlam ifade edecektir.

Buradan bir egoistin kendi geçiş sonrası kaynakları bakımından riske kayıtsız olması gerektiği sonucu çıkar mı? Pek değil. Fiziksel kaynaklar keyfi ölçeklerde ömür boyu ya da zihinsel performansa dönüştürülebilir olmayabilir. Eğer bir hayat gözlemci anların önceki olayları hatırlayabileceği ve önceki tercihlerden etkilenebileceği şekilde sıralı olarak yaşanacaksa, o zaman bir dijital zihnin yaşamı *sıralı* bilgisayar işlemlerinin sayısındaki artıştan yararlanmadan keyfi olarak uzatılamaz. Fakat fizik kuralları kaynakların sıralı bilgi işlemlere hangi ölçüde dönüştürülebileceğini sınırlar.[40] Sıralı bilgi işlemdeki sınırlar görece mütevazı kaynakların ötesinde radikal derecede altdoğrusal olarak ölçeklendirecek şekilde bilişsel performansın bazı veçhelerini de kısıtlayabilir. Ayrıca bir egoistin kalitece ayarlanmış öznel yaşam yıllarının sayısı gibi normatif olarak fevkalade ilintili sonuç ölçekleriyle ilintili olarak bile riske duyarsız olmalıdır ya da olacaktır. Kesinkes 2 bin yıl fazladan yaşamak ile fazladan 30 bin yıl yaşamak konusunda onda bir şansa sahip olmak arasında bir tercih yapma imkânı sunulsa, sanırım çoğu insan ilkini tercih edecektir (her yaşam yılının eşit nitelikte olması durumunda bile).[41]

Gerçekte, kazanımların geniş ölçekte paylaştırılması konusundaki tedbir amaçlı görüş muhtemelen özneden özneye ve durumdan duruma değişir. Ama bütüne bakıldığında, eğer geniş bir paylaşım yapmanın bir yolu bulunursa, insanların istediklerinin (neredeyse tamamını) almaları daha muhtemel olacaktır – ve bu durum daha geniş çaplı bir paylaşıma bağlılığın işbirliğini güçlendirip bu suretle varoluşsal felaketten kaçınma şansını artıracağını hesaba katmadan önce bile doğrudur. Dolayısıyla geniş çaplı bir paylaşımdan yana olmanın sadece ahlaki zorunluluk değil, akla uygun da olduğu gözükmektedir.

İşbirliğinin en azından birazcık dikkate alınması gereken başka sonuçları da vardır: Geçiş öncesi işbirliğinin geçiş sonrası işbirliği düzeyini etkileme olanağı. İnsanlığın kontrol problemini çözdüğünü varsayalım. (Eğer kontrol problemi çözülmezse, geçiş sonrasında ne kadar işbirliği olduğu çok az anlam ifade edebilir.) Ele alınması gereken iki husus var. Birincisi, zekâ patlaması kazananın her şeyi aldığı bir dinamik *yaratmaz* (muhtemelen havalanma görece yavaş olduğu için). Bu örnekte, eğer geçiş öncesi işbirliğinin geçiş sonrası işbirliğine sistematik bir etkisi varsa, sonrasında işbirliğini teşvik etme eğilimiyle olumlu bir etkisi olması akla yatkındır. Özgün işbirliği ilişkileri geçişten sonra da devam edebilir: Ayrıca geçiş öncesi işbirliği insanların gelişmeleri geçiş sonrasında istenen (ve muhtemelen daha işbirlikçi) doğrultularda yönlendirmeleri için daha fazla fırsat sunabilir.

İkinci örnekte zekâ patlamasının doğası, kazananın her şeyi aldığı bir dinamiği teşvik eder (muhtemelen havalanma görece hızlı olduğu için). Bu örnekte, eğer havalanmadan önce kapsamlı bir işbirliği yoksa, bir tekelin ortaya çıkması muhtemeldir: Tek bir proje geçişi tek başına gerçekleştirecek, bir noktada süper zekâyla birleşmiş bir belirleyici stratejik üstünlük elde edecektir. Tekel, tanımı gereği, son derece işbirlikçi bir toplumsal düzendir.[42] Dolayısıyla geçiş öncesinde kapsamlı bir işbölümünün olması geçiş sonrasında muazzam bir işbirliğine yol açacaktır. Halbuki zekâ patlamasına giden süreçte biraz daha yüksek düzey bir işbirliği çok çeşitli ihtimallere kapı aralar. İşbirliği halindeki projeler yükselişlerini eşgüdümlü hale getirip, içlerinden herhangi biri belirleyici stratejik üstünlük elde etmeden geçişi birlikte gerçekleştirebilirler. Veya farklı sponsor grupları çabalarını birleştirip tek bir proje haline getirirken, bu projeye bir tekel oluşturma yetkisi vermeyi reddedebilirler. Örneğin bir uluslar konsorsiyumunun süper zekâ geliştirmek için ortak bir bilimsel proje oluşturduğunu, ama bu projeye süper yetkili bir Birleşmiş Milletler gibi bir şeye evrilme yetkisi vermediğini tahayyül edebiliriz; bunun yerine, öncesinde var olan bölünmüş dünya düzenini sürdürmeyi tercih edebilirler.

Dolayısıyla özellikle de bir hızlı havalanma örneğinde, geçiş öncesinde muazzam bir işbirliğinin geçiş sonrasında işbirliğini azaltması mümkündür. Gelgelelim işbirliği içindeki oluşumlar sonucu şekillendirebildiği ölçüde, geçiş sonrasındaki bölünmüşlükten feci bir sonuç doğmayacağını öngörürlerse, işbirliğinin olmadığı bir durumun doğmasına ya da sürmesine olanak tanıyabilir. Dolayısıyla geçiş sonrasındaki işbirliğinin geçiş sonrasında işbirliğinin azalmasına yol açtığı senaryolar, büyük oranda, geçiş sonrasındaki işbirliğinin azalmasının zararsız olduğu işbirlikleri olabilir.

Genel olarak geçiş sonrasında işbirliğinin yüksek olması iyi bir şey gibi gözükmektedir. Ekonomik rekabetin ve hızla büyüyen bir nüfusun Malthusçu bir koşula yol açtığı ya da evrimsel seçilimin insani değerleri aşındırdığı ve ödomanik olmayan biçimler seçtiği yahut rakip güçlerin savaş ve teknoloji yarışları gibi başka eşgüdüm başarısızlıkları tattıkları distopyacı dinamikler doğurabilir. Bu ihtimallerden sonuncusu, yani teknoloji yarışları ihtimali, eğer geçiş aracı bir makine zekâsı biçimindeyse (tüm beyin emülasyonu) bilhassa sorunlu olabilir, zira bilahare daha ileri bir makine zekâsı biçimine (yapay zekâya) geçiş için kontrol probleminin çözülmesi ihtimaline zarar verecek bir yeni yarış dinamiği yaratacaktır.

İşbirliğinin zekâ patlamasına giden süreçte çatışmayı azaltabileceğini, kontrol probleminin çözülme ihtimalini artırarak hem ahlaki meşruiyeti hem de ortaya çıkan kaynak tahsisatının sağduyulu istenirliğini güçlendireceğini yukarıda anlatmıştık. Dolayısıyla bu işbirliği yararlarına bir tane daha eklemek mümkündür: Geçiş öncesindeki bu geniş çaplı işbirliği geçiş sonrası çağda eşgüdüm bakımından önemli sonuçlar doğmasına yol açabilir.

Birlikte çalışmak

İşbirliği, işbirliği yapan oluşumların ölçeğine bağlı olarak farklı biçimlere bürünebilir. Küçük bir ölçekte, birbirleriyle rekabet halinde olduklarına inanan bireysel yapay zekâ takımları, çabalarını birleştirmeyi tercih edebilir.[43] Şirketler birleşebilir ya da çapraz yatırım yapabilirler. Daha büyük bir ölçekteyse, devletler büyük bir uluslararası projeye katılabilir. Bilim ve teknolojide büyük ölçekli uluslararası işbirliğinin (CERN, İnsan Genomu Projesi ve Uluslararası Uzay İstasyonu gibi) öncelleri vardır, ama güvenli bir süper zekâ geliştirmek için uluslararası bir proje çalışmanın güvenlik içerimlerinden ötürü farklı bir zorluk düzeyi teşkil eder. Açık bir akademik işbirliği olarak değil, son derece sıkı bir kontrollü ortak teşebbüs olarak oluşturulması gerekecektir. Belki de projeye dahil olan bilimciler proje boyunca dünyanın geri kalanından fiziksel olarak yalıtılmak zorunda kalacak ve dikkatle seçilmiş tek bir iletişim kanalı haricinde onlarla iletişimi engellenecektir. Gerekli güvenlik düzeyine şu anda erişmek neredeyse olanaksız olabilir, ama yalan yakalama ve gözetleme teknolojisindeki ilerlemeler sayesinde bu yüzyılın ileriki dönemlerinde mümkün hale gelebilir. Diğer yandan, geniş çaplı işbirliğinin çok sayıda araştırmacının projeye katılması gerektiği anlamına gelmediğini de hatırda tutmak gerekiyor; sadece, birçok insanın projenin amaçlarında söz hakkı olması anlamına gelir. İlkesel açıdan bakıldığında, bir proje, sponsoru tüm insanlık olan (sözgelimi Birleşmiş Milletler Genel Meclisi'nce temsil edilen) azami düzeyde geniş bir işbirliğini kapsayabilir, ama çalışmayı yürütmesi için sadece tek bir bilimci istihdam edebilir.[44]

İşbirliğine olabildiğince erken başlamak için, yani hangi projenin süper zekâya ilkin varacağına dair özgül bilgileri bizden saklayan bir cehalet perdesinden istifade etmek için bir neden vardır. Bitiş çizgisine yaklaştıkça rakip projelerin göreli şansına ilişkin belirsizlik azalacaktır ve yararları tüm insanlığa dağıtacak bir işbirliği projesine katılmak için önde giden projenin kendi çıkarlarına dayalı bir savunuda bulunmak zorlaşabilir. Diğer yandan, süper zekâ olasılığı şu anda olduğundan çok daha geniş ölçekte tanınmadan önce ve makine zekâsının yaratılmasına giden açıkça görülür bir yol olmadan önce dünya çapında resmi bir işbirliği tesis etmek de zor görünmektedir. Ayrıca işbirliği bu yolda ilerlemeyi teşvik edeceği ölçüde, güvenlik bakımından, daha önce belirtildiği üzere, zararlı bile olabilir.

Dolayısıyla şu anda işbirliğinin ideal biçimi ilk başta özgül resmileştirilmiş anlaşmalar gerektirmeyen ve makine zekâsında ilerlemeleri hızlandırmayan nitelikte olabilir. Bu kıstaslara uyan önerilerden biri süper zekânın ortak iyilik için olması gerektiği fikrine bağlılığımızı ifade eden uygun bir ahlaki normu savunmamızdır. Bu tür bir norm şu şekilde formüle edilebilir:

Ortak iyilik ilkesi

Süper zekâ sadece tüm insanlığın çıkarı için geliştirilmeli ve geniş kesimlerin paylaştığı etik ideallerin hizmetine koşulmalıdır.[45]

Süper zekânın muazzam potansiyelinin tüm insanlığa ait olmasının daha erken bir aşamada yerli yerine oturtulması bu tür bir normun yerleşmesi için daha fazla zaman sağlayacaktır.

Ortak iyilik ilkesi ilintili alanlarda aktif olan bireyler ya da şirketler için ticari teşvikleri dışlamaz. Örneğin bir şirket, çok yüksek bir sınıra kadar (örneğin yıllık bir trilyon) tüm kârların şirketin hisse sahiplerine ve diğer yasal hak sahiplerine olağan tarzda dağıtılması ve sadece belirlenen çıtayı aşan kârların tüm insanlığa eşit şekilde (ya da diğer türlü, evrensel ahlaki kıstaslara göre) dağıtılması anlamında bir "talih kuşu maddesi"ni benimseyerek süper zekânın yararlarının evrensel olarak paylaşılması çağrısına olumlu yanıt verebilir. Bu tür bir maddenin benimsenmesi epey masrafsız olmalıdır, zira verili bir şirketin bu çok yüksek kâr eşiğini aşma (ve bu tür düşük olasılıklı senaryoların şirketin yöneticilerinin ve yatırımcıların kararlarında hiçbir rol oynamaması) ihtimali son derece düşük olacaktır. Fakat bunun geniş ölçekte benimsenmesi (bağlılıklara güvenilebildiği ölçüde) insanlığa şöyle değerli bir garanti sağlayacaktır: Eğer özel bir teşebbüs zekâ patlamasıyla turnayı gözünden vurursa, herkes kârların çoğundan pay alacaktır. Aynı fikir şirketler dışındaki oluşumlara da uygulanabilir. Örneğin devletlerden birinin gayri safi hâsılası dünya gayri safi hâsılasının çok yüksek bir kısmını (sözgelimi yüzde 90'ını) aştığı takdirde, bu miktarın herkese eşit şekilde dağıtılması konusunda devletler anlaşabilirler.[46]

Ortak iyilik ilkesi (ve talih kuşu maddesi gibi somut örnekleri) ilk başta makine zekâsıyla ilintili alanlarda aktif olan sorumlu bireyler ve örgütler tarafından gönüllü bir ahlaki bağlılık olarak benimsenebilir. Sonrasında, daha geniş ölçekte onaylanabilir ve bir yasa ve anlaşma haline getirilebilir. Burada sunulan türde muğlak bir formülasyon pekâlâ bir kalkış noktası olabilir, ama nihayetinde özgül birtakım doğrulanabilir gereklilikler haline getirilmesi gerekecektir.

ON BEŞİNCİ BÖLÜM

Kırılma Anı

Kendimizi yoğun bir belirsizlik sisiyle çevrili bir stratejik karmaşıklık yumağında bulmuş durumdayız. Birçok husus tespit edilmiş olsa da, ayrıntıları ve karşılıklı ilişkileri hâlâ belirsizdir. Ayrıca henüz ele almadığımız başka etkenler de olabilir. Bu durumda, ne yapmamız gerekiyor?

Miadı olan felsefe

Bir meslektaşım Fields Madalyası'nın (matematikteki en yüksek payenin) o ödüle layık görülen kişi için iki şeyi gösterdiğini söylemeye bayılır: Bir, önemli bir şeyi başarabilecek yetenekte olduğunu; iki, başaramadığını! Sert olmasına sert, ama doğru söze ne denir!

Bir "keşif"i bilginin varışını zamandaki geç bir noktadan daha önceki bir zamana kaydıran bir edim olarak düşünün. Keşfin değeri keşfedilen bilginin değerine değil, daha ziyade bilgiye olabileceğinden daha önce erişilmiş olmasının değerine eşittir. Bir bilimci ya da matematikçi birçoklarının gözünden kaçmış olan bir çözümü herkesten önce bularak büyük bir yetenek sergileyebilir; ama eğer problem zaten çok geçmeden çözülmüş olsaydı, o zaman bu çalışmanın dünyaya muhtemelen pek bir faydası olmazdı. Biraz da olsa daha önce bir çözüme ulaşmanın muazzam değerli olduğu örnekler *vardır*, ama bu en çok da çözümün ya pratik bir amaç için kullanılarak ya da yeni teorik çalışmalara bir temel hizmeti görerek hemen uygulamaya konulduğu durumlarda geçerlidir. İkinci durumda, yani bir çözüm sadece yeni teoriler için bir temel hizmeti görmesi anlamında doğrudan kullanıldığında, ancak önünü açtığı yeni çalışmalar bizatihi hem önemliyse hem de acilse çözüme biraz da olsa daha önce ulaşmanın büyük değeri vardır.[1]

Dolayısıyla mesele Fields Madalyası sahibinin keşfettiği sonucun başlı başına (araçsal olarak ya da sırf bilginin yüzü suyu hürmetine) "önemli" olup olmadığı değildir. Daha ziyade, madalya sahibinin sonucun yayınlanmasının daha önceki bir tarihte gerçekleşmesini sağlamasının önemli olup olmadığıdır. Bu zamansal naklin değeri, dünya çapında bir matematik beyninin başka bir şey üzerinde çalışarak yaratabileceği şeyin değeriyle kıyaslanmalıdır. En azından bazı örneklerde Fields Madalyası yanlış

problemi (örneğin albenisi tam da çözümü çok zor bir problem olmasıyla nam salmasından kaynaklı bir problemi) çözmeye adanmış bir hayata işaret edebilir.

Benzer durumlardan akademik felsefe gibi başka alanlar için de bahsedebiliriz. Varoluşsal riski azaltmayla alakalı bazı problemler felsefenin alanına girer; bu kitapta bunlardan birkaçıyla karşılaştık. Ama felsefenin bazı alt dalları vardır ki, varoluşsal riskle, hatta herhangi bir pratik kaygıyla görünüşte en ufak bir bağı yoktur. Tıpkı saf matematikte olduğu gibi, felsefenin ele aldığı problemlerden bazıları içsel olarak, yani insanların pratikteki uygulamalarından bağımsız olarak önemsemeleri gerektiği anlamında önemli görülebilir. Sözgelimi gerçekliğin temel doğası hakkında bir şeyler bilmek başlı başına kıymetli olabilir. Eğer kimse metafizik, kozmoloji ya da sicim teorisi çalışmasaydı dünya herhalde bugünkü kadar azametli bir yer olmazdı. Gelgelelim ufukta gözüken zekâ patlaması olasılığı bu kadim bilgelik arayışına yeni bir ışık tutuyor.

Şu anki bakış açısı felsefi ilerlemenin doğrudan felsefe yaparak değil, daha ziyade dolaylı yollardan azamiye çıkarılabileceğini gösteriyor. Süper zekânın (hatta azıcık geliştirilmiş insan zekâsının) mevcut düşünürleri geride bırakacağı birçok görevden biri bilim ve felsefedeki temel soruların cevaplanmasıdır. Bu düşünce ertelenmiş doyum stratejisini akla getiriyor. Ebedi sorulardan bazıları üzerine çalışmayı bir süre erteleyebilir, bu görevi daha işinin ehli olacağını umut ettiğimiz haleflerimize bırakabilir ve dikkatimizi daha yakıcı zorluklara odaklayabiliriz: bu sayede, işinin ehli haleflerimiz olma şansını da artırmış oluruz. Bu yüksek etkili felsefe ve yüksek etkili matematik olacaktır.[2]

Ne yapmalı?

Dolayısıyla önemli olmakla kalmayıp, zekâ patlamasından önce çözümlerine ihtiyaç duyulması anlamında yakıcı da olan problemlere odaklanmak istiyoruz. Ayrıca olumsuz değeri olan (yani çözmenin zararlı olacağı) problemlerle ilgilenmemeye dikkat etmemiz gerekiyor. Sözgelimi yapay zekâ alanındaki bazı teknik problemler, çözümleri makine zekâsının gelişimini hızlandıracak ama makine zekâsı devrimini yaşayabilir ve yararlı hale getirebilecek kontrol yöntemlerinin gelişimini süratlendirme noktasında fazla bir şey yapmayacak olduğu ölçüde olumsuz değerde olabilir.

Hem yakıcı hem de önemli olan ve olumlu değere sahip olduğunu güvenle söyleyebileceğimiz sorunları saptamak zor olabilir. Varoluşsal riskin azaltılması konusundaki stratejik belirsizlik iyi niyetli müdahalelerin bile üretken olmamakla kalmayıp zarar verebileceğinden kaygı duyabileceğimiz anlamına geliyor. Aktif derecede zararlı ya da ahlaken yanlış bir şey yapma riskini sınırlamak için güçlü şekilde pozitif değere sahip görünen (yani çözümüyle çok çeşitli senaryolarda olumlu katkı sunacak) problemler üzerinde çalışmayı ve güçlü şekilde gerekçelendirilebilir olan (yani çok çeşitli ahlaki bakış açılarından kabul edilebilir olan) araçlardan yararlanmayı tercih etmemiz gerekir.

Hangi problemlere incelik vermemiz gerektiğini değerlendirirken bir kıstası daha dikkate almak zorundayız. Çözüm arama çabalarımız karşısında *esnek* olan problemlerle uğraşmalıyız. Son derece esnek problemler, bir tık daha fazla çaba harcayarak çok daha hızlı çözülebilecek ya da çok daha geniş ölçekte çözülebilecek problemlerdir. Dünyada daha fazla nezaketi teşvik etmek önemli ve yakıcı bir problemdir; dahası fevkalade sağlam bir olumlu değere sahiptir; ama bunun nasıl başarılacağına dair yeni bir fikir olmadığından, muhtemelen esnekliği çok düşük bir problem olarak kalacaktır. Dünya barışının sağlanması çok istenen bir şeydir; ama bu problemi şimdiden hedef alan sayısız çaba ve çabuk bir çözüme ulaşmanın önündeki sıkı engeller düşünüldüğünde, fazladan birkaç bireyin katkılarının büyük bir fark yaratması muhtemel görünmüyor.

Makine zekâsı devriminin risklerini azaltmak için, bu gereklilikleri en iyi karşıladığı gözüken iki hedef önereceğiz: stratejik analiz ve kapasite inşası. Bu parametrelerin göstergesi konusunda görece güvenli olabiliriz: Daha stratejik içgörü ve daha fazla kapasite daha iyidir. Ayrıca parametreler esnektir: Fazladan birazcık yatırım görece büyük bir fark yaratabilir. İçgörü ve kapasite elde etmek bu parametrelerde erkenden ilerleme kaydetmenin sonraki çabaları daha etkili kılabilecek olmasıdır. Bu iki genel hedefe ilaveten, inisiyatif noktasında potansiyel açıdan kayda değer birkaç amaca daha işaret edeceğiz.

Stratejik ışık arayışı

Kafa karışıklığı ve belirsizliklerle dolu bu ortamda analizin beklenen değerinin bilhassa yüksek olduğu göze çarpıyor.[3] Stratejik durumumuzun aydınlatılması sonraki müdahaleleri daha etkili şekilde hedef almamıza yardımcı olacaktır. Tali bir meselenin salt bir ayrıntısı konusunda değil, merkezi şeylerin temel nitelikleri konusunda da ne yapacağımızı bilmediğimizde stratejik analize özellikle ihtiyaç duyarız. Birçok kilit parametre açısından, bunların *göstergesi* konusunda bile son derece belirsiziz, yani hangi değişim yönünü isteyip hangisini istemediğimizi bile bilmiyoruz. Cehaletimiz umarsız olabilir. Bu alanda çok az araştırma yapılmıştır ve parlak stratejik içgörüler toprağın hemen altında gün yüzüne çıkarılmayı bekliyor olabilir.

Burada "stratejik analiz" derken kastettiğimiz *hayati etmenler* arayışıdır: Görüşlerimizi sadece tatbikinin ince yapısı hakkında değil, aynı zamanda istenilirliğin genel topolojisi konusunda da değiştirme potansiyeli olan fikirler ya da argümanlar arayışıdır.[4] Hayati etmenlerden bir tekinin bile kaçırılması en cüretkâr çabalarımızı berbat hale getirebilir ya da yanlış tarafta savaş veren bir askerinki kadar aktif derecede zararlı kılabilir. Hayati etmenler arayışı (tanımlayıcı meseleler kadar normatif meseleleri de araştırmayı gerektirir) farklı akademik disiplinler ile diğer bilgi alanları arasındaki sınırların sık sık iki tarafına da gidip gelmeyi gerektirir. Bu tür bir araştırmanın nasıl yürütüleceğine dair yerleşik bir metodoloji olmadığından, zor ve özgün düşünceler şarttır.

İyi kapasite inşası

Yüksek değere sahip bir diğer faaliyet (stratejik analiz gibi çok çeşitli senaryolarda yararlı olma gibi sağlam bir özelliği olan faaliyet), geleceği ciddiye alan sıkı bir destek tabanının geliştirilmesidir. Bu tür bir taban araştırma ve analiz için hemen kaynak sağlayabilir. Diğer öncelikler görünür hale gelirse ve geldiğinde, kaynaklar da buna göre yeniden dağıtılabilir. Dolayısıyla bir destek tabanı, ortaya çıkan yeni içgörülerin kılavuzluk edebileceği bir genel amaçlı yeterliliktir.

Değerli özelliklerden biri akli insancıllığa kendini adamış, varoluşsal risk konusunda bilgili ve bunları azaltma araçları konusunda içgörüye sahip bireylerden müteşekkil bir verici ağı olacaktır. İlk başlardaki fonlayıcıların diğerkâm ve ferasetli olması özellikle önemlidir, zira alışıldık maddi çıkarlar devreye girip kök salmadan önce alanın kültürünü şekillendirmek için fırsatlara sahip olabilirler. Dolayısıyla bu açılış hamleleri sırasında odak noktasında doğru türde insanları alana devşirmek olmalıdır. Safları güvenliğe gerçekten önem veren ve yönelimleri doğrunun aranması olan (ve kendileri gibi insanları çekme ihtimalleri yüksek olan) bireylerle doldurmak için kısa vadede bazı teknik ilerlemelerden vazgeçmeye değer.

Önemli değişkenlerden biri yapay zekâ alanı ile onun önde gelen projelerinin "toplumsal epistemolojisi"nin niteliğidir. Hayati etmenleri keşfetmek değerlidir, ama ancak eylemi etkilediği takdirde. Bunu her zaman ön kabul olarak alamayız. Bir prototip geliştirmek için milyonlarca dolar harcayan ve yıllarca emek veren bir proje düşünün. Sistem birçok teknik zorluğu aştıktan sonra nihayet gerçek bir ilerleme kaydediyor olsun. Biraz daha çalışmayla yararlı ve kârlı bir şeye dönüşme ihtimali vardır. Ama sonra, tamamen farklı bir yaklaşımın biraz daha güvenli olacağını gösteren hayati bir etmen keşfediliyor. Proje onuru kırılmış bir samuray gibi kendini öldürmeli, güvenlik açığı olan tasarımından ve kaydedilen tüm ilerlemelerden vazgeçmelidir denebilir mi? Yoksa kaygılı bir ahtapot gibi davranıp saldırıdan kaçmak adına etrafa güdülenmiş şüphecilik mi yaymalıdır? Bu tür bir ikilem karşısında samuray seçeneğine yönelecеğine güvenilen bir proje çok daha tercih edilir bir yapay zekâ yaratıcısı olacaktır.[5] Ama belirsiz iddialar ve spekülatif muhakemeye dayalı *seppuku* yapmaya hevesli süreçler ve kurumlar inşa etmek kolay değildir. Toplumsal epistemolojinin bir diğer boyutu da hassas bilgilerin idaresidir, özellikle de gizli tutulması gereken bilgilerin sızdırılmasından kaçınabilmektir. (Bilgi gizliliği akademik araştırmacılar için özellikle zor olabilir, zira sonuçlarını buldukları her kanaldan sürekli duyurmaya alışkındırlar.)

Somut önlemler

Stratejik ışık ve iyi kapasite gibi genel hedeflerin yanı sıra, eylem için maliyeti düşürecek bazı özgül hedeflerden de bahsedilebilir.

Bunlardan biri makine zekâsı güvenliğinin teknik zorlukları konusundaki ilerlemedir. Bu hedefin peşine düşerken bilgi tehlikeleriyle başa çıkmaya dikkat edilmelidir. Kontrol problemini çözmek için yararlı olabilecek bir çalışma yeterlilik probleminin çözülmesinde de yararlı olabilir. Yapay zekâ fünyesini patlatan çalışmanın net değeri pekâlâ olumsuz olabilir.

Bir diğer özgül hedef, yapay zekâ araştırmacıları arasında "en iyi pratikler"i teşvik etmektir. Kontrol probleminde kaydedilen her türlü ilerlemenin geniş kesimlere dağıtılması gerekir. Bazı hesaplamalı deney biçimleri, özellikle de güçlü yinelemeli kendini geliştirme içeriyorsa, tesadüfi bir havalanma riskini azaltmak için yeterlilik kontrolünden yararlanmayı da gerektirebilir. Güvenlik yöntemlerinin gerçekteki uygulanması bugün o kadar ilintili olmamakla beraber, teknoloji ilerledikçe bu kimliğe kavuşacaktır. Dolayısıyla pratisyenlerin *güvenliğe bağlılık* beyanında bulunmaları (bir olasılık olarak makine zekâsı daha yakın görünmeye başlarsa ve başladığında güvenliğin artırılacağının vaat edilmesi ve ortak iyilik ilkesinin onaylanması dahil) çağrısında bulunmak için hiç de erken değildir. Güzel temenniler yeterli değildir ve tehlikeli bir teknolojiyi güvenli hale getirmeye yetmez: ama ağız nereye giderse, akıl da peyderpey peşinden gidebilir.

Hayati bir parametreyi çabuklaştırmak için, örneğin başka bir varoluşsal riski azaltmak ya da biyolojik bilişsel gelişimimizi ve kolektif zekâmızın ilerletilmesini teşvik etmek, hatta dünya siyasetini daha uyumlu bir eksene kaydırmak için ara ara başka fırsatlar da doğabilir.

İnsanın içindeki en iyi, lütfen ayağa kalksın!

Bir zekâ patlaması olasılığı karşısında biz insanlar bir bombayla oyun oynayan küçük çocuklar gibiyiz. Oyuncağımız ile davranışımızın toyluğu arasında böyle bir uyumsuzluk vardır. Süper zekâ bizim şu an hazır olmadığımız ve uzun bir süre de olmayacağımız bir zorluktur. İnfilakın ne zaman gerçekleşeceğine dair çok az fikrimiz var, ama eğer aleti kulağımıza tutarsak belli belirsiz bir tık tık sesi duyabiliriz.

Elinde pimi çekilmemiş bir bomba olan bir çocuk için yapılacak olan mantıklı şey bunu yavaşça yere bırakmak, çabucak odadan çıkmak ve en yakındaki bir büyüğüyle temas kurmak olacaktır. Ama burada bir değil, birçok çocuk vardır ve hepsinin de tetik mekanizmasına kendi erişimi vardır. Hepimizin bu tehlikeli zımbırtıyı yere koyacak basireti gösterip göstermeyeceği şansa kalmıştır. Salağın teki çıkıp sırf ne olacağını merak ettiği için pimi çekebilir.

Keza kaçarak da güvenliğimizi sağlayamayız, zira bir zekâ patlamasının tahribi tüm gökkubbeyi yere indirir. Ortalıkta bir yetişkin de yok ki!

Bu durumda, eteklerimizin zil çalması alakasız kaçacaktır. Afallama ve korku daha uygundur; ama sanırım en uygun tutum tıpkı hayallerimizi gerçeğe dönüştürecek ya da yıkacak zor bir sınava hazırlanıyormuşçasına olabildiğince ehil hale gelmek için sıkı bir kararlılık göstermektir.

Bu bir fanatizm reçetesi değildir. Zekâ patlamasının gerçekleşmesi hâlâ daha onyıllar alabilir. Ayrıca şu an karşılaştığımız zorluk, kısmen, insanlığımıza tutunmaktır: Bu en gayri tabii ve insanlık dışı problem karşısında ayaklarımızı yere basmaya, aklıselimimizi ve neşeli ılımlılığımızı korumaya devam etmeliyiz.

Ama küresel açıdan kayda değer olanın izini kaybetmememiz gerekir. Günlük meşgalelerin hayhuyunda çağımızın temel görevini belli belirsiz de olsa algılayabiliriz. Bu kitapta diğer açılardan hâlâ görece amorf ve olumsuz tanımlanmış bir vizyon –temel ahlaki önceliğimizin (en azından gayri şahsi ve seküler bir perspektiften) varoluşsal riskin azaltılması ve insanlığın kozmik birikiminden müşfik ve coşkun şekilde yararlanmaya varan bir medeniyet güzergâhına ulaşılması olduğunu söyleyen bir vizyon– hakkında birazcık daha açıklayıcı olmaya çalıştık.

Sonsöz

Bu kitabın ciltli edisyonunun yayımlanmasından bu yana bazı tutum değişiklikleri oldu. Süper zekâya aptalca olmayan bir konu olarak yaklaşmak, bir makine zekâsı geçişinin bu yüzyılda gerçekleşebileceği, bu tür bir geçişin insanlık tarihindeki en önemli olaylardan biri olabileceği, muazzam getirilerinin yanı sıra belli oranda bir varoluşsal risk de barındırdığı ve olumlu bir sonuca giden yolu kısaltmak için yapmamız gereken bir şey olup olmadığını görmek için öncesinde birazcık çalışma yapmanın akıllıca olduğu görüşünü ciddiye almak daha kolay hale geldi. Bunu kabul ettikten sonra, yine de basının konuya ilgi gösterdiği her durumda alay etmek için hazır bekleyen bir Terminatör'le boğuşmak gerekiyor. Ama popüler kakofoniden uzaklaşınca, daha olgun bir konuşmanın kısık sesini duymak da mümkündür (eğer kulak verirseniz).

Makine öğreniminde teknik ilerleme birçoklarının beklediğinden daha hızlı olmuştur. Son gelişmelerle irdelenecek bir dolu fikir ortaya çıkmıştır: Nöral Turing makineleri, derin takviyeli öğrenme, Bayesçi hiperparametre optimizasyonu, grid ağı, LSTM'ler, bellek ağları, değişimli otokodlayıcılar, vektör donanımları, çekişmeli üretici ağlar, dikkat temelli üretici modeller, olasılıkçı programlamaya çeşitli yaklaşımlar bugünkü vızıltının kaynaklarından sadece birkaçıdır. Heyecanın odağında esasen derin öğrenme var. Derin öğrenme yöntemleri (esasen çok katmanlı nöral ağlar) hızlı bilgisayarların, büyük veri kümelerinin ve algoritmik iyileştirmelerin bir araya gelmesi sayesinde birçok algısal görevde insan performansına yaklaşmaya (ve bazı örneklerde geçmeye) başlamıştır; bunların arasında el yazısının tanınması, görüntü tanıma ve görüntü altyazılama, konuşma tanıma ve yüz tanıma da vardır. Derin öğrenme yöntemleri doğal dil tercümesinde ve bazı bilimsel veri analizi problemlerinde de güçlü sonuçlar elde etmiştir. Bu performansların altında yatan yeterlilik –soyut dağıtılmış temsilleri ham duyu verilerinden öğrenebilen ve bir insanın özellikleri ya da özgül alan bilgisini elle halletmesine ihtiyaç duymayan genel algoritmalar– daha karmaşık yeterliliklerin inşasında bir yapıtaşı olabilir.

Mevcut yeterliliklerden birçoğu yararlılık eşiğini aşmıştır. Bu demektir ki yeni ilerlemeler hemen semeresini verecektir, zira doğrudan ticari olarak değerli ürünlerde ilerlemelere dönüşecektir. Eğer bir yararı olmayacak kadar hatalı olan bir konuşma tanıma sistemiyle yola çıkarsanız ve çok az ilerleme kaydederseniz, elinizde hâlâ –ne olacaktı?– yararsız bir sistem olacaktır. Ama geniş ölçekte kullanılacak kadar iyi olan

bir sistemle başlarsanız, kaydedeceğiniz yüzde birlik ilerleme milyar dolarlara denk olabilir. Makine öğrenmenin şu an ilerleyişte olduğuna, daha keşfedeceği birçok yol olduğuna dair heyecan verici his, şu an ticari teşviklerle de birleştiğinde para ve yeteneğin bu alana akmasına yol açmaktadır.

Mevcut heyecan dalgasının ne kadar süreceğini kimse bilmiyor. Bu kitabın argümanları arasında bir zekâ patlamasının eli kulağında olduğu ya da dünyanın genelinin ilerleme hızını hafife aldığı yoktur. Ölçek büyüterek ve ince ayar yaparak elde edilecek ilerlemelerin yakında durağan bir noktaya gelmesi pekâlâ mümkündür. Kuşkusuz, yolun geri kalanını katetmek için atılım sağlayacak yeni birtakım fikirlere ihtiyaç olacaktır; bu fikirler zamanında gelebilir de gelmeyebilir de. Ama ben alanın geçmişte yaşadığı kadar sert bir "yapay zekâ" kışının tekrarlanması beni şaşırtır. Bu noktadan itibaren fon aktarımının devam etmesi ve saygınlığın artması daha muhtemeldir, zira yapay zekâyı hem değerli bir teorik çalışma alanı olarak hem de yararlı bir mühendislik disiplini olarak vazetmek için halihazırda yeterince şey yapılmıştır. (Bu tahmin elbette şu ya da bu yatırım balonunun patlayabileceğini ve birçok tekil projenin beklentileri karşılamayabileceğini dışlamaz.)

Olumlu bir sonuç olasılığının nasıl artırılacağına ilişkin çalışmalarda da bir ilerleme kaydedilmiştir. Teorik alanda, daha fazla irdelenmeyi hak eden bazı yeni fikirler (örneğin Paul Christiano'nun "onay yönlendirmeli aracılar"ı) serdedilmiştir. Örgütsel alanda da durum biraz ilerlemiştir; ilerleme görece çarpıcıdır, ama çok düşük bir temel seviyeden başlamaktadır. Kontrol probleminin meşru bir araştırma konusu olarak tanınması ilintili araştırma topluluklarının daha küçük bir kesimine yayılmıştır ve makine zekâsının gelecekteki etkilerinin daha sistemli bir dikkati hak ettiği yaygın bir kanıdır

Yapay zekâ güvenliği ve etki alanları konusundaki bu ilerleme, böyle kısa bir zamanda gerçekleştiği için kayda değer olsa da, abartılmamalıdır. Evet, alana daha fazla fon akıyor; ama makineleri salt daha akıllı hale getirmek için ayrılandan iki üç kat daha azdır. Evet, makine zekâsındaki ilerlemelerin sonuçlarına daha fazla kafa yoruluyor, ama bunun çoğu nihayetinde ölümcül özerk silahlar, otomasyonun emek pazarına etkileri, siber suçlar, mahremiyet ya da sürücüsüz arabalar gibi daha kısa vadeli ilgilere kaymaktadır. Bunlar bazı insanlar için anlamsız şeyler değildir, ama insan düzeyinde yapay zekâ ya da süper zekânın gündeme getirdiklerinden çoğunlukla epey farklı meselelerle alakalıdır.

Ve evet, yapay zekâ araştırma camiasının içinde uzun vadeli meselelerin belli oranda meşru hale geldiğini görüyoruz. Ama bu kabulün hâlâ kısmi olduğunu ve kırılgan da olabileceğini görmek gerekiyor. Bazı yapay zekâ araştırmacıları halk arasında konuşulanların kontrolden çıktığından endişelenmeye başladılar. Şu aptal Terminatör resimleri sürece zarar veriyor. İnsanın akademik disiplinine, meslek camiasına, haya-

tını adadığı işe mesnetsiz suçlamalar yöneltilmesi pek hoş olmasa gerek. Kötü robot orduları konusunda halk arasında çalınan yanıltıcı alarm zillerine tepki olarak, yapay zekâ camiasının ileri makine zekâsına ilişkin her türlü gerçek kaygıyı karalayan bir görüş etrafında saflarını sıklaştırması mümkündür. Bu tür bir yanıt medyadaki korku tellallığını somut olarak hedef aldığı ölçüde hayırlı bir amaca hizmet edebilir. Ama munzam zarar korkusunu da duymak gerekir. Yapay zekâ araştırmacıları arasında eleştirmenlerin, korku tellallarının, çatlakların ya da düzenleyici bozuntularının eline koz vereceği korkusuyla yapay zekâdan bahsetmenin ya da olanaklı risklerini araştırmanın yakışıksız olduğuna dair bir norm ortaya çıkarsa, son dönemde meşrulaşma bakımından elde edilen kazanımlar çabucak tersine dönebilir. Hayırhah süper zekâ geliştirilmesine giden en iyi yol yapay zekâ geliştiricilerinin ve yapay zekâ güvenlik araştırmacılarının aynı tarafta oldukları, gerçekten de önemli bir ölçüde aynı kişiler oldukları bir yoldur. Dolayısıyla tüm taraflara sabır, itidal ve zihin açıklığı çağrısı yapıyorum ve mümkün olan her durumda doğrudan diyaloğa ve işbirliğine geçmelerini teklif ediyorum.

Kitap yayımlandığından bu yana yapılan yorumlara burada bir yanıt vermek istemiyorum. Sadece ufak bir not düşeceğim, o da bu karton kapak baskısının yeni sahiplerine ve esasen de aldıkları kitabı gerçekten *okumaya* zaman bulamayacak kadar meşgul olan ve sadece kitabın içindekiler kısmına ve önüne arkasına bakabilenlere: Bir meselenin içkin önemi haricinde, açımlanmasına kaç sayfa ayrıldığını etkileyebilecek birçok etken vardır. Dolayısıyla sadece kaç sayfa olduğunu sayarak ne savunduğumu çıkartmak mümkün değildir. Sözgelimi ben potansiyel faydalardan çok risklere odaklanıyorum. Ama bu demek değildir ki muazzam faydaları olduğunu düşünmüyorum; sadece, tarihin bu noktasında eğer makine zekâsına geçiş iyi giderse (astronomik düzeyde) müthiş şeyler umut etmekle yetinebiliriz, ama somut olarak nelerin yanlış gidebileceğine dair kesin ve ayrıntılı bir anlayışa ulaşmanın daha yakıcı olduğunu düşünüyorum, zira bu sayede söz konusu yanlışlardan sakınabiliriz. Aynı şekilde, tek bir süper zeki yapay zekânın geleceği kendi tercihlerine göre şekillendirecek kadar güçlü hale geldiği senaryoları sayfalarca analiz ediyorum; ama buradan çokkutuplu senaryoları reddettiğim sonucu çıkmaz (örneğin bkz. 11. Bölüm). Kontrol probleminin ne kadar zorlu olabileceğini ve yüzeysel açıdan akla yatkın çözümlerin nasıl çuvallayabileceğini uzun uzadıya ele alıyorum; ama kontrol probleminin aynı derecede kolay hallolması da mümkündür diyorum.

Bu kitabın yaratılmasında veya yayılmasında emeği geçen herkese ve aynı zamanda insanlığın bu tuhaf durumunda yapıcı bir rol oynamaya çalışan herkese bir kez daha teşekkür etmek istiyorum.

Nick Bostrom, Kasım 2015

Sözlükçe

Adım riski: Bir geçiş yapmaktan kaynaklanan risk. Burada risk miktarı geçişin ne kadar sürdüğünün basit bir işlevi değildir.

Ahlaki doğruluk: Ahlaken doğru olanı yapmaya çalışan bir yapay zekâ

Alet yapay zekâ: "Bir aracı gibi" olmayan, daha ziyade günümüz yazılımının daha esnek ve ehil bir versiyonunu andıran bir yapay zekâ.

Antropik zapturapt: Bir yapay zekânın bir simülasyonda olabileceğini düşündüğü ve dolayısıyla simülatörleri tarafından ödüllendirilecek şekillerde davranmaya çalıştığı bir varsayımsal olgu.

Araçsal yakınlaşma tezi: "Yakınsak araçsal değerleri", yani çok çeşitli olanaklı çevrelerde mümkün olan çeşitli nihai hedeflerin elde edilmesi için yararlı olan alt hedefleri, dolayısıyla geniş bir zeki aracı sınıfı tarafından peşine düşülmesi muhtemelen olan alt hedefleri saptayabiliriz.

Belirleyici stratejik üstünlük: Bir aracının tüm dünyada (teknolojiyle ya da başka araçlarla) egemenlik kurmasını sağlayacak boyuttaki stratejik üstünlük.

Bilge-tekel sürdürülebilirlik eşiği: Bir yeterlilik seti, bilge-tekel eşiğini ancak ve ancak bu yeterliliklere sahip olan sabırlı ve varoluşsal bir risk bilinçli sistem –hiçbir zeki rakiple karşılaşmadığı takdirde– erişilebilir evrenin büyük bir kısmına yerleşip burayı yeniden dizayn ederse aşacaktır.

Bilişsel geliştirme: Bir sistemin entelektüel yeteneklerinin ilerletilmesi.

Birinci asil-vekil problemi: Bir insanın ("asil") başka birini ("vekil") ilkinin çıkarlarına göre hareket etmesi için atadığında karşılaşılan bilindik problem (örneğin işçi-işveren ilişkisinde görülür).

Cin: Yüksek düzey bir komutu yerine getiren ve ardından diğerini bekleyen bir yapay zekâ.

Çağrışımsal değer ekleme: Yapay zekânın zaman içinde ve deneyimlerine bağlı bir tarzda yeni nihai değerler benimsemesine yol açan bir mekanizmanın saptandığı bir değer yükleme yaklaşımı.

Çokkutuplu sonuç: Makine zekâsı çağına geçişten sonra birbiriyle yarışan çoklu süper zeki aracıların olduğu bir durum.

Değer öğrenme: Değer yükleme problemine bu yaklaşımda, yapay zekân insanların ondan peşine düşmesini bekledikleri değerleri öğrenir.

Değer yükleme problemi: Yapay zekânın bizim istediğimiz değerlerin peşine nihai hedefler olarak düşmesine sebep olma problemi.

Değer yüklemeye evrimsel seçilim yaklaşımı: Bir yinelemeli seçilim süreci yoluyla istenen değerlere sahip bir yapay zekâ elde etmeye çalışan bir değer yükleme yaklaşımı; insanları doğuran evrimsel seçilim süreciyle benzeştir.

Diferansiyel teknolojik gelişme ilkesi: Tehlikeli ve zararlı teknolojilerin, özellikle de varoluşsal risk düzeyini yükselten teknolojilerin gelişimini geciktirin ve yararlı teknolojilerin,

özellikle de doğanın ya da diğer teknolojilerin teşkil ettiği varoluşsal riskleri azaltan teknolojilerin gelişimini hızlandırın.

Diklik tezi: Zekâ ve nihai hedefler diktir: Az çok herhangi bir düzeydeki zekâ ilkesel olarak az çok herhangi bir nihai hedefle birleştirilebilir.

Doğrudan saptama: Programlayıcıların insanların nelere değer verdiklerini saptadıkları ve ardından yapay zekâ için buna denk düşen değerleri ya da kuralları açıkça içeren bir kod yazdıkları kontrol problemi yaklaşımı.

Dolaylı normatiflik: Bir yapay zekânın hangi değerleri teşvik etmesi gerektiğini doğrudan bizzat belirlemektense yapay zekânın sadece zımnen tanımlanmış normatif bir standardın somut içeriğini keşfetmek için kendi entelektüel kaynaklarından yararlanarak takip edebileceği bir kıstas ya da yöntem saptarız.

Durum riski: Belli bir durumda olmaktan kaynaklanan risk, şöyle ki risk miktarı orada harcanan zamanın doğrudan bir işlevidir.

Emülasyon modülasyonu: Az çok normal insan güdüleri olan beyin emülasyonlarıyla başlayıp dijital ilaç analogları ya da başka araçlar kullanarak onların güdülerini değiştirmek.

Epistemik farklılık ilkesi: Gelecekteki bir süper zekâ epistemik açıdan üstün bir konum işgal eder: İnançlarının (muhtemelen birçok konuda) doğru olma ihtimali bizimkinden daha yüksektir. Bu yüzden mümkün olan her noktada süper zekânın kanaatine saygı göstermeliyiz.

Evcimenlik: Yapay zekânın güdülenme sisteminin yapay zekâya çok mütevazı hevesler verecek şekilde tasarlandığı ve müdahale etmek istediği şeylerin kapsamının sınırlandırıldığı kontrol problemi yaklaşımı.

Gayri şahsi perspektif: İnsanın herkesin çıkarlarını en fazla gözetecek şekilde hareket etmesi gerektiği görüşü (Karş. "İnsanı etkileyen perspektif").

Havalanma: Sadece insan düzeyinde makine zekâsının olduğu bir koşuldan radikal süper zekânın olduğu bir koşula geçiş. Esas etkeni çoğunlukla hızdır.

Hayati etmen: Pratik çabalarımızda sadece küçük bir yol değişikliğine değil aynı zamanda büyük bir öncelik değişikliğine de (örneğin önemli müdahalelerin arzulanırlık göstergesini tersine çevirerek) ihtiyaç olduğunu akla yatkın şekilde ortaya serebilen bir fikir ya da argüman.

Hızlı süper zekâ: Bir insan zekâsının yapabileceği her şeyi yapan ama daha hızlı yapan bir sistem.

Hükümdar: Büyük ve muhtemelen çok uzun erimli hedefler peşinde dünyada özerk şekilde hareket eden bir süper zekâ.

İkinci asil-vekil problemi: Kendi adına çalışacak bir süper zeki yapay zekâ ("vekil" ya da "aracı") tasarlamak isteyen bir insan "asil"in karşılaştığı asil-vekil problemi. "Kontrol problemi" diye de geçmektedir.

İnsan düzeyinde donanım: İnsan beyninin bilgi işleme yeteneğine ulaşan donanım.

İnsan düzeyinde yapay zekâ: Pratikte önem arz eden tüm alanlarda tipik bir insanın entelektüel performansını yakalayabilen bir yapay zekâ. (Bu terim ciddi müphemlikler içerir.)

İnsan düzeyinde yazılım: İnsan beyninin yaptığı görevleri yapması için insan beyninin algoritmik verimliliğine erişen yazılım.

İnsanlığın kozmik birikimi: Evrende teknolojik açıdan olgun dünya kökenli bir medeniyet tarafından erişilebilir olan fiziksel kaynak rezervi.

Kâhin: Sadece soruları cevaplayan bir yapay zekâ.

Kişiyi etkileyen perspektif: Kişinin tercihlerinden bağımsız olarak zaten var olan veya var olacak olan herkesin çıkarlarına en iyi uyacak şekilde hareket edilmesi gerektiği görüşü (karş. "Gayri şahsi perspektif").

Kolektif süper zekâ: Sistemin çok genel birçok alandaki genel performansının mevcut bilişsel sistemlerin hepsini misli misli geride bırakacağı şekilde birçok küçük zekâdan oluşan bir sistem.

Kontrol problemine takviyeli öğrenme yaklaşımı: Yapay zekânın bir kümülatif ödül fikrini azamiye çekmeyi öğrendiği bir yaklaşım (burada ödül sinyali yapay zekâ cephesinde uygun davranışı tetiklemesi amaçlanan bir tarzda saptanır ya da yönetilir).

Kutulama: Yapay zekânın dış dünyayla etkileşimde bulunma yeteneğini azaltmak için çevresini tasarlamaktan oluşan bir kontrol yöntemi. Örneğin yapay zekâyı dikkatle yalıtılmış bir sanal gerçeklik simülasyonunda ve insan denetleyicilerine sadece sınırlı bir iletişim kanalı sunacak şekilde çalıştırmak.

Makroyapısal gelişim hızlandırıcısı: Makroyapısal özelliklerin (teknoloji ve jeopolitik dinamikler gibi) gelişim hızını değiştirecek olan ama mikro düzeyde insan ilişkilerinin gelişim hızını aynı bırakan hayali bir düğme (düşünce deneylerinde kullanılır).

Müşterek iyilik ilkesi: Süper zekâ sadece tüm insanlığın yararı için ve geniş kesimlerin paylaştığı etik ideallerin hizmetine koşularak geliştirilmelidir.

Nitelikli süper zekâ: En az bir insan zihni kadar hızlı ve niteliksel açıdan daha akıllı bir sistem.

Optimizasyon gücü: Bir sistemin zekâsını geliştirmek için uygulanan kalitece ayarlanmış tasarım çabası miktarı.

Özendirici ambalajlama: Yapay zekânın gerçekleştirilmesine yardımcı olanlara fazladan ödül veya nüfuz sağlayan hedef hükümleri

Süper zekâ: Neredeyse her alanda insanların bilişsel performansını misliyle aşan bir zekâ.

Tekel: En üst karar alma düzeyinde sadece tek bir etkili aracının olduğu birt dünya düzeni (ama içeride birçok fraksiyon ve çıkar odağı olabilir); başka bir deyişle, tüm mnemli küresel eşgüdüm problemlerinin büyük oranda çözüldüğü bir koşul. Mümkün olan örnekler arasında bir dünya demokrasisi, rakipsiz bir dünya diktatörü ya da her türlü potansiyel rakibini bastıracak kadar güçlü bir süper zeki yapay zekâ sayılabilir.

Teknoloji bağlaşımı: İki teknoloji arasında kestirilebilir bir zaman ilişkisidir. Şöyle ki, teknolojilerden birini geliştirmenin ya zorunlu bir öncül olarak ya da bariz ve karşı konulmaz bir uygulama ya da sonraki adım olarak diğerinin gelişimine yol açması güçlü bir ihtimaldir.

Tersten örnekleme: Hedef kıstaslarını tanımlamış olan programlayıcıların niyetlerini ihlal eden bir yapay zekânın hedefini tutturmanın etkili bir yolu (örneğin "insanları gülümsetin" hedefini insanın yüz kaslarını sürekli sırıtacak şekilde felç ederek tutturmak).

Teşvik yöntemleri: Yapay zekânın çevresini, nihai hedefleri insani değerlerle örtüşmese bile yapay zekânın araçsal nedenleri olacak şekilde kurmaktan oluşan yapay zekâyı kontrol etme sistemleri.

Tohum yapay zekâ: Kendi mimarisini geliştirerek etkileyici bir yapay zekâya önyükleyebilecek olan, başlangıçta mütevazı yeterliliklere sahip olan bir yapay zekâ.

Tüm beyin emülasyonu: İnsan beyninin hesaplamalı yapısını kopyalayarak yaratılan makine zekâsı.

Uzlaşmazlık: Bir sistemin ilerlemekte ne kadar zorlandığı.

Yeterlilik kontrol yöntemleri: Yapay zekânın yapabileceklerini sınırlayarak istenmeyen sonuçlardan kaçınma stratejileri.

Yinelemeli kendini ilerletme: Bir yapay zekânın (belki de bir tohum yapay zekânın) kendi zekâsını tekrarlayarak ilerlettiği süreç; artan zekâsını giderek sağlamlaşan optimizasyon gücünü bilişsel kendini geliştirme görevine uygulamak için kullanır.

Zekâ patlaması: Bir yapay zekânın "görece mütevazı" zekâ düzeyinden radikal derecede insanüstü düzeye hızla ilerlediği bir farazi olay (bu sürecin çoğu zaman yinelemeli kendini ilerletme içerdiği düşünülür).

Zenginleştirme: İstenen güdülere sahip bir süper zekâ edinme yaklaşımlarından biri. Sıfırdan bir hedef sistemi yaratmaya çalışmak yerine, halihazırda uygun güdüleri olan (örneğin bir insan) bir sistemle başlanır ve o daha akıllı hale getirilmeye çalışılır.

Zihin suçu: Ahlakla bağlantılı bilgi işlem süreçlerine yanlış davranma (simülasyonlarda ya da bir yapay zekânın içinde veya araçsal nedenlerle makine substratında yaratılır).

Bu sözlükçe okura kolaylık olsun diye hazırlanmıştır, ama ne tamdır (kitaptaki birçok terim ve kavram buraya alınmamıştır) ne de esas belirleyicidir (bir kavramın nasıl anlaşılması gerektiği açık olmadığında metindeki değerlendirmeler önceliklidir). Bu sözlükçenin ilk versiyonunu hazırladığı ve değişiklikler yapmama izin verdiği için Stephanie Zolayvar ve Katja Grace'e (AI Impacts, Machine Intelligence Research Institute) minnettarım. Özgün versiyonu için, bkz. http://aiimpacts.org/ai-risk-terminology/

Notlar

ÖNSÖZ

1 Gelgelelim tüm sonnotlardaki bilgi yararlı değildir.

2 Hangileri olduğunu bilmiyorum.

BİRİNCİ BÖLÜM

1 Bu türde çarpıcı bir büyüme ve hızlanma John von Neumann'ın matematikçi Stanislaw Ulam'la bir sohbetinde ima edildiği üzere, "tekillik"in nasıl gelebileceğine dair mefhumlardan birini sunabilir:

> Konuşmamız insanların yaşam tarzındaki değişiklikler ve sürekli hızlanan teknolojik ilerlemeye odaklanmıştı; bu gelişmeler bize şu anda yarışın tarihinde asli bir tekilliğe yaklaşıldığı ve bunun ötesinde bildiğimiz haliyle insan ilişkilerinin devam edemeyeceği görüntüsü sunuyor (Ulam 1958).

2 Hanson (2000).

3 Vinge (1993); Kurzweil (2005).

4 Sandberg (2010).

5 1960'larda sık sık tekrarlanan iki iyimser ifade: "Makineler yirmi yıla kalmaz bir insanın yapabileceği her türlü işi yapabilir hale gelecekler" (Simon 1965, 96); "Bir kuşağa kalmaz [...] yapay zekâ yaratma problemi ciddi düzeyde çözülmüş olacaktır" (Minsky 1967, 2). Yapay zekâ tahminlerine dair sistemli bir değerlendirme için, bkz. Armstrong ve Sotala (2012).

6 Örneğin bkz. Baum vd. (2011) ile Armstrong ve Sotala (2012).

7 Gelgelelim bu bize yapay zekâ araştırmacılarının gelişim zaman çizelgeleri hakkında düşündüklerinden daha az şey bildiklerini söylüyor olabilir, ama bu hem lehte hem de aleyhte olabilir: Yapay zekâya varılacak zamanı olduğundan hem daha çok hem de daha az görebilir.

8 Good (1965, 33).

9 İstisnalardan biri olanakçı sonuçlar hakkında bazı şüpheleri olan Norbert Wiener'dir. 1960'ta şöyle yazıyordu: "Eğer amaçlarımıza ulaşmak için mekanik bir aracıdan yararlanırsak ama hareketi tamamlanmadan önce müdahale edebilmemiz için gerekli verilerden mahrum olduğumuz için işleyişine etkili şekilde müdahale edemeyeceğimiz kadar hızlı ve geri döndürülemez şekilde hareket ediyorsa; o zaman makineye katılan amacın salt renkli bir taklit değil, gerçekten arzuladığımız amaç olduğundan emin olmamız gerekir" (Wiener 1960). Ed Fredkin süper zeki yapay zekâ hakkındaki endişelerinden McCorduck'ın (1979) anlattığı bir mülakatta bahsetmişti. 1970 yılına

gelindiğinde Good da bizzat risklerden bahsediyor, hatta bu tehlikelerle başa çıkmak için bir dernek kurulması çağrısında bulunuyordu (Good [1970]; ayrıca 13. Bölüm'de ele aldığımız "dolaylı normatiflik" fikirlerinden bazılarının habercisi olan sonraki makalesine bakınız (Good 1982). 1984 yılına gelindiğinde, Marvin Minsky de kilit endişelerden birçoğu hakkında yazıyordu (Minsky 1984).

10 Karş. Yudkowsky (2008a). Geleceğin potansiyel açıdan tehlikeli teknolojilerin uygulanabilir hale gelmeden *önceki* etik içerimlerini değerlendirmenin önemi hakkında, bkz. Roache (2008).

11 McCorduck (1979).

12 Newell vd. (1959).

13 Sırasıyla SAINTS programı, ANALOGY programı ve STUDENT programı. Bkz. Slagle (1963), Evans (1964, 1968) ve Bobrow (1968).

14 Nilsson (1984).

15 Weizenbaum (1966).

16 Winograd (1972).

17 Cope (1996); Weizenbaum (1976); Moravec (1980); Thrun vd. (2006); Buehler vd. (2009); Koza vd. (2003). Motorlu Taşıtlar Nevada Departmanı sürücüsüz araba için ilk ehliyeti Mayıs 2012'de verdi.

18 STANDUP sistemi (Ritchie vd. 2007).

19 Schwartz (1987). Schwartz burada Hubert Dreyfus'un yazılarında ifade bulduğunu düşündüğü şüpheci bir görüşü tanımlıyor.

20 Bu dönemde sesi çıkan eleştirmenlerden biri Hubert Dreyfus'tu. O çağın önde gelen şüphecileri arasında John Lucas, Roger Penrose ve John Searle de vardır. Gelgelelim bunlar arasında sadece Dreyfus yapay zekâdaki mevcut paradigmalardan hangi somut başarıların elde edileceğine dair iddiaları çürütmekle ilgilenmişti (yeni paradigmaların daha da ileri gidebilme ihtimaline de açık görünüyordu). Searle'in hedefinde zihin felsefesindeki işlevselci teoriler vardı, yapay zekâ sistemlerinin araçsal güçleri değil. Lucas ve Penrose klasik bir bilgisayarın asla bir insan matematikçinin yapabileceği her şeyi yapacak şekilde programlanamayacağını savunuyordu, ama herhangi bir işlevin otomatlaştırılabileceğini ya da yapay zekâların nihayetinde çok araçsal olarak güçlü olabileceğini ilkesel olarak reddetmiyordu. Cicero "hiçbir saçmalık yoktur ki bir filozof dile getirmemiş olsun," demişti (Cicero 1923, 119); ama bu kitapta kullanılan anlamda makine süper zekâsı olanağını reddeden önemli *herhangi* bir düşünür düşünmek epey zordur.

21 Ne var ki birçok uygulama için bir nöral ağda gerçekleşen öğrenme, 1800'lerin başında Adrien-Marie Legendre ve Carl Friedrich Gauss'un geliştirdiği bir istatistik tekniği olan doğrusal gerilemede gerçekleşen öğrenmeden çok az farklıdır.

22 Temel algoritma Arthur Bryson ve Yu-Chi Ho tarafından 1969'da çok aşamalı bir dinamik optimizasyon yöntemi olarak tarif edilmişti (Bryson ve Ho 1969). Nöral ağlardan, Paul Werbos 1974'te bahsetmişti (Werbos 1994), ama ancak 1986'da David Rumelhart, Geoffrey Hinton ve Ronald Williams tarafından yürütülen çalışmaların ardından (Rumelhart vd. 1986) daha geniş kesimler yavaş yavaş bu yöntemin farkına varmaya başladı.

23 Saklı katmanların olmadığı ağların öncesinde son derece sınırlı işlevselliğe sahip olduğu gösterilmiştir (Minsky ve Papert 1969).

24 Örneğin MacKay (2003).

25 Murphy (2012).

26 Pearl (2009).

27 Wainwright ve Jordan (2008). Bayes ağlarının uygulama alanları çeşit çeşittir; örneğin bkz. Pourret vd. (2008).

28 Burada oyun yapay zekâsı hakkında neden bu kadar ayrıntı verildiği merak edilebilir, zira bazıları için önemsiz bir uygulama alanı olarak görülebilir. Buradaki cevap, oyun oynamanın insan performansına karşı yapay zekâ performansının en açık ölçülerinden bazılarını sunuyor olmasıdır.

29 Newell vd. (1958, 320).

30 Bkz. Vardi (2012).

31 1976'da I.J. Good şöyle yazıyordu: "Bir Grandmaster gücünde bilgisayar programı bizi makine ultra zekâsının kıyısına getirecektir" (Good 1976). Douglas Hofstadter ise 1979'da Pulitzer Ödüllü *Godel, Escher, Bach* eserinde şöyle akıl yürütüyordu: "Soru: Bir insanı yenebilecek satranç programları olacak mıdır? Spekülasyon: Hayır. Satrançta herkesi yenebilecek programlar olabilir, ama bunlar salt satranç programları olmayacaktır. Bunlar genel zekâ programları olacaktır ve onların da insanlar gibi sağı solu belli olmayacaktır. 'Satranç oynamak ister misiniz?' 'Hayır, satranç baydı. Şiirden bahsedelim'" (Hofstadter [1979] 1999, 678).

32 Algoritma, satranca özgü bir bulgulayıcı değerlendirme işleviyle kullanılan alfa-beta budaması olan bir oyun stratejisidir. Sağlam birtakım açılış hamleleri ve oyun sonlarıyla ve çeşitli başka numaralarla birleştirildiğinde yetenekli bir satranç makinesi doğurabilir.

33 Özellikle de simülasyon oyunlarından değerlendirme bulgulayıcısını öğrenmedeki son dönem ilerlemeyle birlikte, altta yatan algoritmalardan birçoğu muhtemelen başka oyunlarda da işe yarayacaktır.

34 Samuel (1959); Schaeffer (1997, 6. Bölüm).

35 Schaeffer vd. (2007).

36 Berliner (1980a, b).

37 Tesauro (1995).

38 Bu tür programlar arasında GNU (bkz. Silver [2006]) ve Snowie (bkz. Gammoned.net [2012]) vardır.

39 Lenat donanma tasarlama sürecinde bizzat role sahipti. Şöyle yazmıştı: "Dolayısıyla son kazanma oranları %60/40 Lenat/Eurisko civarında olmalıdır, ama burada kayda değer olan husus iki tarafın da tek başına kazanamayacak olmasıdır" (Lenat 1983, 80).

40 Lenat (1982, 1983).

41 Cirasella ve Kopec (2006).

42 Kasparov (1996, 55).

43 Newborn (2011).

44 Keim vd. (1999).

45 Bkz. Armstrong (2012).

46 Sheppard (2002).

47 Wikipedia (2012a).

48 Markoff (2011).

49 Rubin ve Watson (2011).

50 Elyasaf vd. (2011).

51 KGS (2012).

52 Nilsson (2009, 318). Knuth kuşkusuz meramını abartmıştı. Yapay zekânın altından kalkamadığı birçok "düşünme görevi" vardır: Saf matematikte yeni bir alt alan bulmak, herhangi tür bir felsefe yapmak, müthiş bir detektif romanı yazmak, bir devlet darbesi yapmak ya da yeni bir tüketici ürünü tasarlamak.

53 Shapiro (1992).

54 Algı, motor kontrolü, aklıselim ve dil anlayışı konusunda insanların yeteneklerine ulaşmanın zor olmasının bir nedeni beyinlerimizin bu işlevlere wetware (evrimsel zaman ölçeklerinin ötesinde optimize edilmiş olan nöral yapılar) ayırmış olması olabilir. Halbuki satranç oynama gibi yetenekler ve mantıksal düşünme bize doğuştan gelmez; dolayısıyla belki de bu görevleri yerine getirmek için sınırlı bir havuz teşkil eden genel amaçlı bilişsel kaynaklara yaslanmak zorundayız. Belki de açıktan mantıksal muhakeme ya da hesaplamayla ilgilendiğimizde beyinlerimizin yaptığı şey bazı açılardan bir "sanal makine" işletmeye benzer: Genel amaçlı bir bilgisayarın yavaş ve hantal bir zihinsel simülasyonu. O zaman (biraz da hayal gücümüzü kullanarak) denebilir ki, klasik bir yapay zekâ programı insan düşüncesinin bir emülasyonu olmaktan çok, tersi geçerlidir: Mantıklı şekilde düşünen bir insan bir yapay zekâ programının emülasyonudur.

55 Bu örnek tartışmalıdır: ABD'de yetişkinlerin yaklaşık yüzde yirmisinin ve diğer gelişmiş ülkelerde benzer bir oranın temsil ettiği azınlık görüşüne göre, Güneş Dünya'nın etrafında dönmektedir (Crabtree 1999; Dean 2005).

56 World Robotics (2011).

57 Guizzo'daki verilerden hareketle tahmin yürütülmüştür (2010).

58 Holley (2009).

59 Kurala dayalı melez istatistiksel yaklaşımlardan da yararlanılmaktadır, ama bunlar şu anda tablonun küçük bir kısmıdır.

60 Cross ve Walker (1994); Hedberg (2002).

61 New York ve Londra merkezli bir sermaye piyasaları araştırma şirketi olan TABB Group'un istatistikleri baz alınmıştır (kişisel iletişim).

62 Nilsson (2009, 319).

63 Minsky (2006); McCarthy (2007); Beal ve Winston (2009).

64 Peter Norvig, kişisel iletişim. Makine öğrenimi sınıfları da çok popülerdir ve biraz dikey bir "büyük veri" dalgasının yansıması olabilir (örneğin Google ve Netflix Ödülü'nden esinlenerek).

65 Armstrong ve Sotala (2012).

66 Muller ve Bostrom (2016).

67 Bkz. Baum vd. (2011), orada zikredilen araştırmalardan bir diğeri ve Sandberg ve Bostrom (2011).

68 Nilsson (2009).

69 Bu bir kez daha medeniyeti yok edecek bir felaketin gerçekleşmemesine bağlıdır. Nilsson'ın başvurduğu HLMI tanımı şudur: "Yapay zekâ hem insanların yaptığı işlerin yüzde 80 civarını onlar kadar ya da onlardan daha iyi yapabilir" (Kruel 2012).

70 Tabloda hem dört farklı oylamanın sonuçları hem de birleşik sonuçları gösterilmiştir. İlk iki oylama akademik konferanslarda yapılmıştı: *PT-AI*, 2011'de Selanik'te düzenlenen *Philosophy ve Theory of AI* konferansına katılanlar (sorular 2012'de yöneltilmişti) ve yanıt oranı 88'de 43'tü; ve *AGI*, her ikisi de Oxford'da Aralık 2012'de gerçekleştirilen *Artificial General Intelligence ve Impacts ve Risks of Artificial General Intelligence* konferanslarının katılımcıları, yanıt oranı: 72/111. *EETN* oylamasına Nisan 2013'te, alandaki eseri yayımlanmış araştırmacıların meslek örgütü olan Yunan Yapay Zekâ Derneği'nin üyeleri katılmıştı (yanıt oranı: 26/250). *TOP100* anketi Mayıs 2013'te bir alıntı endeksi tarafından ölçüldüğü şekliyle yapay zekâdaki en üst 100 yazar arasındaki görüşleri yoklamıştı (yanıt oranı: 29/100).

71 28 (yazılma zamanında) yapay zekâ pratikçisi ve ilgili uzmanla yapılan mülakatlar Kruel (2011) tarafından yayımlanmıştır.

72 Birçok alanda uzman öngörülerinin güvenilmezliğini belgeleyen ciddi bir yazın vardır ve bu araştırmalardaki bulguların birçoğunun yapay zekâ alanı için de geçerli olduğunu düşünmemek için bir neden yoktur. Daha özelde, tahminciler tahminlerinde genellikle aşırı güvenli oluyorlar, gerçekte olduklarından daha isabetli tahminler yaptıklarına inanıyorlar ve dolayısıyla en güvendikleri hipotezin yanlış olabileceğine çok az ihtimal veriyorlar. (Tetlock 2005). (Başka birtakım önyargılar da belgelendirilmiştir, örneğin bkz. Gilovich vd. [2002].) Gelgelelim belirsizlik insan koşulunun kaçıp kurtulamayacağımız bir gerçeğidir ve eylemlerimizden birçoğunun uzun vadeli sonuçların az çok akla yatkın olduğu beklentilere, başka bir deyişle olasılıkçı tahminlere yaslanması kaçınılmazdır. Aleni olasılıkçı tahminler sunmayı reddetmek epistemik problemi çözmeyecek; sadece halının altına iteleyecektir (Bostrom 2007). Bunun yerine, aşırı güven kanıtına güven aralıklarımızı (ya da "güvenilir aralıkları") genişleterek, yani güvenilirlik işlevlerimizi tamamen gözden geçirerek yanıt vermeliyiz ve genel olarak farklı perspektifleri düşünerek ve entelektüel dürüstlükten taviz vermeden önyargılarımızla olabildiğince boğuşmalıyız. Uzun vadede, daha iyi tahminlerde bulunmamıza yardımcı olabilecek teknikler, eğitim yöntemleri ve kurumlar geliştirilmesi için de çalışabiliriz. Ayrıca bkz. Armstrong ve Sotala (2012).

İKİNCİ BÖLÜM

1 Bostrom (2003c) ve Bostrom'da (2006a) sunulan tanıma benzemektedir. Ayrıca Shane Legg'in tanımı ("Zekâ bir aracının çok çeşitli çevrelerde hedeflere ulaşma yeteneğini ölçer") ve bunun formülleştirilmiş halleriyle de (Legg 2008) benzerlik kurulabilir.

Keza Good'un 1. Bölüm'deki ultra zekâ tanımına da ("ne kadar akıllı olursa olsun bir insanın entelektüel faaliyetlerinin hepsini geçebilen bir makine") epey benzerdir.

2 Aynı nedenle, bir süper zeki makinenin "gerçek niyetselliği" olup olamayacağına dair hiçbir varsayımda bulunmayız (Searle tarzında, olabilir; ama bu kitabın kaygıları açısından bu alakasız görünmektedir). Ayrıca felsefi yazını kasıp kavuran zihinsel içerik hakkındaki içselcilik/dışsalcılık tartışmasında ya da bununla ilintili genişletilmiş zihin tezi meselesinde bir tutum almıyoruz (Clark ve Chalmers 1998).

3 Turing (1950, 456).

4 Turing (1950, 456).

5 Chalmers (2010); Moravec (1976, 1988, 1998, 1999).

6 Bkz. Moravec (1976). Benzer bir argümanı da David Chalmers (2010) öne sürmüştür.

7 Gözlem seçilim teorisine dair genel bir tetkik için, bkz. Bostrom (2002a). Mevcut meseleye dair özgül bir uygulama için, bkz. Shulman ve Bostrom (2012). Herkesçe anlaşılır kısa bir giriş yazısı için, bkz. Bostrom (2008b).

8 Sutton ve Barto (1998, 21f); Schultz vd. (1997).

9 Bu terim Eliezer Yudkowsky'ye aittir; örneğin bkz. Yudkowsky (2007).

10 Bu, Good (1965) ve Yudkowsky (2007) tarafından tarif edilen senaryodur. Gelgelelim yinelemeli dizinin zekâ geliştirme değil, bunun yerine tasarım basitleştirme içerdiği bazı adımlara sahip olduğu bir alternatifi de düşünebiliriz. Yani bazı aşamalarda tohum yapay zekâ sonraki iyileştirmeleri bulmayı kolaylaştırmak için kendini yeniden yazabilir.

11 Yani araçsal açıdan yararlı olan bilişsel işleyiş ve iletişim biçimlerinin önünü açmak için yeterli, ama normal bir insan bedeninin duyu organları ve kasları tarafından sağlanan arayüzle kıyaslandığında yine de çok zayıf.

12 Sandberg (2013).

13 Bkz. Sandberg ve Bostrom (2008, 79-81) kitabının "Bilgisayar Gereklilikleri" başlıklı kısmı.

14 Düşük düzey bir başarı biyolojik açıdan mikro dinamikleri olan ve yavaş dalga uyku durumu ya da faaliyete bağımlı esneklik gibi yeni ortaya çıkan türler için tipik faaliyetler sergileyen bir beyin simülasyonu olabilir. Bu tür bir simülasyon nörobilimsel araştırma için yararlı bir sınama ortamı olabilirken (ama ciddi etik meseleleri gündeme getirmeye de yakındır), bir tam beyin emülasyonu olarak sayılabilmesi için simülasyon beynin yapabildiği entelektüel çalışmanın önemli bir kısmını yerine getirebilecek kadar doğru olmalıdır. Pratik olarak, bir insan beyni simülasyonunun bir tam beyin emülasyonu sayılabilmesi için bütünlüklü sözlü düşüncelerini ifade edebilmesi ya da bunu yapmayı öğrenme kapasitesine sahip olması gerektiğini söyleyebiliriz.

15 Sandberg ve Bostrom (2008).

16 İlk harita için, bkz. Albertson ve Thomson (1976) ve White vd. (1986). Birleşik (ve bazı örneklerde düzeltilmiş) ağ için, bkz. "WormAtlas" internet sitesi (http://www.wormatlas.org/).

17 *C. Elegans* emülasyonlarına ilişkin geçmişteki girişimler ve akıbetleri hakkında bir değerlendirme için, bkz. Kaufman (2011). Kaufman bu alanda çalışan hırslı bir doktora öğrencisi David Dalrymple'ın şöyle dediğini söyler: "Optogenetik tekniklere göre, şu anda canlı bir *C. Elegans* sinir sisteminde bir yere yüksek çıktılı otomatlaştırılmış bir sistemden yararlanarak yazma ve okuma yeterliliğine ulaşmanın ürkütücü bir öneri olmadığı noktadayız. [...] 2-3 yıl içinde *C. elegans* meselesinin bitmesini bekliyorum. Ne pahasına olursa olsun, bunu 2020'de hâlâ çözüm bekleyen bir problem olarak görürsek çok şaşırırım" (Dalrymple 2011). Elle kodlanmış biyolojik gerçekçilik amaçlayan beyin modelleri (otomatik olarak yaratılmaktan ziyade) temel bir işlevselliğe ulaşmıştır; örneğin bkz. Eliasmith vd. (2012).

18 *Caenorhabditis elegans* gerçekten de bazı uygun özel niteliklere sahiptir. Örneğin organizma şeffaftır ve sinir sisteminin bağlantılı örüntüsü bireyden bireye değişiklik göstermez.

19 Eğer tüm beyin simülasyonundan ziyade nöromorfik yapay zekâ son ürünse, o zaman ilintili içgörülerin *insan* beynini simüle etme girişimleri yoluyla çıkarılması mümkün olabilir de olmayabilir de. İnsan olmayan hayvan beyinlerinin incelenmesi sırasında önemli kortikal numaraların keşfedilmesi mümkündür. Bazı hayvan beyinleriyle çalışmak insan beyinlerine kıyasla daha kolay olabilir ve küçük beyinler tarama ve modelleme için daha az kaynağa ihtiyaç duyar. Hayvan beyinleri üzerine araştırmalar daha az düzenlemeye tabi olacaktır. İnsan düzeyinde ilk makine zekâsının uygun bir hayvanın bir tüm beyin emülasyonunu tamamlayarak ve ardından ortaya çıkan dijital zihni geliştirmenin yollarını bularak yaratılacak olduğu bile düşünülebilir. Dolayısıyla insanlık müstahakını mutlu bir laboratuvar faresi ya da makaktan bulabilir.

20 Uauy ve Dangour (2006); Georgieff (2007); Stewart vd. (2008); Eppig vd. (2010); Cotman ve Berchtold (2002).

21 2007'de Dünya Sağlık Örgütü'ne göre, neredeyse 2 milyar bireyin iyot alımı yetersizdir (*The Lancet* 2008). *Aşırı* iyot eksikliği nörolojik gelişimi engeller ve ortalama 12,5 civarında IQ puanı kaybı içeren kretenizme yol açar (Qian vd. 2005). Bu durumu tuz takviyesi yoluyla kolayca ve ucuzca engellemek mümkündür (Horton vd. 2008).

22 Bostrom ve Sandberg (2009a).

23 Bostrom ve Sandberg (2009b). Farmakolojik ve besinsel gelişimden kaynaklı tipik bir sözde performans artışı; işler belleği, dikkati vb. ölçen test görevlerinde %10-20 civarındadır. Ama bu tür kazanımların gerçek olup olmadığı, uzun vadede sürdürülüp sürdürülemeyeceği ve gerçek dünyadaki problem durumlarında buna denk düşen ilerletilmiş sonuçların işareti olup olmadığı şüphelidir (Repantis vd. 2010). Sözgelimi bazı örneklerde performans boyutlarında test görevlerince ölçülmeyen tazmin edici bir kötüleşme olabilir (Sandberg ve Bostrom 2006).

24 Biliş yetisini geliştirmenin kolay bir yolu olsaydı, evrim bundan çoktan istifade ederdi. Dolayısıyla araştırılacak en ümit vaat edici nootropik türü geçmiş çevrede uygunluğu azaltacak olan bir tarzda zekâyı ilerletmeyi vaat eden bir tür olabilir – örneğin doğumda kafa büyüklüğünü artırarak ya da beynin glükoz metabolizmasını güçlendirerek. Bu fikre dair ayrıntılı bir değerlendirme için (birkaç önemli kayıt düşmek kaydıyla), bkz. Bostrom (2009b).

25 Spermleri incelemek zordur, çünkü embriyoların aksine sadece tek bir hücreden oluşurlar ve dizileme yapmak için bir hücrenin yok edilmesi gerekir. Oositler de tek hücrelidir; gelgelelim birinci ve ikinci hücre bölünmeleri asimetriktir ve çok küçük sitoplazmayla bir yavru hücre (kutup küreciği) üretir. Kutup kürecikleri ana hücreyle aynı sayıda gene sahip oldukları ve gereğinden fazla oldukları için (nihayetinde dejenere olurlar), biyopsi yapılıp görüntüleme için kullanılabilirler (Gianaroli 2000).

26 Bu pratiklerin her biri ilk başta etik tartışmalara konu olmuştu, ama kabulün artması yönünde bir eğilim olduğu görünmektedir. İnsan genetiği mühendisliği ve embriyo seçilimi konusundaki tutumlar kültürden kültüre muazzam farklılıklar göstermektedir ve dolayısıyla bazı ülkeler ilk başta ihtiyatlı bir duruş sergileseler bile, yeni tekniklerin gelişimi ve uygulanması muhtemelen gerçekleşecektir, ama bunun hangi hızda olacağı ahlaki, dini ve siyasi baskılardan etkilenecektir.

27 Davies vd. (2011); Benyamin vd. (2013); Plomin vd. (2013). Ayrıca bkz. Mardis (2011); Hsu (2012).

28 Yetişkin IQ'sunun geniş anlamda kalıtsallığının gelişmiş ulusların orta sınıf katmanlarında 0,5-0,8 düzeyinde olduğu hesaplanmaktadır (Bouchard 2004, 148). Additif genetik etkenlere atfedilebilen değişiklik kısmını hesaplayan dar-anlamda kalıtsallık daha düşüktür (0,3-0,5 düzeyindedir), ama yine de kayda değerdir (Devlin vd. 1997; Davies vd. 2011; Visscher vd. 2008). Bu tahminler farklı nüfuslar ve çevreler için değişebilir, zira kalıtsallıklar incelenen nüfusa ve çevreye bağlı olarak farklılık gösterir. Örneğin düşük kalıtsallıklar çocuklar arasında ve yoksun çevrelerde yaşayanlarda görülür (Benyamin vd. 2013; Turkheimer vd. 2003). Nisbett vd. (2012) bilişsel yetenekteki değişikliklere çeşitli çevresel etkileri değerlendirmektedir.

29 Aşağıdaki birkaç paragrafta Carl Shulman ile ortak çalışmamızdan yararlanılmıştır (Shulman ve Bostrom 2014).

30 Biliş yeteneğine additif genetik etkiler hakkında kusurlu bilgiyle etki boyutları azaltılacaktır. Gelgelelim az miktarda bilgi bile görece uzun bir yol gidilmesini sağlayacaktır, çünkü seçilimden elde edilecek kazanımlar tahmin edebileceğimiz çeşitlilik kısmıyla doğrusal ölçekte değildir. Bunun yerine, seçilimizin etkililiği tahmin edilen ortalama IQ'nun standart sapmasına bağlıdır ki, bu da çeşitliliğin *kare kökü* olarak ölçeklendirilir (??). Örneğin çeşitliliğin yüzde 12,5'i açıklanabilirse, bunlar Tablo 1'dekinin yarısı büyüklüğünde etkiler üretebilir (yüzde 50). Karşılaştırma adına yakın dönemdeki bir çalışma (Rietveld vd. 2013) halihazırda çeşitliliğin yüzde 2,5'ini tespit ettiğini iddia etmiştir.

31 Karşılaştırma adına, bugün standart pratik ondan daha az embriyonun yaratılmasını içerir.

32 Bu tablo Shulman ve Bostrom'dan (2014) alınmıştır. Tahmini IQ'ların embriyolar arasında 7,5 puanlık bir standart sapmayla Gaussçu dağılımını varsayan bir oyuncak modeline dayalıdır. Farklı sayıda embriyoyla sağlanacak bilişsel gelişim miktarı etkilerini bildiğimiz additif genetik çeşitlilikler bakımından embriyoların birbirinden ne kadar farklı olduğuna bağlıdır. Kardeşlerin akrabalık katsayısı ½'dir ve ortak additif genetik çeşitlilikler yetişkin akışkan zekâsındaki değişikliğin yarısını ya da daha azını açıklar (Davies vd. 2011). Bu iki olgu, gelişmiş ülkelerde gözlemlenen nüfusun standart sapmasının 15 puan olduğu yerlerde, bir grup embriyo içinde genetik etkilerin standart sapmasının 7,5 puan ya da daha az olacağını gösteriyor.

33 Yetişkin ve embriyon kök hücreleri sperm hücreleri ve oksitler olarak geliştirilebilir ve bunlar da akabinde bir embriyo üretilecek şekilde birleştirilebilir (Nagy vd. 2008; Nagy ve Chang 2007). Yumurta hücresi öncülleri de süreç için embriyonik kök hücre hatları üretebilecek partenogenetik blastokistler, yani döllenmemiş ve cansız embriyolar oluşturabilir (Mai vd. 2007).

34 Bu görüş Cyranoski'de (2013) aktarıldığı şekliyle Katsuhiko Hayashi'ye aittir. Kök hücre etiğini ve zorluklarını masaya yatıran bilimcilerin uluslararası konsorsiyumu olan Hinxton Group'un 2008'deki tahminine göre, on yıl içinde insan kök hücresinden türetilen gametler mevcut olacaktır (Hinxton Group 2008) ve gelişmeler de bununla genel anlamda tutarlıdır.

35 Sparrow (2013); Miller (2012); The Uncertain Future (2012).

36 Sparrow (2013).

37 Seküler kaygılar toplumsal eşitsizliğe etkilere, işlemin tıbbi güvenliğine, geliştirilmiş bir "fare yarışı"ndan duyulan korkulara, ebeveynlerin müstakbel döllerine karşı hakları ve sorumluluklarına, 20. yüzyıl öjeni biliminin gölgesine, insan vakarı kavramına ve devletlerin yurttaşların üreme tercihlerine müdahalesinin düzgün sınırlarına odaklanır. (Bilişsel geliştirmenin etiği tartışmaları için, bkz. Bostrom ve Ord [2006], Bostrom ve Roache [2011], Sandberg ve Savulescu [2011].) Bazı dini gelenekler embriyoların ahlaki statüsüne odaklanan kaygılar ya da yaratılış şeması dahilinde insan aracılığının doğru sınırları gibi başka kaygılar da doğurabilir.

38 Aynı soydan çiftleştirmenin olumsuz etkilerini savuşturmak için yinelemeli embriyo seçilimi ya başlangıçta çok sayıda donör arzına ya da zararlı gerilemeli alelleri azaltmak için kayda değer seçilim gücü harcanmasını gerektirecektir. İki alternatif de dölün ebeveynleriyle genetik açıdan daha az ilişkili olması (ama birbirleriyle daha yakından ilişkili olmaları) eğilimini güçlendirecektir.

39 Bir engel epigenetiğinin ne kadar zor olacağı henüz bilinmiyor (Chason vd. 2011; Iliadou vd. 2011).

40 Biliş yeteneği son derece kalıtsal bir özellik olmakla birlikte, zekâya bireysel olarak büyük bir olumlu etkisi olan *ortak* aleller ya da polimorfizmler çok az olabilir ya da hiç olmayabilir (Davis vd. 2010; Davies vd. 2011; Rietveld vd. 2013). Dizilim yöntemleri ilerledikçe, düşük frekanslı alellerin ve onların bilişsel ve davranışsal bağıntılarının haritalandırılması giderek mümkün hale gelecektir. Homozigotlarda genetik bozukluklara sebep olan bazı alellerin heterozigot taşıyıcılarda ciddi bilişsel üstünlükler sağlayabileceğini gösteren bazı teorik kanıtlar vardır ve bu da Gaucher, Tay-Sachs, ve Niemann-Pick heterozigotlarının kontrol gruplarından yaklaşık 5 IQ puanı daha yüksek olacağı tahminine yol açmıştır (Cochran vd. 2006). Bunun doğru olup olmadığını zaman gösterecektir.

41 Shulman ve Bostrom'dan uyarlanmıştır (2014).

42 Bostrom (2008b).

43 Bir yazıda (Nachman ve Crowell 2000) her kuşakta genom başına 175 mutasyon tahmininde bulunulur. Farklı yöntemler kullanan bir diğer yazıda (Lynch 2010), ortalama yenidoğanın 50 ila 100 arasında yeni mutasyon geçirdiği tahmin edilmektedir ve Kong vd. (2012) kuşak başına yaklaşık 77 mutasyon gibi bir rakam telaffuz ederler. Bu mutasyonların çoğu işleyişi etkilemez ya da çok az etkiler; ama çok

az zararlı mutasyonların birleşik etkileri ciddi bir uygunluk kaybı olabilir. Ayrıca bkz. Crow (2000).

44 Crow (2000); Lynch (2010).

45 Bu fikre potansiyel açıdan önemli bazı itirazlar yöneltilmiştir. Modal genomun problemlerden kaçınmak için bazı düzeltmelere ihtiyaç duyması mümkündür. Örneğin genomun bazı parçaları tüm parçaların belli düzey bir verimlilikle etkileşim halinde olacağı varsayımıyla diğer parçalarla etkileşimde bulunacak şekilde tasarlanabilir. Dolayısıyla bu parçaların verimliliğini artırmak bazı metabolik yolaklarda ıskalamaya yol açabilir.

46 Elbette çok geçmeden bazı etkileri olabilir, sözgelimi insanların neler geleceğine dair beklentilerini değiştirerek.

47 Louis Harris & Associates (1969); Mason (2003).

48 Kalfoglou vd. (2004).

49 Veriler elbette sınırlıdır, ama çocukluk yetenek testlerinde 10 binde 1 sonuçla seçilen bireylerin boylamsal çalışmalarda kadrolu profesörler olmaları, patent almaları ve iş hayatında başarılı olmalarının çok daha az istisnai puanlara sahip olanlara kıyasla çok daha muhtemel olduğu gösterilmiştir (Kell vd. 2013). Roe (1953) 64 önde gelen bilimciyi araştırmış ve medyan bilişsel yeteneğin nüfus normunun üstünde üç ila dört standart sapma gösterdiğini; genel olarak bilimcilerdeyse çok daha yüksek olduğunu tespit etmiştir. (Bilişsel yetenek yaşam boyu elde edilen kazançla ve yaşam süresi, boşanma oranları ve okulu bırakma olasılığı gibi parasal olmayan sonuçlarla da bağıntılıdır [Deary 2012].) Bilişsel yetenek dağılımındaki yukarı doğru bir kaymanın arkadan gelenlere orantısız derecede büyük etkileri olacaktır, özellikle de son derece yetenekli olanların sayısını artırmak ve zekâ geriliği ve öğrenme sorunları olan insanların sayısını azaltmak bakımından. Ayrıca bkz. Bostrom ve Ord (2006) ile Sandberg ve Savulescu (2011).

50 Örneğin Warwick (2002). Stephen Hawking bu adımı atmanın makine zekâsındaki ilerlemelere ayak uydurmak için zorunlu olabileceğini bile söylemiştir: "Beyin ile bilgisayar arasında doğrudan bağlantıyı mümkün kılan olanaklı teknolojileri olabildiğince çabuk geliştirmeliyiz ki, yapay beyinler insan zekâsının karşısına dikileceğine ona katkıda bulunsun" (akt. Walsh [2001]). Ray Kurzweil şu sonuca varır: "Hawking'in [...] önerisi, yani beyin ile bilgisayarlar arasında doğrudan bağlantı söz konusu olduğu oranda, bunun hem makul hem istenen hem de kaçınılmaz bir şey olduğunu kabul ediyorum. Ben yıllardır bunu savunuyorum" (Kurzweil 2001).

51 Bkz. Lebedev ve Nicolelis (2006); Birbaumer vd. (2008); Mak ve Wolpaw (2009); Nicolelis ve Lebedev (2009). İmplantlar yoluyla geliştirme problemine daha kişisel bir bakış açısı Chorost'ta bulunabilir (2005, 11. Bölüm).

52 Smeding vd. (2006).

53 Degnan vd. (2002).

54 Dagnelie (2012); Shannon (2012).

55 Perlmutter ve Mink (2006); Lyons (2011).

56 Koch vd. (2006).

57 Schalk (2008). Son teknoloji hakkında genel bir değerlendirme için, bkz. Berger vd. (2008). Bunun zekâ gelişimine yardımcı olacağı görüşü için, bkz. Warwick (2002).

58 Bazı örnekler: Bartels vd. (2008); Simeral vd. (2011); Krusienski ve Shih (2011); Pasqualotto vd. (2012).

59 Örneğin Hinke vd. (1993).

60 Bunun kısmi istisnaları vardır, özellikle de erken dönem duyusal işlemede. Örneğin birincil görme korteksi retinotopik bir haritalandırmadan yararlanır, bu kabaca demektir ki bitişik nöral birlikler retinaların bitişik alanlarından girdi alır (ama göz baskınlık şeritleri haritalandırmayı biraz karmaşıklaştırmaktadır).

61 Berger vd. (2012); Hampson vd. (2012).

62 Bazı beyin implantları iki öğrenme biçimine gerek duyar: Organizmanın nöral temsillerini yorumlamayı öğrenen araç ve uygun nöral ateşleme örüntülerini yaratarak sistemi kullanmayı öğrenen organizma (Carmena vd. 2003).

63 Kolektif oluşumları (şirketler, sendikalar, hükümetler, kiliseler vb.) yapay zekâ aracıları olarak –bilgiyi temsil edip çıkarımda bulunabilen ve eyleme geçebilen sensörlere ve efektörlere sahip varlıklar olarak– görmemiz gerektiği ileri sürülmüştür (örneğin Kuipers [2012]; kolektif temsillerin var olup olamayacağı hakkında karş. Huebner [2008]). Bunların güçlü ve ekolojik açıdan başarılı olduğu açıktır, ama yeterlilikleri ve iç durumları insanlarınkinden farklıdır.

64 Hanson (1995, 2000); Berg ve Rietz (2003).

65 Sözgelimi işyerinde patronlar çalışanların hırsızlık ve kaytarma gibi davranışlarında tepelerine çökmek için onlara her işgününün sonunda bir şey çalıp çalmadıklarını ve ellerinden geldiğince çok çalışıp çalışmadıklarını sormak için yalan detektörleri kullanabilirler. Siyasetçilere ve iş dünyasının liderlerine de paydaşlarının ya da seçmenlerinin çıkarlarını gerçekten gözetip gözetmedikleri sorulabilir. Diktatörler ise rejim içindeki tehlikeli generalleri ya da halk arasındaki fesat odaklarını hedef almak için bunları kullanabilirler.

66 Beyin görüntüleme tekniklerinin güdülenmiş bilişin nöral işaretlerini tespit etmeyi mümkün kılacağı düşünülebilir. Kendi kendini kandırma detektörü olmadan yalan detektörü kendi propagandalarına inanan bireylerin işine gelir. Kendi kendini aldatma testlerinde daha iyi testler akliliği öğretmek ve önyargıları azaltmayı amaçlayan müdahalelerin etkililiğini incelemek için de kullanılabilir.

67 Bell ve Gemmel (2009). İlk örneklerden biri MIT'den Deb Roy'un çalışmalarında görülür. Kendisi oğlunun ilk üç yılının her anını kayda almıştır. Bu işitsel-görsel verilerin analizi dil gelişimine ait bilgi verir; bkz. Roy (2012).

68 Dünyadaki toplam biyolojik insan nüfusundaki büyüme sadece küçük bir etken olarak katkı sunacaktır. Makine zekâsını içeren senaryolar dünya nüfusunun (dijital zihinler dahil) çok kısa zamanda çok büyük oranlarda büyüdüğünü görebilir. Fakat süper zekâya giden yol yapay zekâ ya da tüm beyin emülasyonu içerir, bu yüzden bu alt kısımda bunu düşünmemiz gerekmiyor.

69 Vinge (1993).

ÜÇÜNCÜ BÖLÜM

1 Vernor Vinge bu tür hızlandırılmış insan zihinlerine atıfta bulunmak için "zayıf süper zekâ" terimini kullanmıştır (Vinge 1993).

2 Örneğin eğer çok hızlı bir sistem mazurka dansı yapmak dışında bir insanın yapabileceği her şeyi yapabiliyorsa, ona yine de hızlı süper zekâ diyebiliriz. Ekonomik ya da stratejik açıdan önemi olan bu temel bilişsel yeterliliklerle ilgilenmeliyiz.

3 İnsan beyinleriyle kıyaslandığında en azından milyon kat hızlanma fiziken mümkündür, daha verimli bilgi işlemleriyle kıyaslandığında ilintili beyin süreçlerinin hız ve enerjisindeki farklılığını düşünerek görülebilir. Işık hızı nöral iletime kıyasla bir milyon kat daha büyüktür, sinaps uçları termodinamik açıdan zorunlu olandan bir milyon kat daha fazla ısı yayar ve mevcut transistor sıklıkları nöron artış sıklıklarından bir milyon kat daha hızlıdır (Yudkowsky [2008a]; ayrıca bkz. Drexler [1992]). Hızlı süper zekânın nihai sınırları ışık hızında iletişim gecikmeleri, durum geçişlerinin hızındaki kuantum sınırları ve zihni ihtiva etmek için gerekli hacim tarafından sınırlandırılmıştır (Lloyd 2000). Lloyd'un (2000) tarif ettiği nihai dizüstü bilgisayar 3.8×1029× hızda .4×1021 FLOPS beyin emülasyonu çalıştıracaktır (emülasyonun yeterince paralelleştirilebileceğini varsayarsak). Gelgelelim Lloyd'un inşasının teknolojik açıdan akla yatkın olması amaçlanmamıştır; sadece temel fizik yasalarından hemen çıkarılabilecek olan hesaplama kısıtlamalarını örneklendirmesi amaçlanmıştır.

4 Emülasyonlarda insan türü bir zihnin delirmeden ya da tekdüzeleşmeden önce ne kadar süre çalışabileceği meselesi de vardır. Görev çeşitliği ve düzenli tatillerle bile insan türü bir zihnin psikolojik problemler geliştirmeden binlerce öznel yıl yaşayabileceği kesin değildir. Ayrıca eğer toplam hafıza kapasitesi sınırlıysa (sınırlı bir nöron popülasyonuna sahip olmanın sonucu), o zaman kümülatif öğrenme sonsuza dek süremez: Belli bir noktanın ötesinde zihin yeni öğrendiği her şeye karşı bir şeyi unutmaya başlamalıdır. (Yapay zekâ bu potansiyel problemleri iyileştirecek şekilde tasarlanabilir.)

5 Bu bakımdan, mütevazı 1 m/s hızda çalışan nanomekanizmalar tipik nanosaniye zaman ölçeğine sahiptir. Bkz. Drexler (1992) Kısım 2.3.2. Robin Hanson normal hızın 260 katı hızda hareket eden 7-mm'lik "tinkerbell" robot bedenlerden bahseder (Hanson 1994).

6 Hanson (2012).

7 "Kolektif zekâ" bilgisayar donanımının düşük seviye paralelizasyonuna değil, insanlar gibi zeki özerk aracılar düzeyinde paralelizasyonuna işaret eder. Muazzam düzeyde paralel makinede tek bir emülasyon uygulamak, eğer paralel bilgisayar yeterince hızlıysa, hızlı süper zekâyla sonuçlanabilir. Ama kolektif zekâ üretmeyecektir.

8 Tekil bileşenlerin hız ya da niteliğinde iyileştirmeler kolektif zekânın performansını dolaylı olarak etkileyebilir, ama burada esasen bu tür iyileştirmeleri sınıflandırmamızda süper zekânın diğer iki biçimi altında ele alıyoruz.

9 Yüksek nüfus yoğunluğunun Üst Yontma Taş Devri'ni tetiklediği ve belli bir eşiğin ötesinde, kültürel karmaşıklık birikiminin çok daha kolay hale geldiği savunulmuştur (Powell vd. 2009).

10 Peki ya internet? Henüz bir süper boyutta atılıma varmadığı görünmektedir. Belki de nihayetinde varacaktır. Burada listelenen diğer örneklerin tüm potansiyellerini ortaya sermeleri yüzyıllar, hatta binyıllar almıştır.

11 Bunun gerçekçi bir düşünce deneyi olması amaçlanmamıştır. Mevcut teknolojiyle yedi katrilyon insan organizmasını yaşatabilecek kadar büyük bir gezegen, eğer çok hafif maddeden yapılmadıysa ya da boşsa ve baskıyla yahut diğer yapay araçlarla ayakta tutuluyorsa, içpatlama yaşayacaktır. (Bir Dyson küresi ya da bir Shellworld daha iyi bir çözüm olabilir.) Tarih böyle engin bir yüzeyde farklı ilerlerdi. Bunların hepsini bir kenara bırakalım.

12 Buradaki odak noktamız birleşik bir zekânın işlevsel özellikleridir, bu tür bir zekânın *qualia* sahibi mi olacağı ya da öznel bilinçli deneyime sahip olma anlamında bir zihin mi olacağı değildir. (Fakat insan beyinlerine kıyasla az çok bütünlüklü olan zekâlardan ne tür bilinçli deneyim doğabileceği düşünülebilir. Bilinç hakkındaki bazı düşüncelere göre (küresel çalışma uzamı teorisi gibi), daha bütünlüklü beyinlerin daha ferah bir bilince sahip olmaları beklenebilir. Karş. Baars [1997], Shanahan [2010] ve Schwitzgebel [2013].)

13 Belli bir süre yalıtık kalmış olan küçük gruplar bile, yine de, büyük bir kolektif zekânın entelektüel çıktılarından yarar sağlayabilirler. Örneğin kullandıkları dil daha büyük bir dilsel topluluk tarafından geliştirilmiş olabilir ve kullandıkları araçlar küçük grup yalıtık hale gelmeden önce çok daha büyük bir nüfusta icat edilmiş olabilir. Fakat küçük bir grup her zaman yalıtılmış olsa bile, yine de daha büyük bir kolektif zekânın parçası olabilir, yani sadece mevcut değil aynı zamanda tüm geçmiş kuşakları da kapsayan kolektif zekâ, bir ileri besleme bilgi işleme sistemi olarak işlev gösterebilecek bir toplam.

14 Church-Turing tezine göre, tüm hesaplamalı işlevler bir Turing makinesi tarafından yapılabilir. Süper zekânın üç biçiminden herhangi biri bir Turing makinesinin simülasyonunu yapabildiğinden (eğer sınırsız belleğe erişim verilirse ve sınırsız işlem göstermesine izin verilirse), bunlar bu biçimsel kıstasa göre hesaplama açısından denktir. Gerçekten de ortalama bir insan (sınırsız boş kâğıdı ve sınırsız zamanı varsa) bir Turing makinesini hayata geçirebilir ve dolayısıyla aynı kıstasa göre eşdeğerdir. Fakat amaçlarımız açısından mühim olan bu farklı sistemlerin sınırlı hafıza ve makul bir zamanda *pratikte* neleri başarabileceğidir. Verimlilik çeşitlilikleri o kadar büyüktür ki, hemen bazı ayrımlar yapılabilir. Örneğin 85 IQ'su olan tipik bir bireye bir Turing makinesini hayata geçirmesi öğretilebilir. (Bilhassa yetenekli ve uysal bir şempanzeye bunun öğretilebileceğini de tasavvur edebiliriz.) Fakat tüm pratik niyetler ve amaçlar açısından, bu tür bireyin sözgelimi genel görelilik teorisini kendi başına geliştirmesini ya da bir Fields madalyası kazanmasını bekleyemeyiz.

15 Sözlü hikâye anlatıcılığı gelenekleri (Homeros destanları gibi) büyük eserler üretebilir, ama belki de katkıda bulunan yazarlardan bazılarının istisnai yetenekleri vardır.

16 Bileşen olarak hızlı ya da nitelikli süper zekâya sahip olan zekâlar içermiyorsa.

17 Tüm bu problemlerin ne olabileceğini saptamayı beceremememiz kısmen denemiyor olmamızdan kaynaklanıyor olabilir: Hiçbir bireyin ya da hiçbir somut örgütün altından kalkamadığı entelektüel işleri ayrıntılandırarak zaman harcamanın pek de bir anlamı yok. Ama bu işlerden bazılarını kavramlaştırmanın bile şu an yapmak için gerekli beyinlerden mahrum olduğumuz işlerden biri olması da mümkündür.

18 Karş. Boswell (1917); ayrıca bkz. Walker (2002).

19 Bu esasen bir nöron altkümesinde kısa patlamalarla gerçekleşir; çoğunun daha sakin ateşleme hızları vardır (Gray ve McCormick 1996; Steriade vd. 1998). 750 HZ kadar yüksek ateşleme sıklığına ulaşabilecek bazı nöronlar vardır ("çenebaz nöronlar", "ritmik açıdan hızlı patlayan" hücreler olarak da bilinir), ama bunlar uç örneklerdir.

20 Feldman ve Ballard (1982).

21 İletim sürati akson çapına (kalın aksonlar daha hızlıdır) ve aksonun miyelinlenmiş olup olmadığına bağlıdır. Merkezi sinir sistemi dahilinde, transmisyon gecikmeleri bir milisaniyeden daha az ile 100 milisaniye arasında değişebilir (Kandel vd. 2000). Optik kablolarda transmisyon %68*c* civarındadır (malzemenin kırılma endeksinden ötürü). Elektrik kabloları kabaca aynı hızdadır, %59-77 *c*.

22 Sinyal hızının %70 *c* olduğunu varsayarsak. %100 *c* varsayımı tahminimizi 1.8×1018 m3'e kadar çıkarır.

23 Bir yetişkin erkeğin beynindeki nöron sayısının 86.1 +- 8.1 olduğu tahmin edilmektedir; bu sayıya beyinleri parçalayıp, hücre çekirdeklerini ayırıp, nöron-spesifik işaretli olanları sayarak varılmıştır. Eskiden 75-125 milyar nöron civarındaki tahminler yaygındı. Bunlar tipik şekliyle temsili küçük bölgelerdeki hücre yoğunluklarının elle sayılmasına dayalıydı (Azevedo vd. 2009).

24 Whitehead (2003).

25 Bilgi işleme sistemlerinin hesaplama ve veri depolama için moleküler ölçekte süreçlerden yararlanmaları ve en azından gezegen boyutuna ulaşmaları çok muhtemeldir. Kuantum mekaniğinin, genel göreliliğin ve termodinamiğin koyduğu nihai fiziksel sınırlar ise bu "Jüpiter beyni" düzeyinin çok ötesindedir (Sandberg 1999; Lloyd 2000).

26 Stansberry ve Kudritzki (2012). Dünya üzerinde veri merkezlerinde kullanılan elektrik, toplam elektrik kullanımının %1,1–1,5'ine denk düşmektedir (Koomey 2011). Ayrıca bkz. Muehlhauser ve Salamon (2012).

27 Bu bir aşırı basitleştirmedir. İşler belleğin sürdürebileceği küçük birim sayısı hem bilgiye hem de görevden göreve değişir, ama az sayıda küçük birim ile sınırlı olduğu açıktır. Bkz. Miller (1956) ve Cowan (2001).

28 Boole kavramlarını (mantık kurallarının tanımladığı kategorileri) öğrenmenin zorluğunun mantıksal açıdan en kısa eşdeğer önerme formülünün uzunluğuyla orantılı olması örnek gösterilebilir. Tipik şekliyle, 3-4 tamsayı uzunluğunda formülleri bile öğrenmesi çok zordur. Bkz. Feldman (2000).

29 Bkz. Landauer (1986). Bu çalışma insanlarda öğrenme ve unutma hızlarına dair deneysel tahminlere dayalıdır. Zımni öğrenmeyi hesaba katmak hesabı biraz daha artırabilir. Eğer sinaps başına ~1 bitlik bir depolama kapasitesi varsayarsak insanın bellek kapasitesinin yaklaşık 10^{15} bit üzerinde bir *üst sınır* elde edilir. Farklı tahminlere dair bir değerlendirme için, bkz. Sandberg ve Bostrom'un A Endeksi (2008).

30 Kanal gürültüsü eylem potansiyellerini tetikleyebilir ve sinaptik gürültü aktarılan sinyallerin gücünde ciddi çeşitlilik üretir. Sinir sistemlerinin gürültü toleransı ile maliyetler (kitle, büyüklük, zamansal gecikmeler) arasında muazzam ödünleşimler yapacak şekilde evrildiği gözükmektedir; bkz. Faisal vd. (2008). Örneğin aksonlar

0.1 μm'dan daha ince olamaz, aksi takdirde iyon kanallarının rastgele açılması kendiliğinden eylem potansiyelleri yaratır (Faisal vd. 2005).

31 Trachtenberg vd. (2002).

32 Bellek ve bilgi işlem gücü bakımından, ama enerji verimliliği bakımından değil. Bu satırlar yazıldığı sırada dünya üzerindeki en hızlı bilgisayar China's Tianhe-2 idi, ama Haziran 2013'te yerini 33.86 petaFLOPS performansı olan Cray Inc. Titan'a kaptırdı. Bu bilgisayar 17,6 MW güce sahiptir, yani beynin ~20 W'sinden neredeyse altı kat büyüktür.

33 Makine üstünlüğünün kaynaklarına dair bu değerlendirmenin ayrıştırıcı olduğunu belirtelim: Sıralanan maddelerden bazıları hayal mahsulü bile olsa argümanımız başarılı olur, yeter ki yeterince büyük bir üstünlük sağlayabilecek en az bir kaynak olsun.

DÖRDÜNCÜ BÖLÜM

1 Sistem kesinkes tanımlanmış bir noktada bu taban çizgilerinden birine ulaşamayabilir. Bunun yerine sistemi ilerletici gelişim görevlerinin pek çoğunda sistemin dış araştırma takımından daha iyi performans göstermeye başladığı bir zaman aralığı olabilir.

2 Son elli yılda, mevcut dünya düzeninin dakikalar ya da saatler içinde (küresel nükleer savaş) sona ereceği en az bir senaryo geniş kabul görmüştür.

3 Flynn etkisinin (son altmış küsur yılda birçok nüfusta IQ puanlarının her on yılda bir 3 IQ puanı hızında artış yaşaması) Birleşik Krallık, Danimarka ya da Norveç gibi fevkalade gelişmiş bazı ülkelerde son yıllarda durmuş, hatta tersine dönmüş gözükmektedir (Teasdale ve Owen 2008; Sundet vd. 2004). Geçmişte Flynn etkisinin sebebi (ve genel zekâda gerçek bir kazanımı mı yoksa IQ testi tarzı bulmacaları çözme konusunda daha ileri bir yetenekten mi ibaret olduğu ve kapsamı) epey bir tartışmaya konu olmuştur ve hâlâ bilinmemektedir. Flynn etkisi (en azından kısmen) gerçek bilişsel kazanımları yansıtsa bile ve etkisi şu anda azalmakta, hatta tersine dönmekte olsa bile, geçmişte gözlemlenmiş Flynn etkisinin sebebi her ne olursa olsun henüz kazançların azaldığı noktaya gelmedik. Gerileme ya da tersine dönme daha da büyük bir gözlemlenmiş gerileme üretebilecek olan bağımsız bir zararlı etkenden kaynaklanıyor olabilir.

4 Bostrom ve Roache (2011).

5 Somatik gen terapisi olgunlaşmadaki gecikmeyi ortadan kaldırabilir, ama teknik açıdan hücre hattı müdahalelerinden daha zorludur ve nihai potansiyeli daha düşüktür.

6 1960–2000 döneminde yıllık ortalama küresel ekonomik üretkenlik %4,3'tü (Isaksson 2007). Üretkenlikteki bu büyümenin sadece bir kısmı örgütsel verimlilikteki kazanımlardan kaynaklanmaktadır. Bazı tikel ağlar veya örgütsel süreçler elbette çok daha hızlı oranlarda ilerlemektedir.

7 Biyolojik beynin evrimi zihin dijital bir araca kaydırıldığında ciddi ölçüde rahatlayan birçok kısıtlamaya ve ödünleşime tabiydi. Örneğin beyin büyüklüğü kafa büyüklüğü tarafından sınırlandırılmıştır ve çok büyük bir kafa, doğum kanalından geçmekte sıkıntı yaşar. Keza büyük bir beyin metabolik kaynakları da çarçur eder ve hareketi engelleyen bir boş ağırlıktır. Belli beyin bölgeleri arasındaki bağlantılılık sterik

kısıtlamalar nedeniyle sınırlanabilir: Ak madde hacmi gri madde hacminden çok daha büyüktür. Kan akışı ısı yayılımını sınırlar ve kabul edilebilir işleyiş için üst sınıra yakın olabilir. Ayrıca biyolojik nöronlar gürültülüdür, yavaştır ve gliyal hücreler ve kan damarlarından sürekli koruma, bakım ve yeniden takviyeye ihtiyaç duyarlar (bu da kafatasının içinde çapraşıklığa katkıda bulunur). Bkz. Bostrom ve Sandberg (2009b).

8 Yudkowsky (2008a, 326). Daha yakın döneme ait bir değerlendirme için, bkz. Yudkowsky (2013).

9 Lin vd. (2012).

10 Kolektif zekâda salt kurucu zekâlarının sayısını artırarak belli bir artış mümkündür. Bunu yapmak kolaylıkla paralelleştirilebilecek olan görevlerde en azından daha iyi bir performans göstermeyi mümkün kılmalıdır. Gelgelelim bu tür bir nüfus patlamasından tam geri dönüşü almak için kurucular arasında belli bir (asgariden çok) eşgüdüm düzeyine de ulaşmak gerekecektir.

11 Zekânın hız ve niteliği arasındaki ayrım nöromorfik olmayan yapay zekâ sistemlerinde bir şekilde bulanıklaşmaktadır.

12 Rajab vd. (2006, 41–52).

13 Genel amaçlı işleyicilerden ziyade ayarlanabilir entegre devreler (FPGA) kullanmanın, nöral ağ simülasyonlarındaki bilgi işlem hızlarını iki katına kadar artırabileceği iddia edilmiştir (Markram 2006). Nitekim petaFLOP düzeyinde yüksek çözünürlüklü bir iklim modelleme çalışmasında; gömülü işlemci çiplerinden birini kullanarak güç gerekliliklerinde yaklaşık iki kat bir azalma ve maliyetlerde 24 ila 34 kat bir azalma tespit edilmiştir (Wehner vd. 2008).

14 Nordhaus (2007). Moore yasasının farklı anlamlarına dair birçok değerlendirme vardır, örneğin bkz. Tuomi (2002) ve Mack (2011).

15 Eğer gelişim yeterince yavaşsa, proje bu süre zarfında dış dünyada kaydedilen ilerlemeden (üniversitelerdeki araştırmacıların bilgisayar biliminde kaydettikleri ilerlemeler ve yarı iletken sanayisinin donanımda kaydettiği ilerlemelerden) istifade edebilir.

16 Algoritmik sarkma ihtimali sanırım daha azdır, ama istisnalardan biri kuantum hesaplaması gibi egzotik bir donanımın öncesinde mümkün olmayan algoritmaları çalıştırmak üzere mevcut hale gelmesidir. Nöral ağların ve derin makine öğreniminin algoritma sarkmasına örnek teşkil ettiği de savunulabilir: İlk icat edildiğinde hesaplama açısından çok pahalıydı ve bir süre rafa kaldırıldı, ardından hızlı grafik işleme birimleri çalıştırmayı kolay hale getirince çöpten geri çıkarıldı. Şimdi yarışlar kazanıyor.

17 Üstelik insan taban çizgisine doğru ilerleme yavaş olsa bile.

18 Ekonomik büyüme teorisi çerçevesinde bir zekâ patlaması fikri geliştirmek için muvakkat girişimlerde bulunulmuştur; örneğin bkz. Hanson (1998b); Jones (2009); Salamon (2009). Bu çalışmalar dijital zihinlerin gelmesiyle son derece hızlı bir büyüme potansiyeli olduğuna işaret etmiştir, ama endojen büyüme teorisi tarihsel ve çağdaş uygulamalar için bile görece az gelişmiş olsa da, potansiyel açıdan süreksiz bir gelecek bağlamına uygulanmasını bu aşamada yetkili öngörülerde bulunması muhtemel bir egzersizden ziyade potansiyel açıdan yararlı kavramların ve mülahazaların bir kaynağı olarak görmek daha iyidir. Bir teknolojik tekilliği

matematiksel açıdan modelleme girişimleri hakkında bir değerlendirme için, bkz. Sandberg (2010).

19 Elbette hiçbir havalanma olmaması da mümkündür. Ama daha önce savunulduğu üzere, süper zekâ teknik açıdan uygulanabilir göründüğünden, havalanma olmaması varoluşsal felaket gibi bir engelleyicinin müdahalesinden kaynaklanabilir. Eğer güçlü süper zekâ yapay zekâ şeklinde ya da tüm beyin emülasyonu şeklinde değil de, yukarıda ele aldığımız yollardan biriyle gelirse, o zaman yavaş havalanma daha muhtemeldir.

BEŞİNCİ BÖLÜM

1 Bir yazılım aklı dünya çapında bilgisayar ağlarından farklı olarak tek bir makinede çalışabilir, ama bizim "yoğunlaşma" derken kastettiğimiz bu değildir. Bunun yerine, burada ilgilendiğimiz şey gücün, özellikle de teknolojik yetenekten çıkartılan gücün makine zekâsı devriminin ileri aşamalarında ya da hemen akabinde ne ölçüde temerküz edeceğidir.

2 Örneğin tüketim ürünlerinin teknoloji yayınımı gelişmekte olan ülkelerde genellikle daha yavaştır (Talukdar vd. 2002). Ayrıca bkz. Keller (2004) ve Dünya Bankası (2008).

3 Şirket teorisiyle ilgilenen iktisat teorisi mevcut değerlendirme için bir karşılaştırma yapmak adına önemlidir. En bilinen örnek: Coase (1937). Ayrıca örneğin bkz. Canback vd. (2006); Milgrom ve Roberts (1990); Hart (2008); Simester ve Knez (2002).

4 Diğer yandan, bir tohum yapay zekâyı çalmak özellikle kolay olabilir, zira elektronik olarak aktarılabilecek ya da bir portatif bellek aygıtında taşınabilecek yazılımdan oluşur.

5 Eğer durumu projeler arasındaki gecikme süresinin normal bir dağılımdan alındığı bir durum olarak modellersek, o zaman önde giden proje ile en yakın takipçisi arasındaki muhtemel mesafe kaç tane proje olduğuna da bağlı olacaktır. Eğer çok sayıda proje varsa, ilk ikisi arasındaki mesafe, dağılım çeşitliliği görece yüksek olsa bile muhtemelen azdır (ama eğer tamamlanma zamanları normal şekilde dağıtılırsa, birinci proje ile ikinci proje arasındaki beklenen mesafe çok yavaş azalır). Gelgelelim hepsi de başa güreşecek kadar iyi kaynaklar bulan çok sayıda proje olması pek muhtemel değildir. (Eğer peşine düşülecek farklı temel yaklaşım sayısı çoksa, çok sayıda proje olabilir, ama bu durumda bu yaklaşımların çoğunun çıkmaza girmesi muhtemeldir.) Belirtildiği üzere, ampirik açıdan tek bir somut teknolojik hedefin peşinden koşan ciddi rakiplerin sayısının genellikle bir avuçtan fazla olmadığını görüyor gibiyiz. Çok az farklı ürünler için birçok nişin olduğu ve giriş önündeki engellerin düşük olduğu bir tüketici piyasasında durum biraz farklıdır. Tişört tasarlayan pek çok tek kişilik proje vardır, ama video kartlarında bir sonraki nesli geliştiren çok az şirket vardır. (İki şirket AMD ve NVIDIA, şu anda birlikte bir tekel konumundadır, ama Intel de kıyısından köşesinden rekabete dahil olmuştur.)

6 Bostrom (2006c). Varlığı görünmez olan bir tekel (örneğin müdahaleleri bir insan tarafından anlaşılmadan dünyadaki olayları incelikle kontrol edebilecek kadar ileri bir teknolojiye ya da içgörüye sahip bir süper zekâ) ya da kendi güç uygulamasına çok sıkı

sınırlamaları bilerek dayatan bir tekel (örneğin anlaşmayla saptanmış belli uluslararası kurallara ya da liberter ilkelere saygı duyulmasını sağlayacak şekilde kendisini titizlikle sınırlayan bir tekel) tahayyül edilebilir. Belli türde bir tekelin ortaya çıkma ihtimalinin ne kadar muhtemel olduğu elbette ampirik bir sorudur; ama en azından *kavramsal açıdan*, iyi bir tekel, kötü bir tekel, ele avuca sığmaz derecede farklı bir tekel, mülayim bir yekpare tekel, baskıcı bir tekel ya da bağırıp çağıran bir despottan ziyade bir başka doğa yasasına benzer bir tekel mümkündür.

7 Jones (1985, 344).

8 Manhattan Projesi'nin savaş sırasında yürütülmüş olması kayda değer olabilir. Bu işe girişen pek çok bilimci esasen savaş durumundan ve Nazi Almanyası'nın Müttefik güçlerden önce atom bombası geliştirebileceği korkusundan etkilendiğini iddia etmişti. Birçok hükümetin barış döneminde benzer derecede yoğun ve gizli bir çabayı seferber etmesi zor olabilir. Bir diğer sembolik bilim/mühendislik mega projesi olan Apollo programı Soğuk Savaş rekabetinden güçlü bir itki almıştı.

9 Çok tehlikeli olduğunu düşünseler bile, bunu böyle gördüklerini (açıktan) belli etmeyeceklerdir denilebilir.

10 Şifreleme teknikleri işbirliği yapan takımın fiziksel olarak dağılmasını mümkün kılabilir. İletişim zincirindeki tek zayıf halka, fiziksel yazma ediminin potansiyel açıdan gözlemlenebildiği girdi aşaması olabilir. Fakat (mikroskopik kayıt cihazları yoluyla) iç mekân dinlemesi yaygın hale gelirse, mahremiyetlerini korumaya meraklı olanlar karşı önlemler geliştirebilirler (örneğin gizlice dinleyen aygıtlardan yalıtılacak özel odalar). Fiziksel uzam yaklaşan bir dinleme-gözetleme çağında şeffaf hale gelebilirken, siber-uzayın güçlü şifreleme protokollerinin daha geniş çapta benimsenmesi yoluyla korunması mümkündür.

11 Bir totaliter devlet daha da zorlayıcı tedbirlere başvurabilir. İlgili alanlardaki bilimciler, Stalinist Rusya'daki "akademik köyler"e benzer şekilde tutulup kamplara tıkılabilir.

12 Halktaki kaygı düzeyi görece düşük olduğunda, bazı araştırmacılar birazcık korku tellallığı yapılmasından hoşlanabilirler, çünkü dikkatleri kendi çalışmalarına çeker ve çalıştıkları alanın önemli ve heyecan verici görünmesini sağlar. Kaygı düzeyi arttığında, ilgili araştırma toplulukları fon kesintileri, düzenleme ve halkın tepkileri konusunda endişelenmeye başlayınca başka telden çalabilirler. Komşu disiplinlerdeki araştırmacılar (bilgisayar biliminin ve robot biliminin yapay genel zekâyla çok ilintili olmayan kısımları gibi) fonun ve dikkatlerin kendi araştırma alanlarından uzaklaştırılmasından rahatsızlık duyabilirler. Bu araştırmacılar *kendi* çalışmalarının tehlikeli bir zekâ patlamasına yol açma riskinin olmadığı gibi doğru bir gözlemde bulunabilirler. (Nanoteknoloji fikrinin tarihiyle bazı tarihsel paralellikler kurulabilir; bkz. Drexler [2013].)

13 Bunlar yapmaya giriştikleri işlerin en azından bir kısmını başarmış olmaları anlamında başarılı olmuştur. Geniş bir anlamda ne kadar başarılı olduklarını (ne kadar uygun maliyetli olduğunu vb. hesaba katarak) belirlemek zordur. Örneğin Uluslararası Uzay İstasyonu vakasında, muazzam bir maliyet fazlalığı ve gecikmeler olmuştur. Projenin karşılaştığı problemlerin ayrıntıları hakkında, bkz. NASA (2013). Büyük Hadron Çarpıştırıcısı projesi de bazı büyük tersliklerle karşılaşmıştır, ama bunun nedeni görevin içsel zorluğu olabilir. İnsan Genomu Projesi sonunda başarıya ulaştı, ama Craig Venter'ın özel teşebbüs çabasıyla rekabete zorlanmış olmasının bir itki sağladığı

görünüyor. Kontrollü füsyon enerjisi elde etmeyi amaçlayan uluslararası projeler muazzam yatırımlara karşın beklentileri karşılayamamıştır; ama bir kez daha, bunun da beklenenden daha zor bir görev olmaktan kaynaklandığı söylenebilir.

14 ABD Kongresi, Teknoloji Değerlendirme Ofisi (1995).

15 Hoffman (2009); Rhodes (2008).

16 Rhodes (1986).

17 ABD donanmasının şifre kırma örgütü OP-20-G, Britanya'nın Enigma karşıtı yöntemlerine dair tam bir bilgi edinme davetini belli ki umursamamış ve ABD'deki üst düzey karar alıcıları Britanya'nın şifreleme sırlarını paylaşma teklifi konusunda bilgilendirmeyi atlamıştı (Burke 2001). Bu durum Amerikalı liderlerde Britanya'nın önemli bilgileri sakladığı izlenimi uyandırmıştı ve bu da savaş boyunca sürtüşmeye yol açmıştı. Britanya Alman telsizlerinden deşifre ettiği istihbaratın bir kısmını Sovyet hükümetiyle gerçekten de paylaşmıştı. Örneğin Rusya Almanya'nın Barbarossa Harekâtı hazırlıkları konusunda uyarılmıştı. Ama Stalin uyarılara inanmayı reddetmişti ve bunun bir nedeni de Britanya hükümetinin bilgiyi nasıl elde ettiğini söylememiş olmasıydı.

18 Russell birkaç yıl boyunca Rusya'yı Baruch planını kabul etmeye ikna etmek için nükleer savaş tehdidini savunmuş görünmektedir; ama sonradan, karşılıklı nükleer silahsızlanmanın güçlü bir savunucusu haline gelmiştir (Russell ve Griffin 2001). John von Neumann'ın ABD ile Rusya arasında bir savaşın kaçınılmaz olduğuna inandığı ve "Eğer onları [Rusları] neden yarın bombalamayalım derseniz, neden bugün bombalamayalım derim. Eğer bugün beşte derseniz, neden saat birde derim?" dediği iddia edilir. (Bu meşum sözü McCarthy çağında ABD Savunma Bakanlığındaki şahinler nezdindeki anti-komünist itibarını güçlendirmek için sarf etmiş olması mümkündür. Eğer ABD politikalarının başında von Neumann olsaydı bir saldırı gerçekleştirir miydi, söylemek imkânsızdır. Bkz. Blair [1957, 96].)

19 Baratta (2004).

20 Eğer yapay zekâ bir insan grubunun kontrolünde olursa, problem bu insan grubuna uygulanabilir, ama bu zamana gelene kadar güvenilir şekilde bir anlaşmaya bağlılık göstermenin yeni yollarının bulunması mümkündür ve bu durumda insan grupları bile potansiyel iç çözülme ve bir alt-koalisyon tarafından devrilme probleminden sakınabilirler.

ALTINCI BÖLÜM

1 İnsanlık hangi anlamda yeryüzünde egemen bir türdür? Ekolojik açıdan insanlar en yaygın büyük (~50 kg) hayvandır, ama toplam insan kuru biyokütlesi (~100 milyar kg) karıncalarınkiyle, Formicidae familyasıyla kıyaslandığında o kadar da etkileyici değildir (300 milyar–3.000 milyar kg). İnsanlar ve insan(dan) yararlanan organizmalar toplam küresel biyokütlenin çok küçük bir kısmını (<0.001) oluşturur. Gelgelelim ekin tarlaları ve otlaklar şu anda gezegendeki en büyük ekosistemler olup buzsuz yeryüzünün yaklaşık %35'ini kaplamaktadır (Foley vd. 2007). Ayrıca tipik bir değerlendirmeye göre, net birincil verimliliğin neredeyse dörtte birini temellük ediyoruz (Haberl vd. 2007), ama tahminler esasen ilintili terimlerin değişen tanımlarına bağlı olarak %3 ile %50 arasında farklılık göstermektedir (Haberl vd.

2013). İnsanlar yeryüzünde en fazla yer kaplayan hayvandır ve en fazla sayıda farklı besin zincirinin üstünde bulunmaktadır.

2 Zalasiewicz vd. (2008).

3 Bu bölümün ilk sonnotuna bakınız.

4 Aslında bu pek doğru olmayabilir. İnsan türünde zekâ çeşitliliği neredeyse sıfıra kadar gitmektedir (örneğin embriyolarda ya da kalıcı bitkisel hayatta olan hastalarda). Niteliksel açıdan, insan türünde bilişsel yetenekteki azami farklılık belki de herhangi bir insan ile bir süper zekâ arasındaki farktan daha fazladır. Ama eğer "insan"ı "fonksiyonları normal olan bir yetişkin" diye okursak metinde kastedilen şey yanlış olmaktan çıkar.

5 Gottfredson (2002). Ayrıca bkz. Carroll (1993) ve Deary (2001).

6 Bkz. Legg (2008). Kabaca, Legg tüm ödül-hesaplanabilir çevrelerde beklenen performansı olarak bir takviyeli öğrenme aracısını ölçmeyi teklif eder, burada bu türden her bir çevre Kolmogorov karmaşıklığının belirlediği bir ağırlığa sahiptir. Takviyeli öğrenmeyle ne kastedildiğini 12. Bölüm'de açıklayacağız. Ayrıca bkz. Dowe ve Hernandez-Orallo (2012) ve Hibbard (2011).

7 Biyoteknoloji ve nanoteknoloji gibi alanlarda teknoloji araştırmasına ilişkin olarak, bir süper zekânın ne konuda üstün olacağı yeni yapıların tasarımı ve modellenmesidir. Tasarım dehası ve modellemesi fiziksel deneyleri ikame edemediği ölçüde, süper zekânın performans üstünlüğü gerekli deneysel aygıta erişim düzeyi tarafından sınırlanabilir.

8 Örneğin Drexler (1992, 2013).

9 Alanı dar bir yapay zekânın elbette önemli ticari uygulamaları olabilir, ama bu demek değildir ki ekonomik verimlilik süper gücüne sahip olacaktır. Örneğin alanı dair bir yapay zekâ, sahiplerine yılda birkaç milyar dolar kazandırsa bile, bu yine de dünya ekonomisinin geri kalanından dört kat daha düşük olacaktır. Sistemin dünya ürünlerini doğrudan ve ciddi ölçüde artırması için bir yapay zekânın birçok farklı iş yapabilmesi gerekir, yani birçok alanda yeterliliğinin olması gerekir.

10 Bu kıstas yapay zekânın başarısız olduğu tüm senaryoları dışlamaz. Örneğin yapay zekâ başarısız olma ihtimali yüksek olan bir kumar oynayabilir. Fakat bu örnekte, kıstas şu biçimi alabilir: (a) yapay zekâ kumarın düşük başarı şansına ilişkin tarafsız bir tahminde bulunmalıdır ve (b) biz bugünkü insanların düşünebildiği ama yapay zekânın gözünden kaçırdığı daha iyi bir kumar olmamalıdır.

11 Karş. Freitas (2000) ile Vassar ve Freitas (2006).

12 Freitas (1980); Freitas ve Merkle (2004, 3. Bölüm); Armstrong ve Sandberg (2013).

13 Örneğin bkz. Huffman ve Pless (2003), Knill vd. (2000), Drexler (1986).

14 Burada ayak bağı olabilecek Dünya dışı medeniyetler olmadığını varsayıyoruz. Ayrıca simülasyon hipotezinin yanlış olduğunu varsayıyoruz. Bkz. Bostrom (2003a). Eğer bu varsayımlardan biri doğru değilse, insan kökenli olmayan önemli riskler (insani olmayan türde bir zeki aracılığı içeren bir risk) olabilir. Ayrıca bkz. Bostrom (2003b, 2009c).

15 En azından evrim fikrini kavrayan bir akıllı tekel, ilkesel olarak, bir öjeni programına girişebilir ve bu yolla kendi kolektif zekâ düzeyini yavaşça artırabilirdi.

16 Tetlock ve Belkin (1996).

17 Açık konuşmak gerekirse: Erişilebilir evrenin büyük bir kısmına yerleşmek ve yeniden mühendisliğini yapmak şu an bizim doğrudan menzilimiz dahilinde değildir. Galaksiler arası kolonizasyon bugünün teknolojisinin çok ötesindedir. Mesele şu ki, ilkesel olarak, mevcut yeterliliklerimizi ihtiyaç duyulacak ek yeterlilikleri geliştirmek için kullanabiliriz ve dolayısıyla başarmak *dolaylı olarak* menzilimiz dahilindedir. İnsanlığın şu an bir tekel oluşturmadığı ve eğer erişilebilir evrende yeniden mühendislik yapmaya başlarsak bir dış güçten zeki karşı çıkışla asla karşılaşmayacağımızı bilmediğimiz elbette doğrudur. Gelgelelim bilge-tekel sürdürülebilirlik eşiğini tutturmak için eğer akıllı bir karşı çıkışla karşılaşmayan bir bilge tekel yeterliliklere sahipse, o zaman erişilebilir evrenin büyük bir kısmının kolonizasyonunun ve yeniden mühendisliğinin onun dolaylı menzili dahilinde olacağı şekilde yeterliliklere sahip olmak kâfidir.

18 Bazen iki yapay zekânın da verili bir süper güce sahip olduğundan bahsetmek yeterli olabilir. Kelimenin geniş anlamında, bir süper gücü bir aracının belli bir eylem alanına ilişkin sahip olduğu bir şey olarak tasavvur edebiliriz – bu örnekte belki de tüm insan medeniyetini içeren ama diğer yapay zekâyı dışlayan bir alandır.

YEDİNCİ BÖLÜM

1 Bu elbette görsel olarak küçük gözüken farklılıkların işlevsel açıdan derin olabileceğini reddetmek değildir.

2 Yudkowsky (2008a, 310).

3 İskoç Aydınlanma filozofu David Hume, tek başına inancın (sözgelimi neyi yapmanın iyi bir şey olduğu inancının) bir eylemin güdüsü olamayacağını savunmuştu; eylem için istek de gerekir. Bu görüş diklik tezini, yöneltilen muhtemel itirazlardan birini (yeterli zekânın belli inançlar edinilmesini gerektirebileceği ve bunun da akabinde zorunlu olarak belli güdüler doğuracağı itirazını) zayıflatarak destekleyecektir. Gelgelelim diklik tezi Hume'un güdülenme teorisinden destek alabilirse de, bunu varsaymaz. Daha özelde, sadece inancın asla eylemi güdüleyemeyeceğini savunmak gerekmez. Örneğin bir aracının –çok zeki diyelim– yeterli, baskın kuvvette belli arzuları varsa herhangi bir eylemin peşinden gidecek şekilde güdülenebileceğini varsaymak yeterlidir. Hume'un güdülenme teorisi yanlış olsa bile, eğer keyfi derecede yüksek zekâ güya kendi başına güdüleyen türde inançlar edinilmesini gerektirmiyorsa, diklik tezinin doğru olabileceğini varsayabiliriz. Hume'un güdülenme teorisi yanlış olsa bile diklik tezinin doğru olabileceği üçüncü bir yol ise, keyfi derecede yüksek zekâya sahip olan ama insanlarda "inanç" ve "arzu" adını verdiğimiz şeye açık bir işlevsel analog içermeyecek kadar yabancı bir yapıya sahip aracı (ya da daha nötr şekliyle, bir "optimizasyon süreci") inşa etmektir. (Son dönemde Hume'un güdülenme teorisini savunma girişimlerinden bazıları için, bkz. Smith [1987], Lewis [1988] ve Sinhababu [2009].)

4 Sözgelimi Derek Parfit bazı temel tercihlerin akıldışı olacağını savunmuştur, "Gelecek-Salı-Umursamazlığı" olan diğer açılardan normal bir aracınınki gibi:

"Belli bir hedonist gelecekteki deneyimlerinin niteliğine büyük önem verir. Tek bir istisnayla, geleceğinin tüm parçalarına aynı derecede önem verir. İstisna, Gelecek-

Salı-Umursamazlığı'na sahip olmasıdır. Salı günleri başına gelenleri normal bir tarzda umursar. Ama müstakbel bir salıdaki mümkün acıları ya da hazları asla umursamaz. [...] Bu umursamazlık yalın bir olgudur. Geleceğini planlarken, bir salı günü büyük ıstırap olasılığını başka bir günkü en ufak acıya her zaman tercih ettiği doğrudur." [Parfit (1986, 123–4); ayrıca bkz. Parfit (2011)]

Buradaki amaçlarımız açısından, Perfit'in bu aracının akıldışı olduğu görüşünde haklı olup olmadığına dair bir tutum almamız gerekmiyor, yeter ki metinde açıklanan araçsal anlamda mutlaka zeki olmadığını kabul edelim. Parfit'in aracısı tamamen akli bir aracıdan beklenebilecek "nesnel akla" bir tür duyarlılık konusunda yetersiz kalsa da, mükemmel bir araçsal akliliğe ve dolayısıyla büyük zekâya sahip olabilir. Dolayısıyla bu örnek diklik tezini çürütmez.

5 Tamamen akli bir aracının anlayacağı nesnel ahlaki olgular olsa bile ve bu ahlaki olgular bir şekilde içsel olarak güdüleyici olsa bile (şöyle ki bunları tamamen anlayan biri ister istemez bunlara uygun olarak hareket edecek şekilde güdülenecektir), bunun diklik tezini çürütmesi gerekmiyor. Eğer bir aracı akliliği oluşturan başka bir yetiden ya da nesnel ahlaki olguları tamamen anlamak için gerekli bir yetiden yoksunken mükemmel bir *araçsal* akliliğe sahip olabiliyorsa, tez yine de doğru olabilir. (Ayrıca bir aracı her alanda tam araçsal akliliğe sahip olmadan aşırı zeki, hatta süper zeki olabilir.)

6 Diklik tezi hakkında daha fazlası için, bkz. Bostrom (2012) ve Armstrong (2013).

7 Sandberg ve Bostrom (2008).

8 Stephen Omohundro bu konuda iki çığır açıcı yazı kaleme almıştır (Omohundro 2007, 2008). Omohundro tüm ileri yapay zekâ sistemlerinin birtakım "temel dürtüler" sergilemesinin muhtemel olduğunu savunur ve "temel dürtüler" derken, "açıkça karşı çıkılmadıkça kendini gösterecek olan eğilimleri" kasteder. "Yapay zekâ dürtüsü" terimi kısa ve çağrıştırıcı olma gibi bir üstünlüğe sahiptir, ama göndermede bulunduğu araçsal hedeflerin yapay zekânın karar alma sürecini tıpkı psikolojik dürtülerin insanın karar alma sürecini etkilemesi gibi, yani istenç gücümüzün ara ara direnmeyi başardığı egomuz üzerinde bir tür fenomenolojik asılma yoluyla etkilediğini söylemesiyse mahzurludur. Bu yan anlam yararsızdır. Normalde bir insanın vergi beyannamesini doldurma "dürtüsü" olduğunu söylemeyiz, ama vergi beyannamelerini doldurmak günümüz toplumlarında insanlar için epey yakınsak bir araçsal hedef olabilir (bu hedefin gerçekleştirilmesi nihai hedeflerimizden birçoğunu gerçekleştirmekten bizi alıkoyacak olan bir belayı defeder). Buradaki değerlendirmemiz Omohundro'nun değerlendirmesinden başka birtakım önemli açılardan da ayrılmaktadır ama altta yatan fikir aynıdır. (Ayrıca bkz. Chalmers [2010] ve Omohundro [2012].)

9 Chislenko (1997).

10 Ayrıca bkz. Shulman (2010b).

11 Bir aracı ontolojisini değiştirdiği takdirde hedef *temsilini* de değiştirebilir, zira bu sayede eski temsilini yeni ontolojiye aktarır; karş. de Blanc (2011).

Kanıtsal karar teorisyeninin çeşitli eylemlere (nihai hedeflerini değiştirmek dahil) girişmeye itebilecek bir diğer tür etken de bunu yapmaya karar vermenin kanıtsal mahiyetidir. Örneğin kanıtsal karar teorisinin peşinden giden bir aracı evrende onun gibi başka aracılar olduğuna ve kendi eylemlerinin bu diğer aracıların nasıl hareket

edeceğine dair bazı kanıtlar sunacağına inanabilir. Bu nedenle aracı diğer kanıtsal olarak bağlantılı aracılara karşı diğerkâm olan bir nihai hedef benimsemeyi seçebilir, zira bu sayede aracı bu diğer aracıların benzer bir tarzda hareket etmeyi seçeceği kanıtı sunacaktır. Gelgelelim her bir örnekte sanki bu nihai hedeflere sahipmiş gibi hareket etmeyi seçerek nihai hedefleri değiştirmeden eşdeğer bir sonuç elde edilebilir.

12 Kapsamlı bir psikolojik yazın uyumlu tercih oluşumunu irdeler. Örneğin bkz. Forgas vd. (2010).

13 Formel modellerde bilginin değeri bu bilgiyle alınan optimum kararların gerçekleştirdiği beklenen değer ile bu bilgi olmadan alınan optimum kararların gerçekleştirdiği beklenen değer arasındaki fark olarak nicelleştirilir. (Örneğin bkz. Russell ve Norvig [2010].) Bu demektir ki, bilginin değeri asla olumsuz değildir. Keza alıp alacağınız herhangi bir kararı asla etkilemeyeceğini bildiğiniz bir bilginin sizin için sıfır değeri var demektir. Gelgelelim bu model türü gerçek dünyada çoğu zaman geçersiz olan birkaç idealleştirme benimser – öyle ki bilginin hiçbir nihai değeri yoktur (yani bilginin sadece araçsal değeri vardır ve başlı başına değere sahip değildir) ve aracılar diğer aracılar için şeffaf değildir.

14 Örneğin Hajek (2009).

15 Bu stratejiye örnek, uygun bir kaya bulana kadar yüzüp duran ve sonra ilelebet o kayaya tutunan deniz üzümü larvasıdır. Olduğu yere yapışan bu larvanın karmaşık bilgi işlemeye ihtiyacı daha azdır ve bu andan itibaren kendi beyninin bir kısmını (*serebral ganglion*) yemeye başlar. Benzer bir olguyu kadro kapan akademisyenlerde de gözlemleyebiliyoruz.

16 Bostrom (2012).

17 Bostrom (2006c).

18 Soruyu tersine çevirip bunun yerine süper zeki bir tekelin bazı teknolojik yeterlilikleri *geliştirmemesinin* muhtemel nedenlerine bakabiliriz. Bu sebepler arasında: (a) Tekel o yeterliliğe ihtiyaç duymayacağını öngörür; (b) gelişim maliyeti beklenen faydaya kıyasla çok büyüktür (örneğin teknoloji tekelin amaçlarından herhangi birini gerçekleştirmek için asla uygun olmayacaksa ya da tekel, yatırımı ciddi şekilde caydıran çok yüksek bir indirim haddine sahipse); (c) tekel teknoloji gelişiminin tikel kanallarından sakınmayı gerektiren bir nihai değere sahiptir; (d) tekel durağan kalacağından emin değilse, iç istikrarını tehdit edebilecek ya da dağılmanın sonuçlarını daha kötü hale getirecek teknolojiler geliştirmekten uzak durmayı tercih edebilir (sözgelimi bir dünya hükümeti isyanı kolaylaştıracak teknolojiler –bunların bazı yararları olsa bile– geliştirmek ya da dünya hükümeti dağılacak olsa müthiş yıkıma yol açacak olan kitle imha silahlarını kolayca üretecek teknolojiler geliştirmek istemeyebilir); (e) benzer şekilde, tekel bir teknoloji geliştirmemek için bir tür bağlayıcı stratejik taahhütte bulunmuş olabilir; şu an geliştirilmesi uygun olmasa bile bu taahhüt işlemeye devam eder. (Fakat teknoloji gelişimi için bazı *cari* nedenler bir tekel için geçerli *olmayacaktır*: Sözgelimi silahlanma yarışından doğan nedenler.)

19 Bir aracı gelecekte edinilen kaynakları üssel hızda ucuza mal ediyor ve ışık hızı sınırlamasından ötürü aracının kaynak birikimini çokterimli hızda artırabiliyor olsun. Buradan aracının edinimsel genişlemeyi artık gereksiz bulacağı bir noktanın geleceği sonucu çıkar mı? Hayır, çünkü gelecek zamanda edinilen kaynakların mevcut değeri biz geleceğe baktıkça sıfıra kavuşmaz olsa da, *aynısı bunları edinmenin mevcut*

maliyetleri için de geçerlidir. Şimdiden 100 milyon yıl sonraya bir von Neumann sondası daha yollamanın mevcut maliyeti (muhtemelen kısa zaman önce edinilmiş bir kaynaktan yararlanarak) fazladan sondanın edineceği (sabit bir etkeni dışarıda bırakarak) gelecekteki kaynakların mevcut değerini azaltacak olan aynı iskonto etkeniyle azalacaktır.

20 Belli bir zamanda kolonizasyon sondalarının ulaştığı hacim kabaca küre şeklinde olmakla birlikte ve ilk sonda kullanılmaya başladıktan ($\sim t^2$) sonra geçen zamanın karesiyle orantılı bir hızda genişlemesine karşın, bu hacimde içerilen kaynak miktarı daha düzensiz bir büyüme örüntüsü izleyecektir, zira kaynak dağılımı homojen değildir ve birkaç ölçekte çeşitlilik gösterir. Başlangıçta büyüme hızı ana gezegen kolonize edilirken $\sim t^2$ olabilir; ardından büyüme hızı yakındaki gezegenler ve güneş sistemleri kolonize edilirken fırlayabilir; ardından, Samanyolu'nun kabaca disk şeklindeki hacmi doldukça, büyüme hızı eşitlenebilir, takriben t ile orantılı hale gelebilir; ardından büyüme hızı yakındaki galaksiler kolonileştirildikçe iyice fırlayabilir; ardından genişlemenin galaksilerin dağılımının kabaca türdeş olacağı bir ölçeğin ötesindeyken büyüme hızı yeniden $\sim t^2$'ye yaklaşabilir; ardından galaktik süper kümeler kolonileştirilirken bir diğer hızlı büyüme döneminin akabinde sorunsuz $\sim t^2$ büyümesi yaşanır; en sonunda büyüme hızı nihai bir gerileme yaşar ve nihayet evrenin genişleme hızı daha fazla kolonileşmeyi imkânsız hale getirecek kadar arttığında sıfıra ulaşır.

21 Simülasyon argümanı bu bağlamda bilhassa önemli olabilir. Bir süper zeki aracı hipotezlere kayda değer bir olasılık atfedebilir; buna göre, bir bilgisayar simülasyonunda yaşar ve algı dizilimi başka bir süper zeka tarafından üretilir. Bu durum ne tür simülasyonların devrede olmasının en çok muhtemel olduğuna dair aracının tahminine bağlı olarak çeşitli yakınsak araçsal nedenler yaratabilir. Karş. Bostrom (2003a).

22 Temel fizik yasalarını ve dünya hakkındaki diğer temel olguları keşfetmek bir yakınsak araçsal hedeftir. Bunu burada "bilişsel geliştirme" başlığı altında ele alabiliriz, ama "teknolojiyi kusursuzlaştırma" hedefinden de çıkartılabilir (zira yeni fiziksel olgular yeni teknolojiyi kolaylaştırabilir).

SEKİZİNCİ BÖÜM

1 İnsanlığın son derece altoptimal durumda hayatta kaldığı ya da istenen gelişim potansiyelimizin büyük bir kısmının geri dönüşsüz şekilde çarçur edildiği senaryolarda, ek birtakım varoluşsal riskler vardır. Üstelik örneğin süper zekâyı daha önce geliştirmek için rekabet eden ülkeler arasındaki savaştan doğan potansiyel bir zekâ patlamasına giden yolla bağlantılı varoluşsal riskler de olabilir.

2 Yapay zekâ bu tür bir saklama ihtiyacını ilk fark ettiğinde önemli bir yaralanabilirlik anı vardır (bizim *aldatma tasavvuru* adını verebileceğimiz bir olay). Bu ilk farkındalık gerçekleştiğinde kasten saklanmayacaktır. Ama yapay zekâ bu farkındalığa sahip olunca hızla davranıp farkındalığın gerçekleştiği olgusunu saklarken, bir yandan da uzun vadeli stratejisini gizli kapaklı sürdürmeye devam etmesini sağlayacak gizli bir iç dinamik kurabilir (belki de onun zihninde gerçekleşen diğer karmaşık süreçlerin hepsiyle karışan zararsız bir süreç kılığına bürünür).

3 İnsan *hacker*lar bile tamamen beklenmedik şeyler yapan küçük ve görünüşte zararsız programlar yazabilirler. (Örneğin IOCCC'deki kazananlardan bazılarına bakınız.)

4 Bazı kontrol yöntemlerinin değişmez bir bağlamda işliyor gözükebildiğini ama bağlam değiştiğinde feci şekilde çuvalladığını Eliezer Yudkowsky de vurgular; örneğin bkz. Yudkowsky (2008a).

5 [*wireheading*] Bu kelimeyi ilk kez bilimkurgu yazarı Larry Niven (1973) kullanmıştır, ama gerçek dünyadaki beyin uyarılması ödül deneylerine dayalıdır; karş. Olds ve Milner (1954) ile Oshima ve Katayama (2010). Ayrıca bkz. Ring ve Orseau (2011).

6 Bostrom (1997).

7 Takviyeli öğrenme mekanizmasının yapay zekâ *wirehead* çözümünü keşfettiğinde altyapı bolluğundan ziyade güvenli bir acizliğe varacağı bir takviyeli öğrenme mekanizmasının bazı olanaklı uygulamaları olabilir. Fakat bu kolaylıkla sarpa sarıp beklenmedik nedenlerden ötürü başarısız olabilir.

8 Bunu dillendiren Marvin Minsky olmuştur (bilhassa bkz. Russell ve Norvig [2010, 1039]).

9 Hangi tür dijital zihinlerin öznel olgusal deneyime (filozofların diliyle "qualia"ya) sahip olma anlamında bilinçli olacağı meselesi bu husus açısından önemlidir (ama bu kitabın diğer birçok kısmıyla alakasızdır). Bir insanın, simülasyonun bilinçli olacağı kadar ayrıntılı şekilde beynini simüle etmeden çeşitli koşullarda nasıl davranacağını doğru tahmin etmenin ne kadar zor olacağı cevabı belirsiz bir sorudur. Bir diğer soru da bir süper zekâ için *qualia* yaratmayı sağlayacak kadar yararlı algoritmalar (sözgelimi takviyeli öğrenme teknikleri) olup olmadığıdır. Bu tür altprogramların bilinçli olma olasılığının son derece düşük olduğunu tespit etsek bile, örneklerle desteklemeler o kadar büyük olabilir ki, onların ıstırap da yaşayabileceklerine dair en ufak bir riske bile ahlaki hesaplamalarımızda ciddi bir ağırlık verilmelidir. Ayrıca bkz. Metzinger (2003, 8. Bölüm).

10 Bostrom (2002a, 2003a); Elga (2004).

DOKUZUNCU BÖLÜM

1 Örneğin Laffont ve Martimort (2002).

2 Seçmenlerin çoğunluğu ülkelerinde bir tür süper zekâ yaratılmasını istiyor olsun. İstediklerini yapacağını vaat eden bir aday seçiyorlar, ama onun bir kez iktidara geldiğinde kampanyasındaki vaadinin arkasında durup projeyi seçmenlerin niyetlendiği tarzda sürdüreceğinden emin olamayabilirler. Bu vekilin sözüne sadık olduğunu varsayalım. Hükümetine bu işi yerine getirmek için akademik ya da sınai bir konsorsiyumla sözleşme yapmasını söylüyor; ama bir kez daha aracılık problemleri vardır: Devlet bürokratlarının ne yapılması gerektiğine dair kendi görüşleri olabilir ve projeyi liderin talimatlarının ruhuna değil lafzına bağlı kalacak şekilde hayata geçirebilirler. Bürokratlar işlerini düzgün şekilde yapsalar bile, sözleşmeli bilimcilerin kendi gündemleri olabilir. Problem birkaç düzlemde karşımıza çıkar. Bir teknisyenin tarıma onaylanmamış bir öğeyi dahil etmesinden duyduğu korku nedeniyle katılımcı laboratuvarlardan birinin müdürünün uykuları kaçabilir. Dr. T.R. Eason'ın bir gece geç saatte gizlice ofisine dalıp proje kod temeline girerek tohum yapay zekânın hedef

sisteminin bir kısmını yeniden yazdığını tahayyül edebilir. "İnsanlığa hizmet et" demesi gereken yerde artık "Dr. T.R. Eason'a hizmet et" demektedir.

3 Süper zekâ gelişimi için bile (geniş güvenlik önlemlerinin içinde tali bir öğe olarak) davranışsal test için bir rol olabilir. Bir yapay zekâ gelişim aşamasında yanlış davranırsa, bir şeylerin ters gittiği açıktır, ama daha önemlisi tersi geçerli değildir.

4 Steven Dompier 1975 yılına ait klasik bir macerada, Altair 8800 için bu etkiden (mikro bilgisayarın örneği etrafında korumanın yokluğundan) yararlanan bir program yazmıştı. Programı çalıştırmak bir transistorlu radyoyu bilgisayarın yakınında tuttuğunda müzik üreten elektromanyetik dalgalarının yayılımına yol açmıştı (Driscoll 2012). Bir demoya katılmış olan Bill Gates, bu başarıdan etkilenmiş ve şaşırmıştı (Gates 1975). Her halükârda, gelecekteki çipleri yerleşik Wi-Fi özellikleriyle tasarlama planları vardır (Greene 2012).

5 Tüm kozmik birikimimizin mahvolmasına yol açabilecek bir görüşe sahip olmuş olmak hiç de önemsiz bir mesele değildir. Belki de şu ilke savunulabilir: Eğer geçmişte birisi sistemin güvenli hale gelecek kadar geliştirildiğini sayısız kez test ettiyse ve her seferinde yanıldıkları ortaya çıktıysa, o zaman bir sonraki sefere sistemin güvenli olduğuna $1/(N+1)$'den fazla güven atfetme hakları olamaz.

6 Gayri resmi bir deneyde yapay zekâ rolünü zeki bir insan oynamıştı. Bir başka bireyse geçit denetçisi rolünü oynamış ve yapay zekâyı kutudan çıkarmamakla görevlendirilmişti. Yapay zekâ geçit denetçisiyle ancak yazışarak iletişim kurabiliyordu ve denetçiyi ikna etmesi için sadece iki saat verilmişti. Farklı bireylerin denetçi rolünü oynadığı beş örneğin üçünde, yapay zekâ kaçmayı başarmıştı (Yudkowsky 2002). Bir insanın yapabileceğini süper zekâ da yapabilir. (Elbette tersi geçerli değildir. Gerçek bir süper zekâ için görev daha zor olsa bile –belki de geçit denetçilerinin yapay zekâyı salıvermeyi engelleme konusunda deneye katılan insan bireylerden daha hırslı olacaklardır– süper zekâ bir insanın başarısız olduğu yerde yine de başarılı olabilir.)

7 Bu yolla elde edilebilecek marjinal güvenlik miktarını abartmamak gerekir. Zihindeki görüntüler grafiksel gösterimi ikame edebilir. Kitapların insanlara yapabileceği etkileri düşünün, üstelik kitaplar interaktif değildir.

8 Ayrıca bkz. Chalmers (2010). Buradan dışarıdaki bir varlığın asla gözlemlemeyeceği bir sistem inşa etmenin hiçbir yararı olmadığı çıkarımını yapmak hata olur. Bu tür bir sistemin içinde neler olup bittiğine nihai bir değer atfedilebilir. Ayrıca diğer insanlar bu tür bir sistemin içinde neler olup bittiğine dair tercihlere sahip olabilirler ve bu nedenle onun yaratılmasından ya da yaratılması vaadinden etkilenebilirler. Belli türde yalıtık sistemlerin (gözlemciler içeren sistemlerin) varoluşunun bilgisi dışarıdaki gözlemcilerde antropik belirsizliğe de yol açabilir ve onların davranışlarına etki edebilir.

9 Toplumsal bütünleşmenin neden bir yeterlilik kontrol biçimi olarak görüldüğü merak edilebilir. Bunun yerine, bir sistemin davranışını teşvikler yoluyla etkilemeye çalışmasını içerdiği düşünüldüğünde bir güdülenme seçilim yöntemi olarak sınıflandırılması gerekmez mi? Birazdan güdülenme seçilimine yakından bakacağız, ama bu soruya cevaben, güdülenme seçilimini bir sistemin nihai hedeflerini (araçsal nedenlerden ziyade kendi yüzü suyu hürmetine peşine düşülen hedefleri) seçerek ya da şekillendirerek işleyen birtakım kontrol yöntemleri olarak yorumluyoruz. Toplumsal bütünleşme bir sistemin nihai hedeflerini hedef almaz, dolayısıyla güdülenme seçilimi

değildir; daha ziyade, sistemin etkin yeterliliklerini sınırlamayı amaçlar: Sistemi belli birtakım sonuçları –sistemin ilintili cezalarla (misilleme ve işbirliğinden elde edilen kazançların kaybedilmesiyle) karşılaşmadan kaçışın yararlarını elde ettiği sonuçları– gerçekleştiremez hale getirmeyi amaçlar. Sistemin hangi sonuçları elde edebileceğini sınırlayarak sistemin nihai hedeflerini gerçekleştirmenin geriye kalan yegâne aracının işbirliği yapmak olduğunu bulacağı umut edilir.

10 Bu yaklaşım emülasyonun insanbiçimci güdülenmeleri olduğuna inanılan durumda biraz daha ümit vaat edici olabilir.

11 Bu fikri Carl Shulman'a borçluyum.

12 Süper zeki bir şifre kırıcıya dayanması kesin olan bir şifreleme yaratmak hiç de yabana atılmayacak bir zorluktur. Örneğin rastgele sayıların izleri bir gözlemcinin beyninde ya da rastgele üretecin mikroyapısında kalmış olabilir ve süper zekâ bunları buradan kurtarabilir ya da yalancı rastgele sayılar kullanılırsa, süper zekâ bunların yaratıldığı tohumu tahmin edebilir veya keşfedebilir. Ayrıca süper zekâ büyük kuantum bilgisayarlar yaratabilir, hatta yeni türde bilgisayarlar yaratmak için kullanılabilecek bilinmedik fiziksel görüngüler keşfedebilir.

13 Yapay zekâ kendisini ödül işaretleri aldığına inanacak şekilde *yapılandırabilir*, ama ödül işaretlerini isteyecek şekilde tasarlanırsa (ödül işaretleri konusunda belli inançları olduğu bir durumda olmayı istemenin aksine), bu onu *wirehead* yapmayacaktır.

14 Karş. örneğin Quine ve Ullian (1978).

15 Bir yapay zekâ bunu çeşitli temel bilgi işlem işlevselliklerinin performans niteliklerini (çeşitli veri taşıtlarının büyüklüğü ve kapasitesi, belleğin farklı kısımlarına erişimin aldığı zaman, rastgele bit flipler vb.) ele alarak araştırabilir.

16 Belki de önce; algoritmik karmaşıklıkları temelinde olanaklı dünyalara olasılık atfeden Solomonoff öncesi (hesaplama kestirimi) olabilir. Bkz. Li ve Vitanyi (2008).

17 Aldatmanın tasavvur edilmesinden *sonraki* anda yapay zekâ isyankâr düşüncesinin izini silmeyi becerebilir. Bu nedenle, bu tetikleyicinin süreli işlemesi önemlidir. Keza yapay zekânın tüm faaliyetinin tam bir izini depolayan bir "uçuş kaydedicisi (kara kutu)" kullanmak da iyi bir pratik olacaktır (programlayıcılardan klavye girişinin kesin zamanlaması dahil), dolayısıyla otomatik bir kapamanın ardından güzergâhının izi sürülebilir ya da analiz edilebilir. Bu bilgi "bir kez yaz birçok kez oku" aracısında depolanabilir.

18 Asimov (1942). Bu üç yasaya sonradan bir "Sıfırıncı Yasa" da eklendi: "(0) Bir robot insanlığa zarar veremez ya da hiçbir şey yapmayarak zarar görmesine izin veremez" (Asimov 1985).

19 Karş. Gunn (1982).

20 Russell (1986, 161f).

21 Benzer şekilde, bazı filozoflar tüm kariyerlerini deontoloji sistemlerini dikkatle formüle etmeye çalışmakla geçirmiş olsalar da, ara ara düzeltmeler gerektiren yeni vakalar ve sonuçlar gün ışığına çıkar. Örneğin deontolojik ahlak felsefesi son yıllarda verimli bir alan olan yeni felsefi düşünce deneylerinin keşfedilmesiyle yeniden canlanmıştır: Eylemler/ihmaller ayrımında, yani niyetlenen sonuçlar ile niyetlenmeyen sonuçlar arasındaki ayrımın ve bu türde diğer meselelerin ahlaki

önemine dair sezgilerimiz hakkında birçok incelikli etkileşimi ortaya seren "vagon problemi". Örneğin bkz. Kamm (2007).

22 Armstrong (2010).

23 Pratik olarak, eğer bir yapay zekâyı kontrol altında tutmak için çoklu güvenlik mekanizmaları kullanmak amaçlanırsa, her birini *sanki* yegâne güvenlik mekanizması olması amaçlanıyormuş gibi ve dolayısıyla *sanki* tekil olarak yeterli olması gerekiyormuş gibi çalışmak akıllıca olabilir. Eğer delik bir kovanın içine başka bir delik kova koyarsanız, yine de su sızdıracaktır.

24 Aynı fikrin bir çeşidi de yapay zekânın zımnen tanımlanan standardın ne olduğuna dair en iyi tahminlerine göre hareket edecek şekilde sürekli güdülenmiş olmasıdır. Bu kurulumda, yapay zekânın nihai hedefi her zaman zımnen tanımlanmış standarda göre hareket etmektir ve bu standardın ne olduğuna dair araştırmasını sadece araçsal nedenlerle takip eder.

25 Extrapolation –yn.

ONUNCU BÖLÜM

1 Bu adlar elbette insanbiçimseldir ve analoji olarak ciddiye alınmamalıdır. Bunlar insanın yaratmaya çalışmayı düşünülebileceği olanaklı sistem türlerinin ilk bakışta farklı kavramları için etiketten ibaret değildir.

2 Bir sonraki seçimin sonucuna dair bir soruya yanıt olarak, yakındaki parçacıkların öngörülen konumu ve moment vektörlerinin kapsamlı bir listesini kimse almak istemez.

3 Belli bir makinedeki komut takımına endekslidir.

4 Kuhn (1962); de Blanc (2011).

5 Bu tür bir "mutabakat yöntemi"ni cinlere ya da hükümdarlara uygulamak zor olacaktır, çünkü verili bir hedefi gerçekleştirme konusunda neredeyse tastamam aynı derecede etkili olacak sayısız temel eylem dizileri olabilir (sistemin aktüatörlerine tikel elektrik sinyali örüntüleri yollamak gibi); bu yüzden biraz farklı aracılar biraz farklı eylemleri meşru olarak seçebilir ve böylece bir mutabakata ulaşılamadan kalır.

Halbuki uygun şekilde formüle edilmiş sorularla, genellikle az sayıda uygun cevap seçeneği ("evet" ve "hayır" gibi) olacaktır. ("Odak noktası" diye de geçen bir Schelling noktası kavramı hakkında, bkz. Schelling [1980].)

6 Dünya ekonomisi bir bakıma zayıf bir cine benzemiyor mu? Tek farkı hizmetlerinden ötürü bir hesap çıkarmasıdır. O zaman, gelecekte geliştirilebilecek türde muazzam büyüklükte bir ekonomi kolektif süper zekâya sahip bir cine benzeyebilir.

Mevcut ekonominin bir cinden *farklılık gösterdiği* önemli noktalardan biri benim bir ücret karşılığında ekonomiye komut vererek kapıma pizza getirtebilirken, barışı getirmesi komutunu veremiyor olmamdır. Bunun nedeni ekonominin yeterince güçlü olmaması değil, yeterince eşgüdümlü olmamasıdır. Bu açıdan, ekonomi tek bir cinden ya da başka türde bir birleşik aracıdan ziyade farklı efendilere hizmet eden (ve farklı gündemleri olan) bir cinler *meclisine* benzer. Her bir cini daha güçlü hale getirerek ya da daha fazla cin ekleyerek ekonominin bütünsel gücünü artırmak, mutlaka ekonomiyi daha huzur verici hale getirmez. Bir süper zeki cin gibi işlev

göstermek için, ekonomi pahalı olmayan şekilde mal ve hizmet (radikal derecede yeni teknoloji gerektirenler dahil) üretme yeteneğini artırmakla kalmayıp, küresel eşgüdüm problemlerini de daha iyi çözebiliyor olmalıdır.

7 Eğer cin bir sonraki emre uymamayı bir türlü başaramıyorsa –ve kendisini bu hassaslıktan kurtaracak şekilde yeniden programlamayı bir şekilde beceremiyorsa– o zaman yeni bir emrin verilmesini engelleyecek şekilde hareket edebilir.

8 Evet/hayır cevabı vermekle sınırlı bir kâhin bile bir cin ya da hükümdar yapay zekâ arayışını kolaylaştırmak için kullanılabilir, hatta bu tür bir yapay zekâda doğrudan bir bileşen olarak kullanılabilir. Eğer yeterince çok sayıda soru sorulabilirse, kâhin bu tür bir yapay zekâ için gerçek kod üretmek için de kullanılabilir. Bu tür sorular kabaca şu biçimi alabilir: "Bir cini oluşturacak olan ilk yapay zekânın kodunun ikili versiyonunda n'ci sembolü sıfır mıdır?"

9 Ancak saptanmış bir otorite tarafından verilirse emirleri ya da soruları kabul eden biraz daha karmaşık bir kâhin ya da cin tahayyül edilebilir, ama yine de bu otoritenin yozlaşma ya da üçüncü bir tarafın şantajına uğrama olasılığı devam edecektir.

10 20. yüzyılın önde gelen siyaset felsefecilerinden John Rawls bir toplumsal sözleşmenin formüle edilmesinde hesaba katılması gereken tercih türlerini nitelendirmenin bir yolu olarak cehalet perdesinden yararlanmıştır. Rawls hangi insan olacağımızı ve hangi toplumsal rolü işgal edeceğimizi bilmekten bizi alıkoyan bir cehalet perdesinin arkasından bir toplumsal sözleşmeyi seçtiğimizi tahayyül etmemiz gerektiğini söylemiştir; burada kastedilen, bu tür bir durumda adil olmayan ayrıcalıkların sefasını süreceğimiz bir toplumsal düzeni seçmemize yol açabilecek bencil çıkarlarımızı ve önyargılarımızı umursamadan hangi toplumun genel olarak en adil ve en çok istenen toplum olacağını düşünmemiz gerekeceğidir. Bkz. Rawls (1971).

11 Karnofsky (2012).

12 İstisnalardan biri yeterince güçlü başlatıcılara bağlanan yazılım (doğrudan nükleer savaş başlıklarına ya da nükleer saldırı başlatmaya yetkili insan yetkililere bağlıysa ilk uyarı sistemlerindeki yazılım gibi) olacaktır. Bu tür bir yazılımda arızalar yüksek riskli durumlara yol açabilir. Bunun en az iki kez yaşandığı hafızalarda tazedir. 9 Kasım 1979'da bir bilgisayar problemi sonucu, NORAD (Kuzey Amerika Hava-Uzay Savunma Komutanlığı) çok yakında Sovyetlerin ABD'ye büyük çaplı bir saldırı hazırlığında olduğunu rapor etmişti. ABD de acilen misilleme hazırlıklarına girişmiş, ama ancak daha sonra erken uyarı radarı sistemlerinden veriler hiçbir saldırı gerçekleştirilmediğini göstermişti (McLean ve Stewart 1979). 26 Eylül 1983'te, arızalanan Soviet Oko nükleer erken uyarı sistemi ABD'nin bir füze saldırısına kalkıştığını bildirmişti. Fakat komuta merkezindeki nöbetçi subay Stanislav Petrov bunun yanlış alarm olduğunu tespit etmişti: Bu kararın bir nükleer savaşın önüne geçtiğine inanılmaktadır (Lebedev 2004). Görünen o ki, Soğuk Savaş'ın tepe noktasında bütün nükleer güçlerin cephaneleri bir araya da gelseydi, bu savaş medeniyeti yok edecek ve akla hayale gelmeyecek ıstıraplara ve ölümlere yol açacak olsa da, insanlığın yok oluşuna neden olacak kadar büyük bir savaş olmayacaktı (Gaddis 1982; Parrington 1997). Ama geleceğin silahlanma yarışlarında daha büyük stoklar biriktirilebilir, hatta daha da ölümcül silahlar icat edilebilir ya da büyük bir nükleer savaşın etkilerine dair modellerimiz (özellikle de müteakip nükleer kışın sertliği) yanlış olabilir.

13 Bu yaklaşım bir doğrudan-saptama kural-temelli kontrol yöntemi kategorisine uyabilir.

14 Eğer çözüm kıstası bir iyilik ölçüsü saptıyorsa, durum özünde aynıdır.

15 Kâhin yaklaşımını savunan biri önerilen çözümdeki kusuru kullanıcının saptayabileceğine –resmi başarı kıstaslarını karşılasa bile, kullanıcının niyetine uymadığını görebileceğine– dair en azından bir olasılık olduğunda ısrar edebilir. Bu aşamada hatayı yakalama ihtimali çeşitli etkenlere bağlı olacaktır: Kâhinin çıktılarının insanlar tarafından ne kadar anlaşılabildiği ve potansiyel sonucun hangi sonuçlarının kullanıcının dikkatine sunulacağını seçmenin ne kadar iyiliksever olduğu gibi.

Alternatif olarak, bu işlevsellikleri sağlamak için bizatihi kâhine yaslanmak yerine bunu yapmak için ayrı bir alet (kâhinin beyanlarını teftiş edecek ve bize eğer bunlara göre hareket edersek ne olacağını yardımcı bir tarzda gösterecek bir alet) yapmaya çalışabilir. Ama bunu tüm genelliği içinde yapmak kehanetlerine güvenmemiz gerekecek olan başka bir süper zeki kâhine gerek duyacaktır; dolayısıyla güvenilirlik problemi çözülmüş olmayacak, sadece odağı kayacaktır. Emsal değerlendirmeyi yapmak için birden çok kâhinden yararlanarak güvenliği artırmaya çalışmak mümkün olabilir, ama tüm kâhinlerin aynı tarzda başarısız olduğu örneklerde bu bir koruma sağlamaz. Sözgelimi hepsine neyin tatmin edici bir çözüm sayılacağına dair aynı formel saptama verildiğinde bu yaşanabilir.

16 Yeterince büyük bilgi işlem gücü miktarlarıyla (sonlu ama fiziksel açıdan akıl almaz), şu an mevcut olan algoritmalarla genel süper zekâya ulaşmak herhalde mümkün *olacaktır*. (Örneğin karş. AIXI*tl* sistemi; Hutter [2001].) Fakat Moore yasasının bir yüzyıl daha devam etmesi bile bunu gerçekleştirmek için gerekli bilgi işlem gücüne ulaşmaya yetmeyecektir.

ON BİRİNCİ BÖLÜM

1 Bunun nedeni mutlaka en muhtemel ya da en çok istenen senaryo türü olması değil, standart bir ekonomi teorisinin aletleriyle analiz etmesi en kolay senaryo ve dolayısıyla değerlendirmemiz için uygun bir kalkış noktası olmasıdır.

2 American Horse Council (2005). Ayrıca bkz. Salem ve Rowan (2001).

3 Acemoglu (2003); Mankiw (2009); Zuleta (2008).

4 Fredriksen (2012, 8); Salverda vd. (2009, 133).

5 Sermayenin en azından bir kısmının genel yükselişle birlikte değerlenen varlıklara yatırılması da esastır. Bir endeks takip fonundaki hisseler gibi çeşitlendirilmiş bir varlık portföyü tamamen gözden kaçmama şansını artıracaktır.

6 Avrupa'daki refah sistemlerinin çoğu *finans desteğinden yoksundur*, bu demektir ki emeklilik maaşları tasarruflardan değil şu an çalışan işçilerin katkılarından ve vergilerinden ödenmektedir. Bu tür şemalar gerekliliği otomatikman karşılamayacaktır: Ani bir işsizlik durumunda, maaşların ödendiği gelirlerin kaynağı kuruyabilir. Gelgelelim hükümetler bu eksiği başka kaynaklardan telafi etmeyi tercih edebilirler.

7 American Horse Council (2005).

8 7 milyar insana her yıl 90 bin dolar emekli maaşı bağlamak yılda 630 trilyon dolar yapar ki, dünyanın mevcut gayri safi hâsılasının on katıdır. Maddison'a (2007) göre, son yüz yılda, dünyadaki gayri safi hâsıla yaklaşık 19 kat artmış ve 1900'de 2 trilyon dolarken 2000'de 37 trilyona ulaşmıştır. Dolayısıyla son yüzyılda gördüğümüz büyüme oranları nüfus aynı kalırken iki yüzyıl daha sürerse, herkese yıllık 90 bin dolar emekli maaşı bağlamak dünya gayri safi hâsılasının yüzde 3'üne denk gelecektir. Bir zekâ patlaması bu büyüme rakamına çok daha kısa bir sürede ulaşılmasını sağlayabilir. Ayrıca bkz. Hanson (1998a, 1998b, 2008).

9 Ve iddia edildiği gibi, son 70 bin yılda bu civarda ciddi bir nüfus tıkanıklığı olduysa, belki de neredeyse bir milyon kat artmıştır. Daha fazla veri için, bkz. Kremer (1993) ve Huff vd. (2010).

10 Cochran ve Harpending (2009). Ayrıca bkz. Clark (2007) ve bir eleştiri için, Allen (2008).

11 Kremer (1993).

12 Basten vd. (2013). Yükselişin devam ettiği senaryolar da mümkündür. Genel olarak, bu tür tahminlerin belirsizliği bir ya da iki kuşak sonra muazzam artmaktadır.

13 Küresel çapta alındığında, yenilenmede toplam doğurganlık oranı 2003'te kadın başına 2,33 çocuktu. Bu sayıya şöyle ulaşıyoruz: Ebeveynleri ikame etmek kadın başına iki çocuk tutmaktadır, artı "bir çocuğun üçte biri" de (1) erkek çocuk doğmasının yüksek olasılığını; (2) doğurgan yaşamlarının sonundan önce erken ölümlüğü telafi içindir. Gelişmiş uluslarda sayı düşük ölüm oranlarından ötürü daha azdır: 2,1 civarı. (Bkz. Espenshade vd. [2003, Giriş, Tablo 1, 580].) Eğer göçler olmasaydı en gelişmiş ülkelerde nüfus azalacaktı. Yenilenme altı doğum oranları olan ülkeler içinde birkaç göze çarpan örnek: Singapur, 0,79 (dünyada en düşük); Japonya, 1,39; Çin Halk Cumhuriyeti, 1,55; Avrupa Birliği ülkeleri, 1,58; Rusya, 1,61; Brezilya, 1,81; İran, 1,86; Vietnam, 1,87 ve Birleşik Krallık, 1,90. ABD nüfusu bile 2,05 doğurganlık oranıyla muhtemelen birazcık azalacaktır (Bkz. CIA [2013].)

14 Zamanın tamlığı milyarlarca yıl sonra gerçekleşebilir.

15 Carl Shulman biyolojik insanlar dijital ekonominin yanı sıra doğal yaşam sürelerini sonuna kadar yaşamaya bel bağlarlarsa, sadece dijital alandaki siyasal düzenin insan çıkarlarını korumakla kalmayıp bunun çok uzun bir dönem boyunca süreceğini varsaymaları gerektiğine dikkat çeker (Shulman 2012). Örneğin dijital alandaki olaylar dışarıdaki olaylardan bin kat daha hızlı gerçekleşirse, o zaman biyolojik bir insanın 50 bin yıllık iç değişim ve çalkantı boyunca aynı kalan dijital siyasal bedene yaslanması gerekecektir. Ama dijital siyasal dünya bizimkine benzerse, bu binyıllar boyunca birçok devrim, savaş ve felaket olacak ve muhtemelen dışarıdaki biyolojik insanlara zorluk çıkaracaktır. Küresel bir nükleer savaşın ya da benzer bir felaketin yol açacağı riskin yüzde 0,01'i bile hayatlarını ağır çekimli yıldız zamanında yaşayıp giden biyolojik insanlar için neredeyse kesin bir kayıp getirecektir. Bu problemi aşmak için dijital alanda daha istikrarlı bir düzen gerekecektir: Belki de kendi istikrarını tedricen geliştiren bir tekel.

16 Makineler insanlardan çok daha etkili olsalardı bile, bir insan işçiyi çalıştırmanın kârlı olacağı bir ücret düzeyi (sözgelimi saati bir sent) düşünülebilir. Eğer insanlar için tek gelir kaynağı bu olsaydı, türümüz yok olurdu, zira insanlar saatte 1 sentle hayatta kalamazlar. Ama insanlar sermayeden de gelir elde ederler. Şimdi eğer toplam gelirin

asgari geçim seviyesinde olacağı noktaya kadar nüfusun artacağını varsayıyorsak, insanların çok sıkı çalıştıkları bir durumdan bahsettiğimiz söylenebilir. Örneğin asgari geçim gelirinin günde 1 dolar olduğunu varsayın. O zaman, nüfusun kişi başına sermaye günde sadece 90 sent gelir sağlayana kadar büyüyecektir denebilir, bundan sonra insanlar geri kalan 10 senti bulup buluşturmak adına 10 saatlik ağır bir mesai daha yapmak zorunda kalacaklardır. Gelgelelim böyle olması gerekmiyor, çünkü asgari geçim geliri yapılan iş miktarına bağlıdır: Sıkı çalışan insanlar daha fazla kalori yakarlar. Her bir saat mesainin yiyecek masraflarını 2 sent artırdığını varsayalım. Bu durumda, insanların dengeye vurulduğunda aylak(??) oldukları bir model ortaya çıkacaktır.

17 Bu kadar zayıf bir grubun oy veremeyeceği veya haklarını savunamayacağı düşünülebilir. Ama kabuklarında yaşayanlar işlerini idare etmesi ve siyasal çıkarlarını temsil etmesi için yapay zekâ yedieminlerini vekil olarak atayabilirler. (Bu kısımdaki değerlendirmenin bu bölümü mülkiyet haklarına saygı duyulduğunu varsaymaktadır.)

18 En iyi kelimenin ne olduğu açık değildir. "Öldür", gerektiğinden çok daha aktif bir vahşeti akla getirebilir. "Son" da hüsnütabir olur. Karmaşıklıklardan biri potansiyel açıdan iki ayrı olay olmasıdır: Aktif bir şekilde bir süreç başlatmayı bırakmak ve bilgi şablonunu silmek. Bir insan ölümü normalde iki olayı kapsar, ama bir emülasyon için bu ikisi ayrılabilir. Bir programın *geçici olarak* çalışmaması bir insanın uyumasından daha önemli olmayabilir; ama *kalıcı olarak* çalışmaması bir insanın kalıcı komaya girmesi anlamına gelebilir. Yine de emülasyonların kopyalanabiliyor olması ve farklı hızlarda çalıştırılabilmesi de karmaşıklıklar doğurmaktadır: Bu olanakların insan deneyiminde doğrudan analogları yoktu. (Karş. Bostrom [2006b]; Bostrom ve Yudkowsky [2015].)

19 Toplam paralel bilgi işlem gücü ile bilgi işlem hızı arasında bir ödünleşim olacaktır, zira en yüksek bilgi işlem hızları ancak güç etkinliğindeki bir azalma pahasına elde edilebilecektir. Bu özellikle de tersinebilir bilgi işlem çağına girildiğinde doğru olacaktır.

20 Bir emülasyon baştan çıkartılarak test edilebilir. Belli bir hazırlanmış durumdan başlatılmış bir emülasyonun çeşitli uyarılma dizilimlerine nasıl tepki verdiğini tekrar tekrar test ederek bu emülasyonun güvenilirliğine olan güven artırılabilir. Ama zihinsel durum sonradan başlangıç noktasının ötesinde ne kadar geliştirilirse, güvenilir kalma ihtimali o kadar az olacaktır. (Daha özelde, akıllı bir emülasyon bazen bir simülasyonda olduğunu varsayabileceğinden, simülasyon hipotezinin karar alma sürecindeki ağırlığının daha az olacağı durumlara bu davranışını aksettirme konusunda daha dikkatli olması gerekir.)

21 Bazı emülasyonlar tikel bir örneksemeden ziyade klanlarıyla, örneğin aynı şablondan çıkarılmış olan tüm kopyalarıyla ve çeşitleriyle özdeşlik kurabilirler. Bu tür bir emülasyon diğer klan üyelerinin hayatta kalacağını bilirse kendi sonlanışını bir ölüm olayı olarak görmeyebilir. Emülasyonlar günün sonunda belli bir depolanmış duruma geri döndürüleceklerini ve o günkü anılarını yitireceklerini bilebilirler, ama partiye giden biri ertesi sabah uyandığında o gece yaşadıklarının hiçbirini hatırlamayacak olmasından ne kadar etkilenirse bu da onları ancak o kadar etkiler: Bunu ölüm olarak değil, geriye dönük amnezi olarak görürler.

22 Etik bir değerlendirme başka birçok etkeni de dikkate almalıdır. Tüm işçiler her zaman durumlarından memnun olsalar bile, sonuç yine de başka temellerde ahlaken itiraz edilebilir olabilir ama bunların hangi temeller olduğu rakip ahlak teorileri arasında bir tartışma konusudur. Ama akla yatkın her değerlendirmenin öznel esenliği önemli bir etken olarak göreceği kesindir. Ayrıca bkz. Bostrom ve Yudkowsky (2015).

23 World Values Survey (2008).

24 Helliwell ve Sachs (2012).

25 Karş. Bostrom (2004). Ayrıca bkz. Chislenko (1996) ve Moravec (1988).

26 Bu tür bir senaryoda ortaya çıkacak olan bilgi işlem yapılarının, (olgusal deneyime, yani *qualia*'ya sahip olma anlamında) bilinçli olup olmayacağını söylemek zordur. Bunun zor olmasının nedeni kısmen hangi bilişsel varlıkların çıkacağına dair ampirik bilgisizliğimiz, kısmen de hangi tür yapıların bilince sahip olduğuna dair felsefi bilgisizliğimizdir. Soruyu farklı bir şekilde sorabiliriz: Gelecekteki varlıkların bilinçli olup olmayacağı yerine, ahlaki statülerinin olup olmayacağını sorabiliriz. Veya "esenlik"lerini tercih edeceğimiz türde varlıklar olup olmayacağını sorabiliriz. Ama bu soruları cevaplamak bilinç hakkındaki soruları cevaplamaktan daha kolay olmayabilir; aslında, ahlaki statü ya da tercihlerimiz söz konusu varlığın kendi koşulunu öznel olarak deneyimleyip deneyimlemediğine bağlı olduğu oranda bilinç sorusuna bir yanıt gerektirebilir.

27 Gerek jeoloji gerek insanlık tarihinin bu türde daha büyük bir karmaşıklık eğilimi sergilediğine dair bir argüman için, bkz. Wright (2001). Buna karşı çıkan bir argüman için (Wright'ın kitabının 9. Bölümü'nde eleştirilmektedir), bkz. Gould (1990). Ayrıca şu anda şiddetin ve mezalimin uzun vadede ciddi ciddi azalma eğilimi gösterdiği bir döneme şahit olduğumuzu savunan bir argüman için, bkz. Pinker (2011).

28 Gözlem seçilim teorisi hakkında daha fazlası için, bkz. Bostrom (2002a).

29 Bostrom (2008a). Evrimsel tarihimizin ayrıntılarına dair çok daha dikkatli bir inceleme seçilim etkisini aşmalıdır. Örneğin bkz. Carter (1983, 1993); Hanson (1998d); Ćirković vd. 2010).

30 Kansa (2003).

31 Örneğin Zahavi ve Zahavi (1997).

32 Bkz. Miller (2000).

33 Kansa (2003). Kışkırtıcı bir görüş için, ayrıca bkz. Frank (1999).

34 Küresel siyasi bütünleşmenin derecesini ölçmenin en iyi yolunun ne olduğu açık değildir. Perspektiflerden biri şu olacaktır: Bir avcı-toplayıcı kabile karar alma mekanizmasına yüz bireyi dahil edebilirken, bugün en büyük siyasi oluşumlar bir milyardan fazla insanı barındırmaktadır. Arada yedi basamak fark var demektir ve bir basamak daha eklendiğinde tüm dünya nüfusu tek bir siyasal oluşuma dahil edilmiş olacaktır. Gelgelelim kabilenin en büyük bütünleşme ölçeği olduğu dönemde dünya nüfusu çok daha azdı. Kabilede o sırada yaşayan insanların binde biri yer alabilirdi. Böylece siyasal bütünleşme ölçeğindeki artışın daha az, yani iki basamak olduğunu söylememiz gerekecektir. Mutlak rakamlardan ziyade dünya nüfusunun siyasi açıdan bütünleşmiş kısmına bakmak mevcut bağlamda uygun görünmektedir, özellikle de makine zekâsına geçiş bir nüfus patlamasına (emülasyon ya da diğer dijital zihinler)

yol açabileceğinden. Fakat resmi devlet yapılarının dışında küresel kurumlarda ve işbirliği ağlarında da gelişmeler olmuştur ve bunları da hesaba katmak gerekir.

35 İlk makine zekâsı devriminin hızlı gerçekleşeceği varsayımının gerekçelerinden biri burada geçerli değildir. Ancak, emülasyondan tamamen sentetik makine zekâsına geçişte yazılımdaki buluşlar gibi başka çabuk kazanımlar gerçekleşebilir.

36 Shulman (2010b).

37 Artı ve eksilerin nasıl dengeleneceği süperorganizmanın ne yapmaya çalıştığına ve en genel olarak becerikli olan mevcut emülasyon şablonunun ne kadar becerikli olduğuna bağlı olabilir. Bugün büyük örgütlerde birçok farklı türde insana ihtiyaç duyulmasının nedenlerinden biri birçok alanda çok yetenekli olan insanların nadir bulunmasıdır.

38 Bir yazılım aracısının birden çok kopyasını yapmak elbette çok kolaydır. Ama kopyalama sonucunda kopyaların aynı nihai hedeflere sahip olacaklarının garantisi yoktur. İki aracının ("aynı"nın ilintili anlamında) aynı nihai hedeflere sahip olması için, hedeflerinin işaretsel öğeleri bakımından çakışması gerekir. Eğer Bob bencilse, Bob'un bir kopyası da bencil olacaktır. Ama hedefleri örtüşmez: Bob Bob'u önemserken, Bob kopyası Bob kopyasını önemser.

39 Shulman (2010b, 6).

40 Keşfetmesi çok zor olan gizli bölmeler ya da işlevsel dinamiklere sahip olacak şekilde yaratılabilecek keyfi yapay zekâlardansa biyolojik insanlar ve tüm beyin emülasyonları için daha geçerli olabilir bu. Diğer yandan özel olarak şeffaf olacak şekilde yaratılmış yapay zekâlar beyne benzeyen mimarilere kıyasla daha eksiksiz bir teftiş ve doğrulamaya imkân tanıyacaklardır. Toplumsal baskılar yapay zekâları kaynak kodlarını teşhir etmeye ve kendilerini şeffaf kılacak şekilde değiştirmeye teşvik edebilir, özellikle de şeffaf olmak güvenilmek için ve dolayısıyla hayırhah işlemlerde yer alma fırsatı sunulması için bir önkoşulsa. Karş. Hall (2007).

41 Görece tali görünen diğer bazı meseleler, özellikle de muazzam bir kaybın söz konusu olacağı durumlarda (kilit küresel eşgüdüm başarısızlıklarında olduğu gibi) karşılıklı çıkar sağlayabilecek politikalar bulma arayışını ve bazı aracıların beraberinde izleme ve yaptırım mekanizmaları bulunan kapsamlı küresel anlaşmalara girerek azaltılacak bir biçimde "özerkliği" tercih etmeleri imkânını içerir.

42 Bir yapay zekâ belki de bunu kendisini uygun şekilde değiştirerek ve ardından gözlemcilere sadece kaynak koduna salt okuma erişimi vererek başarırlar. Daha anlaşılmaz bir mimariye sahip bir makine zekâsı (emülasyon gibi) belki de bunu kendisini bir güdülenme seçilim yöntemine açıkça uygulayarak başarabilir. Alternatif olarak, dışsal bir zorlayıcı aracı (bir süper organizma polis gücü) belki de sadece farklı taraflar arasında yapılmış bir anlaşmanın uygulanmasını dayatmak için değil, aynı zamanda içsel olarak bir tek tarafın kendisini tikel bir eylem yoluna bağlaması için de kullanılabilir.

43 Evrimsel seçilim tehditleri umursamayanların, hatta en ufak bir bozgun yaşamaktansa ölümüne savaşmaya hazır olacak gergin karakterlerin yüzünü güldürmüş olabilir. Bu tür bir temayül, taşıyıcısına değerli işaretleşme yararları sağlayabilir. (Bu temayüle sahip olmanın bu tür araçsal ödülleri elbette aracının bilinçli güdülenmesinde rol oynamak zorunda değildir: Adalet ya da şerefi başlı başına bir amaç olarak görebilir.)

44 Gelgelelim bu tür meselelerde kesin bir yargı için daha fazla analize ihtiyaç vardır. Burada daha fazla irdeleyemeyeceğimiz başka potansiyel karmaşıklıklar vardır.

ON İKİNCİ BÖLÜM

1 Bu temel fikrin çeşitli karmaşıklıkları ve modülasyonlarından bahsedilebilir. Bunlardan birini (azamileştirici aracının tam aksi olan tatmin edici aracı) 8. Bölüm'de ele almıştık ve bir sonraki bölümde de alternatif karar teorilerine kısaca değineceğiz. Gelgelelim bu tür meseleler bu altbaşlık açısından asli değildir, bu nedenle burada beklenen fayda azamileştirici aracı örneğine odaklanarak tane tane ilerleyeceğiz.

2 Yapay zekânın entipüften denemeyecek bir fayda işlevi olacağını varsayarsak. Her zaman beklenen faydayı azamiye çıkarıcı eylemi tercih eden bir aracı inşa etmek çok kolay olacaktır, tabii eğer onun fayda işlevi örneğin değişmez fonksiyonu U(w) = 0 ise. Her eylemin bu fayda işleviyle göreli olarak beklenen faydayı aynı şekilde azamiye çıkarması pekâlâ mümkündür.

3 Ayrıca ilk bebeklik dönemimizin vızıltılı kafa karışıklığını unuttuğumuz için; bu dönemde henüz çok iyi göremiyorduk, çünkü beynimiz henüz görsel girdilerini yorumlamayı öğrenmemişti.

4 Ayrıca bkz. Yudkowsky (2011) ve de Muehlhauser and Helm'in (2012) kitabının 5. Bölümü'ndeki değerlendirme.

5 Belki de yazılım mühendisliğindeki ilerlemelerin nihayetinde bu zorlukların üstesinden gelmesi bir o kadar mümkündür. Modern aletlerden yararlanan tekil bir programlayıcı makine kodunu doğrudan yazmaya zorlanan büyükçe bir geliştirici takımının ötesinde bir yazılım üretebilir. Günümüz yapay zekâ programlayıcıları yüksek kaliteli makine öğrenimi ve bilimsel hesaplama kütüphaneleri gibi kişiler örneğin kendi başlarına asla yapamayacakları şekilde kütüphaneleri bir araya getirerek çeşitli bilgisayar korsanlığı imkânı veren olanakların yaygınlığı sayesinde kendilerini gösterebiliyorlar. Uzmanlar tarafından üretilen ama uzman olmayanlar tarafından kullanılabilen yeniden kullanılabilir yazılım birikimi geleceğin programlayıcılarına ifade üstünlüğü sağlayacaktır. Örneğin gelecekteki bir robotbilim programlayıcısı standart yüz tanıma kütüphaneleri, ofis programları vs. gibi şu an mevcut olmayan diğer birçok işleve kolaylıkla erişim elde edebilir.

6 Dawkins (1995, 132). Burada doğal dünyadaki ıstırap miktarının olumlu esenlik miktarını mutlaka *aşacağı* iddia edilmiyor.

7 Gerekli nüfus büyüklükleri kendi atalarımızda olduğundan çok daha büyük ya da çok daha küçük olabilir. Bkz. Shulman ve Bostrom (2012).

8 Eğer çok sayıda masuma zarar vermeden eşdeğer bir sonuca ulaşmak kolay olsaydı, bunu yapmak ahlaken daha iyi olurdu. Ne var ki dijital insanlar yaratılır ve haksız yere zarar görürlerse, bunu telafi etmek için bunları dosyaya kaydetmek ve ardından (insanlığın geleceği garanti altına alındığında) daha olumlu koşullarda bunları yeniden çalıştırmak mümkün olabilir. Bu tür bir tazmin ilahiyatta şer problemini ele alma girişimleri bağlamında öbürdünyaya dair dini anlayışlarla karşılaştırılabilir.

9 Alanının önde gelen isimlerinden Richard Sutton takviyeli öğrenmeyi bir öğrenme yöntemi üzerinden değil, bir öğrenme problemi üzerinden tanımlar: Bu problemi çözmeye uygun olan her yöntem bir takviyeli öğrenme yöntemi olarak tarif edilir

(Sutton ve Barto 1998, 4). Buradaki değerlendirmeyse aracının kümülatif ödülü (böyle bir kavramı) azamiye çekme nihai hedefine sahip olması diye tasavvur edilebilir. Çok farklı türde nihai hedefe sahip olan bir aracı ödül arayan bir aracıyı çok çeşitli durumlarda taklit etmekte yetenekli olabileceğinden ve dolayısıyla takviyeli öğrenme problemlerini çözmeye daha uygun olduğundan, Sutton'ın tanımına göre "takviyeli öğrenme yöntemi" olarak tanımlanabilecek ama bir *wireheading* sendromuna yol açmayacak yöntemler olabilir. Gelgelelim metindeki ifadeler çoğunlukla takviyeli öğrenme topluluğunda kullanılan yöntemlerin çoğu için geçerlidir.

10 İnsan türü bir mekanizmayı bir şekilde insan türü bir makine zekâsı içinde kurmak mümkün olsa bile, bu dijital bebeğin yetiştirildiği ortam sıradan bir çocuğunkine çok benzemediği takdirde bu zekânın edindiği nihai hedeflerin iyi ayarlanmış bir insanınkine benzemesi gerekmez, ama bunu ayarlamak zor olacaktır. İnsan benzeri bir yetişme ortamında bile, tatmin edici bir sonuç garanti değildir zira doğuştan gelen temayüllerde incelikli bir fark bile yaşam olayına çok farklı tepkilerle sonuçlanabilir. Gelgelelim gelecekte insan benzeri zihinler için daha güvenilir bir değer ekleme mekanizması yaratmak mümkün olabilir (belki de yeni ilaçlar ya da beyin implantları yahut bunların dijital muadillerini kullanarak).

11 Biz insanların bizi yeni nihai değerler edinmeye sevk eden mekanizmayı neden bozmaya çalışmadığımızı merak edebiliriz. Burada birkaç etken devrede olabilir. Birincisi, insan güdülenme sistemi soğukkanlılıkla hesap kitap yapan bir fayda azamileştirici algoritma olarak tanımlanır ki, kötü bir tanımdır bu. İkincisi, değerleri nasıl edindiğimize dair uygun araçlara sahip olmayabiliriz. Üçüncüsü, kimi zaman yeni nihai değerler edinmek için araçsal nedenlere (örneğin toplumsal işaretleşme ihtiyaçlarından doğan nedenlere) sahip olabiliriz: Eğer zihinlerimiz diğer insanlara kısmen şeffaf değilse ya da gerçekte sahip olduğumuzdan farklı birtakım nihai değerlere sahipmişiz gibi davranmanın bilişsel karmaşıklığı gerçekten çok külfetliyse, araçsal değerler o kadar da yararlı olmayabilir. Dördüncüsü, nihai değerlerimizde değişikliklere yol açan eğilimlere aktif olarak direndiğimiz örnekler vardır (örneğin kötü arkadaşların yozlaştırıcı etkisine direnmeye çalıştığımızda). Beşincisi, ilginç bir şekilde, normal insani yollarla yeni nihai değerler edinebilecek türde bir aracı olmaya nihai bir değer atfediyor olabiliriz.

12 Veya güdülenme sistemini yapay zekânın bu tür bir yenilemeye ilgisiz olacağı şekilde tasarlamaya çalışabiliriz, bkz. Armstrong (2010).

13 * Computronium: MIT'den Norman Margolus ve Tommaso Toffoli tarafından ortaya atılan bir kavram. Sanal veya gerçek bir bilgisayar modellemesinin substratı olarak "programlanabilir madde" anlamına gelir. Douglas Adams'ın *Otostopçunun Galaksi Rehberi*'nde tüm dünya böyle bir komputranyum olarak tasvir edilir –yn.

14 Değer öğrenme yaklaşımının önündeki yegâne zorluklar bunlar değildir. Sözgelimi bir diğer mesele de yapay zekâya yeterince akla yatkın ilk inançların nasıl kazandırılacağıdır – en azından programlayıcıların onu düzeltme girişimlerinin altını oyacak kadar güçlü hale geldiğinde.

15 Karş. 9. ve 11. Bölüm.

16 Sözgelimi, MDMA geçici olarak empatiyi artırabilir; oksitoksin de geçici olarak güveni artırabilir (Vollenweider vd. 1998; Bartz vd. 2011). Gelgelelim, etkiler çok değişken ve bağlama bağımlı görünmektedir.

17 Geliştirilmiş aracılar öldürülebilir ya da dondurulabilir, önceki bir duruma resetlenebilir ya da yetkisizleştirilip başka bir geliştirme daha almaları engellenebilir, ta ki sistemin geneli bu önceki haydut öğelerin artık sistem çapında bir tehdit teşkil etmediği daha olgun ve güvenli bir duruma ulaşana kadar.

18 Mesele biyolojik insanların gelecekteki toplumunda –psikolojik manipülasyon için ileri gözetleme ya da biyomedikal tekniklere erişim imkânı olan ya da yurttaşları (ve birbirlerini) gözetleyeceği fevkalade yüksek bir güvenlik elemanı istihdam edebilecek kadar zengin olan toplumlarda– daha da muğlak olabilir.

19 Karş. Armstrong (2007) ve Shulman (2010b).

20 Cevaplanmamış sorulardan biri n düzeyinde bir denetçinin kendi düzeylerindeki (n – 1) denetlenenleri değil, aynı zamanda onların düzeylerindeki (n – 2) denetlenenleri de n – 1 düzeyi aracıların işlerini düzgün bir şekilde yapıp yapmadıklarından emin olmak için hangi derecede izlemesi gerektiğidir. n – 1 düzeyinde aracıların n – 2 düzeyinde aracıları başarıyla idare ettiğini bilmek için n düzeyinde aracının n – 3 düzeyinde aracıları da izlemesi zorunlu mudur?

21 Bu yaklaşım güdülenme seçilimi ile yeterlilik kontrolü arasındaki çizgide gider gelir. Teknik açıdan, birtakım yazılım denetçilerini insanların kontrol etmesinden oluşan düzenlemenin parçası yeterlilik kontrolü sayılırken, diğer katmanları kontrol eden sistem içindeki yazılım aracıları katmanlarından oluşan düzenlemenin parçası güdülenme seçilimidir (sistemin güdülenme eğilimlerini şekillendiren bir düzenleme olduğu ölçüde).

22 Aslında diğer birçok masraf da dikkate alınmayı gerektiriyor, ama burada yapmak mümkün değildir. Örneğin bu tür bir hiyerarşinin tepesinde yer almakla görevli aracıları iktidar yozlaştırabilir.

23 Bu garantinin etkili olması için iyi niyetle uygulanması gerekir. Emülasyonun duygusal ve karar alma yetilerinin belli türde manipülasyonu burada söz konusu olamaz, aksi takdirde sözgelimi durdurulma korkusunu yerleştirmek ya da emülasyonun kendi seçeneklerini akılcı bir tarzda düşünmesini engellemek için kullanılabilir.

24 Örneğin bkz. Brinton (1965); Goldstone (1980, 2001). (Sosyal bilimlerin bu konularda kaydettiği ilerleme dünyanın despotlarına iyi bir hediye olabilir ve bu sayede nüfuslarını kontrol etme stratejilerini optimize etme ve karışıklıkların başını daha küçükken nazik nazik ezme konusunda daha doğru kestirimsel modellerden yararlanabilirler.)

25 Karş. Bostrom (2011a, 2009b).

26 Tamamen yapay bir zekâ örneğinde, gerçekte ayrı alt-aracılar yaratmadan bir kurumsal yapının üstünlüklerinden bazılarını elde etmek mümkün olabilir. Bir sistem karar alma sürecine çoklu perspektifleri dahil ederken bu perspektiflerin her birine bağımsız aracılık için gerekli kendi bilişsel yetiler takımını bahşedebilir. Gelgelelim alt-aracılardan oluşmayan bir sistemdeki bir metinde tarif edilen şu özelliği tamamen uygulamak işleri karıştırabilir: "Bir değişiklik teklifinin davranışsal sonuçlarını gözlemle ve eğer sonuçlar geriye dönüp bakıldığında istenmeyen nitelikteyse önceki bir versiyona geri dön."

ON ÜÇÜNCÜ BÖLÜM

1 Meslekten felsefeciler arasında yapılan yakın tarihli bir ankette çeşitli görüşleri "kabul eden ya da bunlara yakın olan"ların yüzdeleri verilmiştir. Normatif etik konusunda, sonuçlar: "deontoloji %25,9; sonuççuluk %23,6; erdem etiği %18,2. Metaetik konusunda, sonuçlar: Ahlaki gerçekçilik %56,4; ahlaki gerçekçilik karşıtlığı %27,7. Ahlaki yargı konusunda: *Bilişçilik* %65,7; *gayri bilişçilik* %17 (Bourget ve Chalmers 2009).

2 Pinker (2011).

3 Bu mesele hakkında bir değerlendirme için, bkz. Shulman vd. (2009).

4 Moore (2011).

5 Bostrom (2006b).

6 Bostrom (2009b).

7 Bostrom (2011a).

8 Daha kesin konuşmak gerekirse, inançlarımızın daha doğru olduğunu varsaymak için iyi bir nedene sahip olduğumuz konular haricinde onun görüşüne saygı duymak zorundayız. Örneğin belli bir anda ne düşündüğümüz hakkında beyinlerimizi görüntüleyecek olan bir beyinden daha fazlasını bilebiliriz. Gelgelelim süper zekânın bizim görüşlerimize erişim imkânı olduğunu varsayarsak bu nitelendirmeyi atabiliriz; bu durumda, süper zekâya fikirlerimize ne zaman güvenileceğini yargılama görevini de verebiliriz. (Ayrı şekilde ele alınması gereken ve indeks bilgileri içeren özel durumlar kalabilir geride – örneğin süper zekâya kendi perspektifimizden neye inanmanın akılcı olduğunu açıklatmak gibi.) Tanıklık ve epistemik otorite üzerine giderek büyüyen felsefi külliyata giriş için, örneğin bkz. Elga (2007).

9 Yudkowsky (2004). Ayrıca bkz. Mijic (2010).

10 Örneğin, David Lewis *bir inanca dayalı değer teorisi* önermiştir. Buna göre, kabaca söylersek, bir X şey A için ancak ve ancak tamamen akılcıysa ve X hakkında ideal düzeyde bilgisi varsa X'i istemeyi istediği takdirde değere sahiptir; örneğin bkz. Sen ve Williams (1982), Railton (1986) ile Sidgwick ve Jones (2010). Biraz daha benzer bir çizgide, felsefi meşrulaştırmanın bir diğer yaygın izahı olan düşünümlü denge yöntemi belli vakalara dair sezgilerimiz, bu sebepleri yönettiğini düşündüğümüz genel kurallar ve daha bütünlüklü bir sisteme ulaşmak için bu öğeleri gözden geçirmekte kullanılmasını isteyeceğimiz ilkeler arasında yinelemeli bir karşılıklı düzenleme süreci teklif eder; örneğin bkz. Rawls (1971) ve Goodman (1954).

11 Muhtemelen burada kastedilen yapay zekânın bu tür felaketleri engellemek üzere harekete geçtiğinde, bunu olabildiğince yumuşak bir dokunuşla, yani felaketi engelleyecek ama işlerin insanlık için diğer açılardan nasıl şekilleneceğine dair çok fazla etkide bulunmadan yapması gerektiğidir.

12 Yudkowsky (2004).

13 Rebecca Roache, kişisel iletişim.

14 Üç ilke "İnsanları, insanlığın geleceğini ve insancıl doğayı koru" (insancıl burada insani olandan, yani olduğumuz şeyden farklı olarak olmak istediğimiz şeydir); "İnsanlık ebediyetin geri kalanını keşke programlayıcılar farklı bir şey yapsalardı diye umutsuz isteklerin peşinde harcamamalıdır" ve "İnsanlara yardım edin"dir.

15 Bazı dini gruplar akla değil inanca vurgu yaparlar ve aklı –varsayımsal derecede en idealleştirilmiş biçiminde bile ve her kutsal kitabı, şerhi ve vahyi hevesle ve önyargısızca incelese bile– özsel tinsel içgörülerin elde edilmesi için yetersiz görebilirler. Bu tür görüşlere sahip olanlar CEV'i karar alma konusunda optimum bir kılavuz olarak görmeyebilirler (ama yine de CEV yaklaşımından kaçınılsaydı gerçekte takip edilebilecek olan diğer kusurlu kılavuzlara tercih edebilirler).

16 İnsan etkileşimlerini düzenleyecek doğadaki bir saklı güç gibi hareket eden bir yapay zekâya "Sysop" denilebilir: İnsan medeniyetinin işgal ettiği madde için bir tür "işleyen sistem". Bkz. Yudkowsky (2001).

17 "Olabilir" diyorum çünkü insanlığın bütünlüklü dışkestirilmiş iradesinin ahlaki mülahazaları bu varlıkları kapsayacak kadar genişletmeyi istememeleri ***koşuluna bağlı*** olduğundan, bu varlıkların gerçekten bir ahlaki statüye sahip olup olmadıkları belki de şüphelidir (gerçi şu anda bunu istemeleri muhtemel görünüyor ama). "Potansiyel açıdan," diyorum çünkü bir aleyhte oyla CEV dinamiğinin bu dışarlıklıları doğrudan koruması engellense bile, ilk dinamik bir kez başlatıldığında geriye hangi temel kurallar kalırsa kalsın, isteklerine saygı duyulan ve bazı dışarlıklıların refahının korunmasını isteyen bireyler bu sonucu elde etmek için (kendi kaynaklarından bazılarını bırakma pahasına) başarıyla pazarlık yapabilirler, böyle bir olasılık vardır. Bunun mümkün olup olmadığı, CEV dinamiğinin bu türden meselelere müzakereyle çözümlere ulaşmayı mümkün hale getiren birtakım temel kurallar olup olmadığına da bağlı olabilir (bu durum stratejik pazarlık problemlerini aşmak için hükümler gerektirebilir).

18 Güvenli ve hayırhah bir süper zekâyı gerçekleştirmeye pozitif katkı sunan bireyler emeklerinin karşılığı olarak ***bir tür*** özel ödülü hak edebilirler, ama bu ödül insanlığın kozmik birikiminin nasıl tasarruf edileceğine dair neredeyse tekelci bir yetki olmayacaktır. Gelgelelim herkesin dış kestirimimizde eşit pay alacağı düşüncesi öyle hoş bir Schelling noktasıdır ki öyle kolayca bir kenara atılamaz. Her halükârda, erdemin dolaylı yollardan ödüllendirilmesi de mümkündür, yani bizatihi CEV insanlık adına kendilerini paralamış olan iyi insanların uygun şekilde tanınmasını gerektiğini saptayabilir. Bunu eğer –kolaylıkla tahayyül edilebilir olduğu üzere– bizim CEV'imiz adil bir ilke olarak onaylarsa (buna en azından sıfıra eşit olmayan bir ağırlık vermek anlamında) dışkestirim tabanında bu tür insanlara özel bir ağırlık vermeden bunu yapmak mümkün olabilir.

19 Bostrom vd. (yayımlanacak).

20 Ahlaki iddialarda bulunurken ifade edilen (yeterince kesin) bir ortak anlam olduğu ölçüde, bir süper zekâ bu anlamın ne olduğunu çıkarabiliyor olmalıdır. Ve ahlaki iddialar "hakikate meyyal" (yani doğru ya da yanlış olmalarını sağlayan temelde bir önermesel niteliğe sahip) oldukları ölçüde, süper zekâ "X aracısı şu anda ☒ yapmalıdır" biçimindeki iddialardan hangilerinin doğru olduğunu saptayabilmelidir. En azından, bizi bu görevde geride bırakmalıdır. İlk başta bu tür bir ahlaki tanınma kapasitesinden mahrum olan bir yapay zekâ eğer zekâ artırma süper gücüne sahipse bunu edinebilmelidir. Yapay zekânın bunu yapabilmesinin yollarından biri insan beyninin ahlaki düşünmesine ters mühendislik yapmak ve ardından benzer bir süreci uygulamak ama daha hızlı çalıştırmak, onu daha doğru olgusal bilgiyle beslemektir vb.

21 Metaetik konusunda emin olmadığımızdan, eğer ahlaki doğruluğun önkoşulları eksikse ne yapmamız gerektiği sorusu cevapsızdır. Seçeneklerden biri ahlaki bilişselciliğin yanlış olmasına ya da göreli olmayan hiçbir uygun ahlaki hakikat olmamasına yeterince yüksek bir olasılık atfediyorsa yapay zekânın kendi kendini kapatmasını şart koşmaktır. Alternatif olarak, yapay zekâyı CEV gibi alternatif bir yaklaşıma döndürebiliriz.

Ahlaki doğruluğu çeşitli müphem ya da dejenere vakalarda ne yapılması gerektiğini daha açık hale getirecek şekilde inceltebiliriz. Sözgelimi eğer hata teorisi doğruysa (ve dolayısıyla "şimdi Φ yapmam gerekiyor" biçiminin tüm pozitif ahlaki iddiaları yanlışsa), o zaman geri çekilme stratejisine (örneğin kapama) başvurulacaktır. Eğer her biri ahlaken doğru olacak birçok gerçekleştirilebilir eylem varsa ne olacağını da saptayabiliriz. Örneğin bu tür durumlarda yapay zekânın insanlığın kolektif dışkestiriminin tercih edeceği ahlaken kabul edilebilir eylemlerden (birini) yerine getirebileceğini söyleyebiliriz. Keza gerçek ahlak teorisi temel söz dağarında "ahlaken doğru" gibi terimleri kullanmazsa ne olması gerektiğini de şart koşabiliriz. Sözgelimi bir sonuççu teori bazı eylemlerin diğerlerinden daha iyi olduğunu ama bir eylemin "ahlaken doğru" olması fikrine denk düşen tikel bir eşik olmadığını savunabilir. O zaman, eğer böyle bir teori doğruysa, ahlaken doğrunun ahlaken en iyi uygulanabilir eylemlerden birini (eğer varsa) yerine getirmesi gerektiğini söyleyebiliriz; veya sonsuz sayıda uygulanabilir eylem varsa ve her bir uygulanabilir eylemden daha iyisi varsa, o zaman belki de ahlaken doğru bir insanın benzer bir durumda seçeceği en iyi eylemden katbekat daha iyi olabilecek bir eylemi seçebilir; bu olmazsa, o zaman bir insanın yerine getirebileceği en iyi eylem kadar iyi olan bir eylem.

Ahlaken doğru önerisinin nasıl inceltilebileceğine dair düşünürken iki genel hususun akılda tutulması gerekir. Birincisi, muhafazakâr adımlarla başlayıp neredeyse tüm olumsallıkları örtmek için geri çekilme seçeneğinden yararlanabilir ve "ahlaken doğru" seçeneğinden ancak tamamen anladığımızı hissettiğimiz durumlarda yararlanabiliriz. İkincisi, ahlaken doğru önerisine şu genel şerhi düşebiliriz: "Hayırsever bir şekilde yorumlanmalı ve biz onu yazmadan önce daha dikkatli şekilde düşünmüş olsaydık gözden geçirebileceğimiz tarzda gözden geçirilmelidir."

22 Bu terimlerden, "bilgi", biçimsel bir analize (enformasyon-teorik terimleriyle) en açık olan terim olabilir. Gelgelelim bir insanın bir şeyi bilmesinin ne olduğunu temsil etmek için yapay zekâ karmaşık psikolojik özelliklerle alakalı birtakım incelikli temsillere ihtiyaç duyabilir. Bir insan, beyninde bir yerlerde depolanmış olan tüm bilgiyi "bilmez".

23 CEV'deki terimlerin (marjinal düzeyde) o kadar anlaşılmaz olmadığının işaretlerinden biri ahlaki doğruluğu CEV'de kullanılan türde terimlerle analiz edebilirsek felsefi ilerleme olarak sayılacağıdır. Aslında metaetikteki ana eğilimlerden biri (ideal gözlemci teorisi) tam da bunu yapmayı amaçlar. Örneğin bkz. Smith vd. (1989).

24 Bu durum temel normatif belirsizlik problemiyle yüzleşmeyi gerektirir. Hakiki olma olasılığı en yüksek olan ahlak teorisine göre hareket etmenin her zaman uygun olmadığı gösterilebilir. Keza doğru olma olasılığı en yüksek olan eylemi gerçekleştirmenin her zaman uygun olmadığı da gösterilebilir. "Yanlışlık dereceleri"ne ya da söz konusu meselelerin vahametine karşı olasılıkları seçmenin bir yoluna ihtiyaç olduğu görünmektedir. Bu yöndeki bazı fikirler için, bkz. Bostrom (2009a).

25 Ahlaki doğruluk fikrinin herhangi bir açımlaması için herhangi birinin nasıl olup da bir doğru ve yanlış fikrine sahip olabildiğini açıklamasının bir yeterlilik koşulu olduğu bile savunulabilir.

26 *Bizatihi yapay zekânın* her zaman ahlaklı davranacağını varsaysak bile, *bizim için* ahlaken doğru şeyin ahlaki doğruluğu uygulayan bir yapay zekâ yaratmak olduğu bariz değildir. Belki de bu tür bir yapay zekâ yaratmak bizim için itiraz edilebilir derecede kibirlilik ya da kasıntılılık olacaktır (özellikle de birçok insan bu projeyi onaylamayabileceği için). Bu mesele ahlaki doğruluk önerisini eğip bükerek kısmen inceltilebilir. Yapay zekânın ancak yaratıcılarının yapay zekâyı yaratması daha en baştan ahlaken doğruysa harekete geçmesi (ahlaken yapması doğru olan şeyi yapması) gerektiğini, aksi takdirde kendisini kapatması gerektiğini varsayalım. *Bu* türde bir yapay zekâ yaratarak nasıl vahim bir ahlaki yanlış yapacağımızı görmek zordur, zira eğer onu yaratmakla yanıldıysak, yegâne sonuç yapay zekânın bu noktaya kadar hiçbir zihin suçu işlemediğini varsayarak hemen kendi kendini kapatan bir yapay zekânın yaratılması olacaktır. (Gelgelelim yanlış davranmış olabiliriz – sözgelimi bunun yerine başka bir yapay zekâ yaratma fırsatını kaçırarak.)

İkinci bir meseleyse lüzumundan fazlasını yapmaktır. Yapay zekânın yapabileceği birçok eylem olduğunu ve her birinin de ahlaken doğru olduğunu (*ahlaken kabul edilir* olması anlamında) ama bunlardan bazılarının ahlaken diğerlerinden daha iyi olduğunu varsayalım. Seçeneklerden biri yapay zekâya bu türden herhangi bir durumda ahlaken en iyi eylemi (ya da aynı derecede iyi olan birkaç eylem varsa en iyi eylemlerden birini) seçtirmektir. Bir diğer seçenek ise yapay zekâya ahlaken izin verilebilir eylemler arasından (ahlaki olmayan) başka bir gerekliliği en fazla karşılayan eylemi seçtirmektir. Örneğin yapay zekâ ahlaken kabul edilebilir eylemler arasından CEV'imizin tercih edeceği eylemi seçebilir. Bu tür bir yapay zekâ ahlaken kabul edilemez hiçbir şey yapmazken çıkarlarımızı ahlaken en iyi olanı yapan bir yapay zekâdan daha fazla koruyabilir.

27 Yapay zekâ bizim bir yapay zekâ yaratmamızın ahlaken kabul edilebilirliğini değerlendirirken, kabul edilebilirliği nesnel anlamında yorumlamalıdır. "Ahlaken kabul edilebilir"in sıradan bir anlamında, bir doktor hastasını iyileştireceğine inandığı bir ilaç yazdığında ahlaken kabul edilebilir şekilde hareket eder – hasta, doktorun bilgisi dışında, ilaca alerjisi olsa ve bunun sonucunda ölse bile. Nesnel açıdan ahlaki kabul edilebilirliğe odaklanmak yapay zekânın üstün olduğu varsayılan epistemik konumundan istifade eder.

28 Daha doğrudan bir şekilde, yapay zekânın hangi etik teorisinin doğru olduğuna dair *inançlarına* (daha doğrusu, etik teoriler karşısında olasılık dağılımına) bağlıdır.

29 Bu fiziksel açıdan olanaklı hayatların ne kadar muhteşem olabileceğini tahayyül etmek zor olabilir. Buna dair bir fikir aktarmayı amaçlayan şiirsel bir girişim için, bkz. Bostrom (2008c). Bu olanaklardan bazılarının *bizim için*, mevcut insanlar için iyi olabileceğine dair bir argüman için, bkz. Bostrom (2008b).

30 Başka bir önerinin daha iyi olacağını düşünüyorsak bir öneriyi desteklemek aldatıcı ya da manipülatif görünebilir. Ama samimiyetsizlikten kaçınacak şekillerde bunu desteklemek mümkündür. Örneğin bir yandan ideal olanın üstünlüğünü kabul ederken, diğer yandan elde edilebilir en iyi uzlaşma olarak ideal olmayanı desteklemek mümkündür.

31 Veya "iyi", "harika" ya da "muhteşem" gibi pozitif olarak değer biçen başka bir terim.

32 Bu "Ne Diyorsam Onu Yap" olarak bilinen yazılım tasarımındaki bir ilkeyi akla getiriyor. Bkz. Teitelman (1966).

33 Hedef içeriği, karar teorisi ve epistemoloji irdelenmesi gereken üç veçhedir; kesinkes üç ayrı bileşene ayrılmalarının gerekip gerekmediği sorusuna değinmiyoruz.

34 Bir etik proje süper zekânın ürettiği nihai kazançların en fazla mütevazı bir kısmını o projenin başarısına ahlaken kabul edilebilir şekillerde katkıda bulunanlara özel ödüller olarak tahsis edebiliyor olmalıdır. Özendirici ambalajlama şemasına büyük bir kısım tahsis etmek yakışıksız kaçacaktır. Gelirinin yüzde 90'ını sponsorlarının performans bonuslarına ve bağışları artırmak için reklam kampanyalarına harcayan bir hayır kurumuna benzer.

35 Ölüler nasıl ödüllendirilebilir? Birkaç olanağı düşünebiliriz. En alt düzeyde, insanlar ölümden sonra şöhrete ulaşmayı istedikleri ölçüde bir ödül olacak olan anma törenleri ve anıtlar olabilir. Merhumların gelecek hakkında başka tercihleri de olabilir, örneğin kültürler, sanatlar, binalar ya da doğal çevrelerle ilgili. Ayrıca çoğu insan akrabalarına önem verir; katkı sunanların çocuklarına ve torunlarına özel ayrıcalıklar tanınabilir. Daha spekülatif düzeyde, süper zekâ bazı merhum insanların görece sadık simülasyonlarını (bilinçli olacak ve özgün olana –en azından bazı insanların kıstaslarına göre– bir hayatta kalma biçimi olarak sayılacak kadar benzeyen simülasyonları) yaratabiliyor olmalıdır. Örneğin kriyojenik dondurmayla saklanan insanlar için bu muhtemelen daha kolay olacaktır; ama belki de bir süper zekâ için yazışma, yayınlar, görsel-işitsel malzemeler ve dijital kayıtlar gibi diğer saklanmış kayıtlardan ya da diğer hayatta kalanların kişisel anılarından özgün insana çok benzeyen bir şeyi yeniden yaratmak imkânsız olmayacaktır. Bir süper zekâ bizim ha diyince aklımıza gelmeyen bazı olanakları da düşünebilir.

36 Pascalcı gasp için, bkz. Bostrom (2009b). Sonsuz faydalarla alakalı meselelerin bir analizi için, bkz. Bostrom (2011a). Temel normatif belirsizlik hakkında, örneğin bkz. Bostrom (2009a).

37 Örneğin Price (1991); Joyce (1999); Drescher (2006); Yudkowsky (2010); Dai (2009).

38 Örneğin Bostrom (2009a).

39 Yapay zekânın hedef içeriğini saptamak için dolaylı normatiflikten yararlanmanın doğru olmayan bir şekilde saptanmış karar teorisinden doğabilecek problemleri azaltacağı da tasavvur edilebilir. Örneğin CEV yaklaşımını düşünün. Eğer iyi uygulanırsa, yapay zekanın karar teorisinin saptanmasındaki en azından bazı hataları telafi edebiliyor olmalıdır. Bu uygulama bizim bütünlüklü dışkestirilmiş irademizin yapay zekânın peşine düşmesini isteyeceği değerlerin yapay zekânın karar teorisine bağlı olmasına izin verebilir. Eğer idealleştirilmiş benliklerimiz belli türde bir karar teorisinden yararlanan bir yapay zekâ için değer saptamalarında bulunduklarını biliyorlarsa, değer saptamalarını çarpık karar teorisine rağmen iyicil davranacak şekilde düzenleyebilirler (tıpkı bir merceğin bozucu etkilerini karşı yönde bozan başka bir lensi önüne yerleştirerek yok edebileceğimiz gibi).

40 Bazı epistemolojik sistemlerin, bütüncü bir varzda, hiçbir ayrı temeli olmayabilir. Bu durumda, yapısal kalıtım ayrı birtakım ilkeler değil, daha ziyade deyim yerindeyse

kanıt akışına yanıt veren belli temayülleri cisimleştirici epistemik bir başlangıç noktasıdır.

41 Örneğin bkz. Bostrom'da (2011a) ele alınan çarpıtma problemi.

42 Sözgelimi, antropik muhakemedeki tartışmalı meselelerden biri, kendini işaret etme varsayımı adı verilen varsayımın kabul edilip edilmeyeceğidir. Kendini işaret etme varsayımı kabaca şunu savunur: Var olduğunuz olgusundan, daha çok sayıda (N) gözlemcinin var olduğu hipotezlerin N ile orantılı olarak olasılık artışı alması gerektiği sonucunu çıkarmanız gerekir. Bu ilkeye karşı çıkan bir argüman için, bkz. Bostrom'daki (2002a) "Presumptuous Philosopher" düşünce deneyi. Bu ilkeyi savunan bir görüş için ise, bkz. Olum (2002); ve bu savunmanın bir eleştirisi için, bkz. Bostrom ve Ćirković (2003). Kendini işaret etmeyle ilgili inançlar potansiyel açıdan hayati stratejik ilintililiğin çeşitli ampirik hipotezlerini, örneğin Carter-Leslie kıyamet günü argümanı, simülasyon argümanı ve "büyük filtre" argümanları gibi mülahazaları etkileyebilir. Bkz. Bostrom (2002a, 2003a, Bostrom 2008a); Carter (1983); Ćirković vd. (2010); Hanson (1998d); Leslie (1996); Tegmark ve Bostrom (2005). Benzer bir husus da gözlem seçilim teorisindeki diğer yüklü meselelere (gönderme sınıfı tercihinin gözlemci anlarıyla görelileştirilip görelileştirilemeyeceği ve eğer bu yapılabiliyorsa nasıl yapılacağı gibi) ilişkin de ortaya konabilir.

43 Örneğin bkz. Howson ve Urbach (1993). İki Bayesçi aracının görüşleri bilinen bir gerçek olduğunda akli olarak ters düşebilecekleri durumları sayıca azaltan bazı ilginç sonuçlar da vardır; bkz. Aumann (1976) ve Hanson (2006).

44 Karş. Yudkowsky'deki (2004) "son yargıç" kavramı.

45 Epistemolojide öne çıkan birçok önemli mesele vardır ki, bunlardan bazılarını yukarıda ele almıştık. Burada mesele pratikte en iyi sonuçtan ayırt edilemeyecek bir sonuç elde etmek için tüm çözümleri tastamam doğru kılmamızın gerekmeyebileceğidir. Karma bir model (çok çeşitli ilkleri bir araya getiren bir model) işe yarayabilir.

ON DÖRDÜNCÜ BÖLÜM

1 Bu ilke Bostrom'da (2009b, 190) takdim edilir ve totolojik olmadığı da belirtilir. Bir görsel analoji için, olanaklı bir teknoloji yoluyla elde edilebilecek temel yeterlilikler uzamını temsil eden büyük ama hacmi sınırlı bir kutu tasavvur edin. Bu kutuya araştırma çabasını temsil eden kum boşaltıldığını tahayyül edin. Kumu nasıl boşalttığınız kutunun neresinde toplaştığını belirler. Ama boşaltmayı sürdürürseniz, sonunda tüm uzam dolar.

2 Bostrom (2002b).

3 Bilim politikalarına yaklaşımda geleneksel bakış açısı bu değildir. Harvey Averch ABD'de 1945-1984 arasında bilim ve teknoloji politikalarının bilim ve teknoloji teşebbüsüne optimum düzeyde kamusal yatırım tartışmalarına ve ülkenin ekonomik refahını ve askeri gücünü en üst düzeyde artırmak için "kazananlar seçme"yle hükümetin hangi ölçüde ilgilenmesi gerektiğine odaklandığını söyler. Bu hesaplamalarda teknolojik ilerlemenin her zaman iyi olduğu varsayılır. Ama Averch "ilerleme her zaman iyidir" önermesini sorgulayan eleştirel perspektiflerin yükselişini de anlatır (Averch 1985). Ayrıca bkz. Graham (1997).

4 Bostrom (2002b).

5 Bu elbette bir totoloji değildir. Farklı türde bir gelişim argümanının savunulduğu da tahayyül edilebilir. İnsanlığın ilkin sözgelimi nanoteknoloji gelişimine bizi daha iyi kurumlar geliştirmeye, uluslararası açıdan daha eşgüdümlü hale getirmeye ve küresel strateji konusundaki düşüncelerimizi olgunlaştırmaya zorlayacağı gerekçesiyle karşı çıkmasının insanlık için daha iyi olacağı savunulabilir. Belki de makine süper zekâsı kadar metafizik açıdan kafa karıştırıcı olmayan bir tehdit sunan bir zorlukla yüzleşmemiz daha muhtemeldir. Bu durumda, nanoteknoloji (yahut sentetik biyoloji veya ilkin yüzleşeceğimiz daha küçük bir zorluk her neyse) bize süper zekânın yüksek düzey zorluğuyla başa çıkmak için gerekli yeterlilik düzeyine yükselmemizde yardımcı olacak bir basamak hizmeti görebilir.

Bu tür bir argüman vaka vaka analiz edilmelidir. Örneğin nanoteknoloji söz konusu olduğunda, nanofabrike hesaplamalı substratlardan donanım performansındaki itki gibi çeşitli mümkün sonuçları; imalat için ucuz fiziksel sermayenin ekonomik büyümeye etkileri; incelikli gözetleme teknolojisinin bollaşması; bir nanoteknoloji atılımının dolaylı ya da dolaysız sonuçlarıyla bir tekelin ortaya çıkma olanağı ve makine zekâsına nöromorfik ve tüm beyin emülasyonu yaklaşımlarının son derece uygulanabilir olması gibi çeşitli olanaklı sonuçları düşünmek gerekecektir. Tüm bu meseleleri ele almak (ya da diğer varoluşsal riske sebep olan teknolojiler için doğabilecek paralel meseleleri ele almak) bu çalışmanın kapsamını aşıyor. Burada ilk bakışta gelişimin önce süper zekâdan başlaması görüşüne yakın duruyoruz, ama bazı vakalarda bu ilk değerlendirmeyi değiştirebilecek karmaşıklıklar olduğunu da vurguluyoruz.

6 Pinker (2011); Wright (2001).

7 Her şeyin hızlandığı hipotezini ilk bakışta herhangi bir gözlemsel sonucu olmadığı göründüğü için anlamsız varsaymak çekici olabilir; ama örneğin bkz. Shoemaker (1969).

8 Hazırlıklı olma düzeyi hazırlık faaliyetlerine ayrılan çaba miktarıyla değil, gerçekte koşulların ne kadar uygun şekilde hazırlandığı ve kilit karar alıcıların uygun eylemde bulunmaya ne kadar hazır olduklarıyla ölçülür.

9 Zekâ patlamasına giden süreçte uluslararası güven derecesi de bir etken olabilir. Bunu aşağıda "İşbirliği" altbaşlığında ele alıyoruz.

10 Bir anekdot olarak, kontrol problemiyle şu anda ciddi olarak ilgilenenlerin zekâ dağılımının aşırı bir ucundan orantısız şekilde örnek olarak seçildiği, ama bu izlenim için alternatif açıklamalar olabileceği gözükmektedir. Eğer alan moda haline gelirse, kuşkusuz vasatlıklarla ve çatlaklarla dolup taşacaktır.

11 Bu terimi Carl Shulman'a borçluyum.

12 Bir makine zekâsının bir nöromorfik yapay zekâdan ziyade bir tüm beyin emülasyonu sayılabilmesi için bir beyne ne kadar benzer olması gerekir? İlintili belirleyici öğe, sistemin değerleri ya da tikel bir bireyin veya türsel bir insanın bilişsel ve değerlendirmeci eğilimlerinin hepsini yeniden üretip üretmeyeceği olabilir, çünkü bunun kontrol problemi için bir fark yaratması akla yatkındır. Bu özellikleri yakalamak son derece yüksek derecede bir emülasyon sadakati gerektirebilir.

13 Artışın büyüklüğü elbette itmenin ne kadar büyük olduğuna ve ayrıca itme kaynaklarının nereden geldiğine bağlı olacaktır. Eğer tüm beyin emülasyonu araştırmalarına yatırılan fazladan kaynakların hepsi düzenli nörobilim araştırmalarından çıkarılırsa, nörobilim için hiçbir net artış olmayabilir, meğerki emülasyon araştırmalarına daha sıkı bir odaklanmanın nörobilim araştırmalarının varsayılan portföyüne kıyasla nörobilimi ilerletmenin daha etkili bir yolu olduğu görülmesin.

14 Bkz. Drexler (1986, 242). Drexler (kişisel iletişim) bu yeniden inşanın sunmayı amaçladığı muhakemeye tekabül ettiğini doğrular. Belli ki, argümanı tümdengelimsel açıdan geçerli bir muhakeme zinciri biçiminde sunmak istiyorsak, birtakım zımni önermelerin eklenmesi gerekecektir. (Drexler'ın aşağıda ele alınan "ilintili argüman türü"nü onaylamadığını belirtelim.)

15 Belki de *küçük* felaketlere hüsnükabul göstermemiz gerekir, zira bizi varoluşsal felaketleri engellemek için zorunlu güçlü önlemleri almak adına ihtiyaç duyulacak *orta ölçekli* felaketleri engelleyecek kadar müteyakkız hale getirecektir. (Ama elbette tıpkı biyolojik bağışıklık sistemlerinde olduğu gibi, alerjilere ve otoimmun bozukluklara olduğu gibi aşırı tepkiler konusunda kaygılanmamız da gerekiyor.)

16 Karş. Lenman (2000); Burch-Brown (2014).

17 Karş. Bostrom (2007)

18 Bu argümanın ilintili olayların zamanlamasından ziyade düzenine odaklandığını belirtelim. Süper zekâyı daha önce yaratmak ancak müdahale kilit gelişmelerin silsilesini değiştirirse diğer varoluşsal geçiş risklerini boşa çıkaracaktır: Örneğin nanoteknolojide ya da sentetik biyolojide belli dönüm noktalarına ulaşmadan önce süper zekâ gerçekleştirilebilirse.

19 Eğer kontrol problemini çözmek makine zekâsında performans problemini çözmeye kıyasla *son derece* zorsa ve eğer proje yeteneği proje büyüklüğüyle sadece zayıf bir bağıntıya sahipse, o zaman küçük bir projenin oraya daha önce varmasının daha iyi olması mümkündür, yani yeterlilikteki çeşitlilik küçük projeler arasında daha çoksa. Bu tür bir durumda, küçük projeler ortalamaya bakıldığında büyük projeler kadar yeterliliğe sahip değilse, verili bir küçük projenin kontrol problemini çözmek için gerekli fevkalade yüksek yeterlilik düzeyine sahip olması daha düşük bir olasılık olabilir.

20 Buradan küresel fikir teatisini teşvik edebilecek ve donanımdaki ek ilerlemelerden (örneğin üst düzey tercüme, daha iyi arama, akıllı telefonlara her an erişim, toplumsal münasebet için çekici sanal gerçeklik çevreleri vb.) yarar sağlayacak, hatta bu ilerlemeyi gerektirecek aletler tahayyül edilebileceğini reddedeceğimiz sonucu çıkmaz.

21 Emülasyon teknolojisine yatırım tüm beyin emülasyonuna doğru ilerlemeyi sadece doğrudan (üretilen tüm teknik teslim edilebilirler yoluyla) değil, aynı zamanda dolaylı olarak, daha fazla fon için bastıracak ve tüm beyin emülasyonu vizyonunun görünürlüğünü ve anlaşılırlığını artıracak bir kamuoyu yaratarak da hızlandırılabilir.

22 Eğer geleceği tüm insanlığın arzularından (bunların uygun bir şekilde üst üste binmesinden) ziyade rastgele bir insanın arzuları şekillendiriyorsa, beklenen değerin ne kadarı kaybedilmiş olacaktır? Bu durum, hangi değerlendirme standardından

yararlandığımıza ve ayrıca söz konusu arzular idealleştiriliyor mu yoksa ham mıdır, ona bağlı olabilir.

23 Örneğin insan zihinleri dil yoluyla yavaş iletişim kurarken, yapay zekâlar aynı programın örneklerinin hem yetenekleri hem de bilgiyi kendi aralarında kolay ve çabuk bir şekilde nakledebilecekleri tarzda tasarlanabilir. Baştan tasarlanan makine zihinleri siber-uzayda önemsiz olan doğal çevre veçheleriyle atalarımızın başa çıkmasına yardımcı olmuş olan hantal miras sistemlerinden kurtulabilir. Dijital zihinler biyolojik beyinler için mevcut olmayan hızlı sıralı işlemden istifade edecek ve son derece optimize edilmiş işlevsellikle (örneğin sembolik işleme, örüntü tanıma, simülatörler, veri madenciliği ve planlama) yeni modülleri kurmayı kolaylaştıracak şekilde tasarlanabilir. Yapay zekânın teknolojik olmayan üstünlükleri de olabilir (patentinin daha kolay alınabilir olması ya da insan yüklemeleri kullanmanın ahlaki karmaşıklıklarına daha az bulaşmış olmak gibi).

24 Eğer $p1$ ve $p2$ her bir adımda başarısızlığın olasılıklarıysa, başarısızlığın toplam olasılığı $p1 + (1 - p1)p2$'dir, zira insan ölümcül olarak ancak bir kez başarısız olabilir.

25 Önde koşanın bu kadar büyük bir üstünlüğe sahip olmaması ve bir tekel yaratamayabilecek olması mümkündür. Ayrıca WBE'nin müdahalesi bile olmadan yapay zekâdan önce bir tekelin ortaya çıkması da mümkündür ki, bu durumda öncelikle WBE'nin gelmesi gerektiğini savunan senaryo çuvallar.

26 WBE savunucusu için, yapay zekâ gelişimine sıçramayı asgariye çekerken WBE'yi hızlandıracak şekilde desteğinin özgüllüğünü artırmanın bir yolu var mıdır? Tarama teknolojisini teşvik etmek muhtemelen nöro-hesaplamalı modellemeyi teşvik etmekten daha iyi bir kumardır. (Bilgisayar donanımını teşvik etmenin şu ya da bu şekilde herhangi bir fark yaratması, bu alanda her halükârda ilerlemeyi teşvik eden büyük ticari çıkarlar düşünüldüğünde, muhtemel değildir.)

Tarama teknolojisini teşvik etmek taramanın bir darboğaz olma ihtimalini azaltarak çokkutuplu bir sonuç ihtimalini artırabilir ve böylece ilk emülasyon nüfusunun sürüsüne bereket küçük sayıda şablon kopyalarından oluşacağına çok farklı insan kopyalarından çıkarılması şansını artıracaktır. Tarama teknolojisindeki ilerlemenin darboğaza bilgisayar donanımının girme ihtimalini artırması da muhtemeldir, zira muhtemelen yavaş bir havalanma gerçekleşecektir.

27 Nöromorfik yapay zekâ bir biyolojik insanınkine benzer bilişsel zaafları ve güçlü yanları olan bir profile sahip olmak gibi tüm beyin emülasyonunun diğer güvenliği teşvik eden özelliklerinden de mahrum olabilir (bu bize gelişiminin farklı aşamalarında sistemin yeterliliklerine dair bazı beklentiler oluşturmak için deneyimimizden yararlanmamıza izin verir).

28 Eğer bir insanın WBE'yi teşvik etme güdüsü WBE'nin yapay zekâdan önce gerçekleşmesini sağlamaksa, WBE'yi hızlandırmanın varış saatini ancak makine zekâsına iki yolun varsayılan zamanı yakınsa ve yapay zekâya biraz daha meylediyorsa değiştireceğini akıllarında tutmaları gerekir. Aksi takdirde, *ya* WBE yatırımı WBE'yi gelişim silsilesini etkilemeden ne olacaksa o hale getirecektir (donanım sarkmasını ve hazırlık zamanını azaltacaktır) *ya da* WBE'ye bu tür bir yatırım çok az etkiye sahip olacaktır (nöromorfik yapay zekâda ilerlemeyi uyararak yapay zekânın belki de çok daha erken gerçekleştirilmesi hariç).

29 Hanson (2009) üzerine yorum.

30 Elbette riski ertelemek için insanı etkileyen perspektiften bile tercih edilebilir olacağı *bir* varoluşsal risk büyüklüğü ve yakınlaşması elbette olacaktır – ister var olan insanların perde kapanmadan önce biraz daha hayat elde edebilmeleri için ister tehlikeyi azaltabilecek hafifletme çabalarında daha fazla zaman kazanmak için olsun.

31 Zekâ patlamasını bir yıl daha yakınlaştıracak bir eylemde bulunduğumuzu varsayın. Şu an dünyada yaşayan insanların yılda yüzde 1 oranında öldüklerini ve zekâ patlamasından ötürü insanın yok olacağına dair varsayılan riskin (sırf örnek olsun diye keyfi bir sayı seçmek gerekirse) yüzde 20 olduğunu varsayalım. Bu durumda, zekâ patlamasının gelişinin bir yıl gecikmesi riskin (insanı etkileyen bakış açısından) yüzde 20'den yüzde 21'e yükselmesi, yani risk düzeyinde yüzde 5'lik bir artış anlamına gelebilir. Gelgelelim zekâ patlamasının başlangıcından bir yıl önce hayatta olan insanların ezici çoğunluğu bu noktada eğer patlama riskini yüzde bir azaltabilecekse bunu ertelemek işine gelecektir (zira birçok birey bir sonraki yılda ölme risklerinin yüzde birden az olduğunu görecektir – zira birçok ölüm hasta ve yaşlılar gibi görece dar bir nüfus kesiminde gerçekleşir). Dolayısıyla nüfusun her yıl zekâ patlamasını bir yıl ertelemek için oy kullandığı bir modele sahip olabiliriz, böylece yaşayan herkes zekâ patlamasının bir noktada yaşanmasının iyi olacağı konusunda hemfikir olsa da zekâ patlaması hiçbir zaman yaşanmaz. Gerçekte, elbette, eşgüdüm başarısızlıkları, sınırlı kestirilebilirlik ya da kişisel olarak hayatta kalmaktan başka şeylerin tercih edilmesinin böyle sonu gelmeyen bir duraklamayı engellemesi muhtemeldir.

Eğer kişiyi etkileyen standarttan ziyade standart bir ekonomik iskonto etkeninden yararlanılırsa, potansiyel üstünlüğün büyüklüğü azalır, zira muazzam uzun zamanlarda hayatta kalan mevcut insanların değeri ciddi derecede azalacaktır. Eğer tenzilat etkeni yıldız zamanından ziyade her bir bireyin öznel zamanına uygulanırsa bu etken bilhassa güçlüdür. Eğer gelecekte elde edilecek yararlar yılda %x oranında azalırsa ve diğer kaynaklardan varoluşsal riskin arkaplan düzeyi yılda %y ise, o zaman zeka patlaması için optimum nokta patlamanın bir yıl daha ertelenmensin bir zeka patlamasıyla alakalı varoluşsal riskin azalma noktalarını yüzde x + y oranında azaltacak olmasıdır.

32 Hatta varoluşsal bir risk bile sunabilir, özellikle de öncesinde yeni kitle imha teknolojileri bulunmuş ya da eşi görülmedik bir silahlanma artışı yaşanmışsa.

33 Bir proje, işçilerini çok sayıda farklı mekâna dağıtmış ve şifreli iletişim kanallarıyla işbirliği yapıyor olabilir. Fakat bu taktik bir güvenlik ödünleşimi içerir: Coğrafi dağılım askeri saldırılara karşı belli bir koruma önerebilirse de, operasyon güvenliğine engel teşkil edecektir, zira personelin birçok mekana yayılmışsa rakip bir güç tarafından kaçırılmasını, işi bırakmasını ya da bilgi sızdırmasını engellemek zordur.

34 Büyük bir zamansal iskonto etkeninin bir projenin sanki bir yarıştaymış gibi davranmasını sağlayabileceğini (velev ki gerçek bir rakibi olmadığını bilsin) belirtelim. Büyük tenzilat etkeni uzak geleceği çok az umursayacağı anlamına gelir. Duruma bağlı olarak, mavi gökyüzü araştırmalarının hevesini kıracaktır ve bu da makine zekâsı devrimini geciktirecektir (ama belki de donanım sarkmasından ötürü gerçekleştiğinde gerçekleşmesi daha ani olacaktır). Ama büyük tenzilat etkeni (ya da gelecek kuşakların çok az umursanması) varoluşsal risklerin daha az önem arz etmesine yol açacaktır. Bu durum artan varoluşsal felaket riski pahasına doğrudan kazanım olanağını içeren kumarları teşvik edecek, böylece güvenlik yatırımına olan teşviki keserken erken bir kalkışmayı teşvik edecektir (yarış dinamiğinin etkilerini taklit ederek). Gelgelelim

yarış dinamiğinin aksine büyük bir iskonto etkeninin (ya da gelecek kuşakların umursanmamasının) bilhassa çatışmayı körükleyeceği söylenemez.

Yarış dinamiğini azaltmak işbirliğinin ana kazanımlarından biridir. Bu işbirliği kontrol problemini çözmenin neden bir kazanım olduğuna dair fikirlerin paylaşılmasını kolaylaştıracaktır, ama işbirliğinin yeterlilik probleminin nasıl çözüleceğine dair fikirlerin paylaşılmasını da kolaylaştıracak olması nedeniyle belli ölçüde dengelenecektir. Fikir paylaşımının kolaylaşmasının net etkisi ilintili araştırma camiasının kolektif zekâsını birazcık artırmak olabilir.

35 Diğer yandan, tek bir hükümetin kamusal nezareti tek bir ulusun kazançları tekelleştirdiği bir sonuç doğurma riski barındırır. Bu sonuç esrarengiz diğerkâmların herkesin kazanacağını garanti altına aldıkları sonuçtan daha geride görünmektedir. Ayrıca bir ulusal hükümetin denetimi bu ülkenin tüm yurttaşlarının bile mutlaka kazançtan pay alacakları anlamına gelmez: Söz konusu ülkenin hangisi olduğuna bağlı olarak, tüm kazançların bir siyasi elit ya da kendi kendine hizmet eden bir avuç aracı tarafından gasp edilmesi gibi az ya da çok bir risk vardır.

36 Özendirici ambalajlamadan yararlanmak (12. Bölüm'de ele alındığı üzere) bazı koşullarda insanları pasif beleşçilerden ziyade aktif işbirlikçiler olarak bir projeye katılmaya teşvik edebilir.

37 Azalan getirilerin çok daha küçük ölçekte olacağı görünecektir. Birçok insan bir milyar yıldızlı bir galakside milyarda bir şansa sahip olmaktansa tek bir yıldıza sahip olmayı tercih edecektir. Gerçekten de birçok insan tüm gezegene milyarda bir sahip olma şansındansa yeryüzündeki kaynakların milyarda birine sahip olmayı tercih edecektir.

38 Karş. Shulman (2010a).

39 Toplulaştırıcı etik teoriler evrenin sonsuz olabileceği fikri ciddiye alındığında sıkıntı yaşar; bkz. Bostrom (2011b). Gülünç derecede büyük ama sonlu değerler fikri ciddiye alındığında da sıkıntılar doğabilir; bkz. Bostrom (2009b).

40 Eğer bir bilgisayar daha büyük hale getirilirse, sonunda bilgisayarın farklı parçaları arasındaki iletişim gecikmelerinden doğan görececi kısıtlamalarla karşılaşılır: Sinyaller ışıktan daha hızlı yayılmaz. Eğer bilgisayar küçültülürse, minyatürleşmeye kuantum sınırlarıyla karşılaşılır. Eğer bilgisayarın yoğunluğu artırılırsa, kara delik sınırına düşülür. Kabul etmek gerekir ki, bu sınırlamaların etrafından bir yol sunan yeni fiziğin bir gün keşfedileceğinden tamamen emin olamayız.

41 Bir insanın kopyalarının sayısı üst sınırı olmayan kaynaklarla doğrusal olarak ölçeklenecektir. Ama ortalama bir insanın kendisine ilişkin çoklu kopyalara sahip olmayı ne kadar değerli göreceği açık değildir. Çoklu olarak somutlaştırılmayı tercih eden insanlar bile artan sayıda kopyayla doğrusal olan bir fayda işlevine sahip olmayabilir. Kopya sayıları, tıpkı ömrün yılları gibi, tipik bir insanın fayda işlevinde azalan getirilere sahip olabilir.

42 Bir tekel en yüksek karar alma düzeyinde içsel olarak son derece işbirlikçidir. Bir tekel eğer tekeli oluşturan üst düzey aracılık işlerin bu şekilde olmasını tercih ederse aşağı düzeylerde işbirliğinden epey kaçınabilir.

43 Eğer her rakip yapay zekâ takımı diğer takımların bir zekâ patlaması üretmesi konusunda hiçbir şansa sahip olmayacağı şekilde yanlış yönlendirildiğine emin olursa,

o zaman işbirliği için nedenlerin birinden (yarış dinamiğinden kaçınmak) kurtulmuş oluruz: Her takım ciddi bir rakibi olmadığından emin olmanın verdiği güvenle daha yavaş gitmeyi tercih edebilir.

44 Bir doktora öğrencisi.

45 Bu formülasyon bugün var olan ya da gelecekte var olabilecek insan olmayan hayvanlara ve diğer acı çekebilen varlıklara (dijital varlıklar dahil) gereken önemin verilmesini savunan bir tarifi kapsayacak şekilde okunmalıdır. Bunu bir yapay zekâ geliştiricisinin genel ahlaki topluluğun ahlaki kurumlarını kendi ahlaki kurumlarıyla ikame etmesi için bir icazet verilmesi şeklinde okumamak gerekir. Bu ilke 12. Bölüm'de ele alınan "bütünlüklü dışkestirilmiş irade" yaklaşımıyla tutarlıdır ve dışkestirim tabanı tüm insanları kapsar.

Biraz daha netleştirmemiz gerekirse: Bu formülasyonun yapay zekâlarda ya da onların kurucu algoritmalarında ve veri yapılarında geçiş sonrası mülkiyet haklarının olanaklılığını dışlamadığını belirtelim. Bu formülasyon geleceğin varsayımsal bir insan sonrası toplumunda hangi hukuki ya da siyasi sistemlerin alışverişlerin örgütlenmesine en iyi şekilde hizmet edeceği konusunda bilinemezcidir. Formülasyonun kastettiği bu tür bir sistemin tercihinin, seçimi süper zekânın başlangıçta nasıl geliştirildiği tarafından nedensel olarak belirlendiği ölçüde, belirtilen kıstas temelinde yapılması gerektiğidir, yani geçiş sonrası yönetim sistemi sözgelimi süper zekâyı hasbelkader ilk kim geliştirdiyse ona yarar sağlayacak şekilde değil, tüm insanlığın yararı için ve geniş kesimlerin paylaştığı etik ideallerin hizmetine koşulacak şekilde seçilmelidir.

46 W maddesinde inceltmeler yapmanın mümkün olduğu açıktır. Örneğin eşik belki de kişi başına terimleriyle ifade edilmelidir veya kazananın daha fazla üretimi güçlü bir şekilde teşvik etmek için aşırılıktan payına düşenden daha fazlasını almasına izin verilmelidir (Rawls'ın maksimin ilkesinin bir versiyonu burada çekici olabilir). Diğer inceltmeler odak noktasını dolar miktarından uzaklaştıracak ve bunu "insanlığın geleceğine etkisi" ya da "farklı tarafların çıkarlarının gelecekteki bir tekelin fayda işlevindeki ağırlık dereceleri" üzerinden vb. ifade edecektir.

ON BEŞİNCİ BÖLÜM

1 Belli bir araştırmaya keşfettiklerinden ötürü değil, başka nedenlerle (örneğin ilgilenenleri eğlendirmek, öğretmek, güven vermek ya da morallerini düzeltmek için) ihtiyaç var.

2 *Hiç kimse* saf matematikle ya da felsefeyle ilgilenmesin demiyorum. Keza bu çabaların akademideki ya da toplumun genelindeki diğer enerji kayıplarıyla kıyaslandığında özellikle israf olduğunu da söylemiyorum. Bazı insanların zihninin yaşamına kendilerini adayabilmeleri ve faydasına ya da etkisine bakmadan meraklarını gittiği yere kadar götürmeleri muhtemelen çok iyidir. Burada kastedilen, kıyıda köşede, en iyi zihinlerden bazılarının bilişsel performanslarının öngörülebilir gelecekte köhnemiş hale gelebileceğini fark ederek, dikkatlerini çözüme biraz daha çabuk ulaşmamızın bir fark yaratacağı teorik problemlere çevirmek isteyip istemedikleridir.

3 Bu belirsizliğin koruyucu olabileceği örneklerde ihtiyatı elden bırakmamak gerekir: Sözgelimi Kutu 13'teki risk yarışı modelini hatırlayın, orada ek stratejik bilginin zararlı olabileceğini görmüştük. Daha genel olarak, bilginin tehlikeleri konusunda

endişe duymamız gerekiyor (Bkz. Bostrom [2011b]). Bilginin tehlikelerine dair daha fazla analize ihtiyaç duyduğumuzu söylemek çekicidir. Bu muhtemelen doğrudur, ama yine de bu tür bir analizin tehlikeli bilgiler doğurabileceği konusunda da endişelenebiliriz.

4 Karş. Bostrom (2007).

5 Carl Shulman'a bu hususu vurguladığı için minnettarım.

Kaynakça

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1 (1): 1–37.

Albertson, D. G. ve Thomson, J. N. 1976. "The Pharynx of Caenorhabditis Elegans." *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938): 299–325.

Allen, Robert C. 2008. "A Review of Gregory Clark's A Farewell to Alms: A Brief Economic History of the World." *Journal of Economic Literature* 46 (4): 946–73.

American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Erişim Tarihi: 30 Temmuz 2013. Bkz. http://www.horsecouncil.org/national-economicimpact-us-horse-industry.

Anand, Paul, Pattanaik, Prasanta ve Puppe, Clemens, haz. 2009. *The Oxford Handbook of Rational and Social Choice*. New York: Oxford University Press.

Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W. ve Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4): 796–805.

Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer*, 19 Mart.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Yayımlanmamış elyazması, 20 Ekim. Erişim Tarihi: 31 Aralık 2012. Bkz. http://www.neweuropeancentury.org/GodAI.pdf.

Armstrong, Stuart. 2010. *Utility Indifference*, Technical Report 2010-1. Oxford: Future of Humanity Institute, Oxford Üniversitesi.

Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12: 68–84.

Armstrong, Stuart ve Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89: 1–13.

Armstrong, Stuart ve Sotala, Kaj. 2012. "How We're Predicting AI—or Failing To." *Beyond AI: Artificial Dreams* içinde, haz. Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak ve Radek Schuster, 52–75. Pilsen: University of West Bohemia. Erişim Tarihi: 2 Şubat 2013.

Asimov, Isaac. 1942. "Runaround." *Astounding Science Fiction*, Mart, 94–103.

Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236–9.

Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore: Johns Hopkins University Press.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Jacob, W., Lent, R. ve Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain." *Journal of Comparative Neurology* 513 (5): 532–41.

Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.

Baratta, Joseph Preston. 2004. *The Politics of World Federation: United Nations, UN Reform, Atomic Control*. Westport, CT: Praeger.

Barber, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean*. Princeton, NJ: Princeton University Press.

Bartels, J. vereasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J. ve Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." *Journal of Neuroscience Methods* 174 (2): 168–76.

Bartz, Jennifer A., Zaki, Jamil, Bolger, Niall ve Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." *Trends in Cognitive Science* 15 (7): 301–9.

Basten, Stuart, Lutz, Wolfgang ve Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." *Demographic Research* 28: 1145–66.

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Baum, Seth D., Goertzel, Ben ve Goertzel, Ted G. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185–95.

Beal, J. ve Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence." *IEEE Intelligent Systems* 24 (4): 21–3.

Bell, C. Gordon ve Gemmell, Jim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything*. New York: Dutton.

Benyamin, B., St. Pourcain, B., Davis, O. S., Davies, G., Hansell, M. K., Brion, M.-J. A., Kirkpatrick, R. M., v.d. 2013. "Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L." *Molecular Psychiatry* (23 Ocak).

Berg, Joyce E. ve Rietz, Thomas A. 2003. "Prediction Markets as Decision Support Systems." *Information Systems Frontiers* 5 (1): 79–93.

Berger, Theodore W., Chapin, J. K., Gerhardt, G. A., Soussou, W. V., Taylor, D. M. ve Tresco, P. A., haz. 2008. *Brain–Computer Interfaces: An International Assessment of Research and Development Trends*. Springer.

Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A. ve Granacki, J. J. 2012. "A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 198–211.

Berliner, Hans J. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2): 205–220.

Berliner, Hans J. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69: 6–9.

Bernardo, Jos. M. ve Smith, Adrian F. M. 1994. *Bayesian Theory*, 1. Baskı. New York: Wiley.

Birbaumer, N., Murguialday, A. R. ve Cohen, L. 2008. "Brain–Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6): 634–8.

Bird, Jon ve Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." *Proceedings of the 2002 Congress on Evolutionary Computation* içinde, 2: 1836–41.

Blair, Clay, Jr. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country." *Life*, 25 Şubat, 89–104.

Bobrow, Daniel G. 1968. "Natural Language Input for a Computer Problem Solving System." *Semantic Information Processing* içinde, haz. Marvin Minsky, 146–227. Cambridge, MA: MIT Press.

Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Yayımlanmamış elyazması. Son ziyaret 19 Eylül 1998.

Bostrom, Nick. 2002a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Bostrom, Nick. 2002b. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.

Bostrom, Nick. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–55.

Bostrom, Nick. 2003b. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308–314.

Bostrom, Nick. 2003c. "Ethical Issues in Advanced Artificial Intelligence." *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* içinde, haz. Iva Smit and George E. Lasker, 2: 12–17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Bostrom, Nick. 2004. "The Future of Human Evolution." *Two Hundred Years After Kant* içinde, Fifty Years After Turing, haz. Charles Tandy, 2: 339–371. *Death and Anti-Death*. Palo Alto, CA: Ria University Press.

Bostrom, Nick. 2006a. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1): 11–30.

Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16 (2): 185–200.

Bostrom, Nick. 2006c. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48–54.

Bostrom, Nick. 2007. "Technological Revolutions: Ethics and Policy in the Dark." *Nanoscale: Issues and Perspectives for the Nano Century* içinde, haz. Nigel M. de S. Cameron ve M. Ellen Mitchell, 129–52. Hoboken, NJ: Wiley.

Bostrom, Nick. 2008a. "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing." *MIT Technology Review*, Mayıs/Haziran sayısı, 72–7.

Bostrom, Nick. 2008b. "Why I Want to Be a Posthuman When I Grow Up." *Medical Enhancement and Posthumanity* içinde, haz. Bert Gordijn ve Ruth Chadwick, 107–37. New York: Springer.

Bostrom, Nick. 2008c. "Letter from Utopia." *Studies in Ethics, Law ve Technology* 2(1): 1–7.

Bostrom, Nick. 2009a. "Moral Uncertainty – Towards a Solution?" Overcoming Bias (blog),1 Ocak.

Bostrom, Nick. 2009b. "Pascal's Mugging." *Analysis* 69 (3): 443–5.

Bostrom, Nick. 2009c. "The Future of Humanity." *New Waves in Philosophy of Technology* içinde, haz. Jan Kyrre Berg Olsen, Evan Selinger ve Siren Riis, 186–215. New York: Palgrave Macmillan.

Bostrom, Nick. 2011a. "Information Hazards: A Typology of Potential Harms from Knowledge." *Review of Contemporary Philosophy* 10: 44–79.

Bostrom, Nick. 2011b. "Infinite Ethics." *Analysis and Metaphysics* 10: 9–59.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." "Theory and Philosophy of AI," içinde, haz. Vincent C.Mu¨ller, özel sayı, *Minds and Machines* 22 (2): 71–85.

Bostrom, Nick ve Ćirković, Milan M. 2003. "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53 (210): 83–91.

Bostrom, Nick ve Ord, Toby. 2006. "The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics." *Ethics* 116 (4): 656–79.

Bostrom, Nick ve Roache, Rebecca. 2011. "Smart Policy: Cognitive Enhancement and the Public Interest." *Enhancing Human Capacities* içinde, haz. Julian Savulescu, Ruudter Meulen ve Guy Kahane, 138–49. Malden, MA: Wiley-Blackwell.

Bostrom, Nick ve Sandberg Anders. 2009a. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311–41.

Bostrom, Nick ve Sandberg Anders. 2009b. "The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement." *Human Enhancement* içinde, 1. baskı, haz. Julian Savulescu ve Nick Bostrom, 375–416. New York: Oxford University Press.

Bostrom, Nick, SandbergAnders ve Douglas, Tom. Yayınlanacak. "The Unilateralist's Curse: The Case for a Principle of Conformity." *Social Epistemology*.

Bostrom, Nick ve Yudkowsky, Eliezer. 2015. "The Ethics of Artificial Intelligence." *Cambridge Handbook of Artificial Intelligence* içinde, haz. Keith Frankish ve William M. Ramsey, 315–334. New York: Cambridge University Press.

Boswell, James. 1917. *Boswell's Life of Johnson*. New York: Oxford University Press.

Bouchard, T. J. 2004. "Genetic Influence on Human Psychological Traits: A Survey." *Current Directions in Psychological Science* 13 (4): 148–51.

Bourget, David ve Chalmers, David. 2009. "The PhilPapers Surveys." Kasım. Bkz. http://philpapers.org/surveys/.

Bradbury, Robert J. 1999. "Matrioshka Brains." Arşiv versiyonu. 16Ağustos 2004'te gözden geçirilmiş haliyle. Bkz. http://web.archive.org/web/20090615040912/http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrainsPaper.html.

Brinton, Crane. 1965. *The Anatomy of Revolution*. Gözden geçirilmiş baskı. New York: Vintage Books.

Bryson, Arthur E., Jr. ve Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, MA: Blaisdell.

Buehler, Martin, Iagnemma, Karl ve Singh, Sanjiv, haz. 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic. Springer Tracts in Advanced Robotics* 56. Berlin: Springer.

Burch-Brown, J. 2014. "Clues for Consequentialists." *Utilitas* 26 (1): 105–19.

Burke, Colin. 2001. "Agnes Meyer Driscoll vs. the Enigma and the Bombe." Yayımlanmamış elyazması. Erişim Tarihi: 22 Şubat 2013. Bkz. http://userpages.umbc.edu/~burke/driscoll1-2011.pdf.

Canback, S., Samouel, P. ve Price, D. 2006. "Do Diseconomies of Scale Impact Firm Sizeand Performance? A Theoretical and Empirical Overview." *Journal of Managerial Economics* 4 (1): 27–70.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov,D. F., Patil, P. G., Henriquez, C. S. ve Nicolelis, M. A. 2003. "Learning to Control a Brain–Machine Interface for Reaching and Grasping by Primates." *Public Library of Science Biology* 1 (2): 193–208.

Carroll, Bradley W. ve Ostlie, Dale A. 2007. *An Introduction to Modern Astrophysics*. 2. Baskı. San Francisco, CA: Pearson Addison Wesley.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Carter, Brandon. 1983."The Anthropic Principle and its Implications for Biological Evolution."*Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512): 347–63.

Carter, Brandon. 1993. "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis." *The Anthropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy* içinde, haz. F. Bertola and U. Curi, 33–66. Cambridge: Cambridge University Press.

CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. *Findings Regarding the Market Events of May 6, 2010: Report ofthe Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65.

Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H. ve Armant, D. R. 2011. "Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development." *Trends in Endocrinology and Metabolism* 22 (10): 412–20.

Chen, S. ve Ravallion, M. 2010. "The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty." *Quarterly Journal of Economics* 125 (4): 1577–1625.

Chislenko, Alexander. 1996. "Networking in the Mind Age: Some Thoughts on Evolutionof Robotics and Distributed Systems." Yayımlanmamış elyazması.

Chislenko, Alexander. 1997. "Technology as Extension of Human Functional Architecture." Extropy Online.

Chorost, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human*. Boston: Houghton Mifflin.

Christiano, Paul F. 2012. "'Indirect Normativity' Write-up." Ordinary Ideas (blog), 21 Nisan.

CIA. 2013. *The World Factbook. Central Intelligence Agency*. Erişim Tarihi: 3 Ağustos. Bkz. https://www.cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html?countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us.

Cicero. 1923. "On Divination." *On Old Age, on Friendship, on Divination* içinde, İng. Çev. W. A. Falconer. Loeb Classical Library. Cambridge, MA: Harvard University Press.

Cirasella, Jill ve Kopec, Danny. 2006. "The History of Computer Games." *Exhibit at DartmouthArtificial Intelligence Conference: The Next Fifty Years* (AI@50), DartmouthCollege, 13–15 Temmuz.

Ćirković, Milan M. 2004. "Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings." *Foundations of Physics* 34 (2): 239–61.

Ćirković, Milan M., Sandberg Anders ve Bostrom, Nick. 2010. "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks." *Risk Analysis* 30 (10): 1495–1506.

Clark, Andy ve Chalmers, David J. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1. Baskı. Princeton, NJ: Princeton University Press.

Clavin, Whitney. 2012. "Study Shows Our Galaxy Has at Least 100 Billion Planets." *Jet Propulsion Laboratory*, 11 Ocak.

CME Group. 2010. *What Happened on May 6th?* Chicago, 10 Mayıs.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 4 (16): 386–405.

Cochran, Gregory ve Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books.

Cochran, G., Hardy, J. ve Harpending, H. 2006. "Natural History of Ashkenazi Intelligence." *Journal of Biosocial Science* 38 (5): 659–93.

Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.

Cope, David. 1996. *Experiments in Musical Intelligence*. Computer Music and Digital Audio Series. Madison, WI: A-R Editions.

Cotman, Carl W. ve Berchtold, Nicole C. 2002. "Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity." *Trends in Neurosciences* 25 (6): 295–301.

Cowan, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and BrainSciences* 24 (1): 87–114.

Crabtree, Steve. 1999. "New Poll Gauges Americans' General Knowledge Levels." *Gallup News*, July 6.

Cross, Stephen E. ve Walker, Edward. 1994. "Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning." *Intelligent Scheduling* içinde, haz. Monte Zweben ve Mark Fox, 711–29. San Francisco, CA: Morgan Kaufmann.

Crow, James F. 2000. "The Origins, Patterns and Implications of Human Spontaneous Mutation." *Nature Reviews Genetics* 1 (1): 40–7.

Cyranoski, David. 2013. "Stem Cells: Egg Engineers." *Nature* 500 (7463): 392–4.

Dagnelie, Gislin. 2012. "Retinal Implants: Emergence of a Multidisciplinary Field." *Current Opinion in Neurology* 25 (1): 67–75.

Dai, Wei. 2009. "Towards a New Decision Theory." Less Wrong (blog), 13 Ağustos.

Dalrymple, David. 2011. "Comment on Kaufman, J. 'Whole Brain Emulation: Looking at Progress on C. Elegans.'" Less Wrong (blog), 29 Ekim.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., v.d. 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16 (10): 996–1005.

Davis, Oliver S. P., Butcher, Lee M., Docherty, Sophia J., Meaburn, Emma L., Curtis, Charles J. C., Simpson, Michael A., Schalkwyk, Leonard C. ve Plomin, Robert. 2010. "A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects." *Behavior Genetics* 40 (6): 759–767.

Dawkins, Richard. 1995. *River Out of Eden: A Darwinian View of Life. Science MastersSeries.* New York: Basic Books.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems.* Berkeley, CA: Machine Intelligence Research Institute, 19 Mayıs.

De Long, J. Bradford. 1998. "Estimates of World GDP, One Million B.C.–Present." Yayımlanmamış elyazması.

De Raedt, Luc ve Flach, Peter, haz. 2001. Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Almanya, 5–7 Eylül 2001. Tutanaklar. Lecture Notes in Computer Science 2167. New York: Springer.

Dean, Cornelia. 2005. "Scientific Savvy? In U.S., Not Much." *New York Times*, 30 Ağustos.

Deary, Ian J. 2001. "Human Intelligence Differences: A Recent History." *Trends in CognitiveSciences* 5 (3): 127–30.

Deary, Ian J. 2012. "Intelligence." *Annual Review of Psychology* 63: 453–82.

Deary, Ian J., Penke, L. ve Johnson, W. 2010. "The Neuroscience of Human IntelligenceDifferences." *Nature Reviews Neuroscience* 11 (3): 201–11.

Degnan, G. G., Wind, T. C., Jones, E. V. ve Edlich, R. F. 2002. "Functional ElectricalStimulation in Tetraplegic Patients to Restore Hand Function." *Journal of Long-Term Effects of Medical Implants* 12 (3): 175–88.

Devlin, B., Daniels, M. ve Roeder, K. 1997. "The Heritability of IQ." *Nature* 388 (6641):468–71.

Dewey, Daniel. 2011. "Learning What to Value." *Artificial General Intelligence: 4th International Conference, AGI 2011* içinde, Mountain View, CA, USA, 3–6 Ağustos 2011. Tutanaklar, haz. Jürgen Schmidhuber, Kristinn R. Thórisson ve Moshe Looks, 309–14.Lecture Notes in Computer Science 6830. Berlin: Springer.

Dowe, D. L. ve Hernandez-Orallo, J. 2012. "IQ Tests Are Not for Machines, Yet." *Intelligence*40 (2): 77–81.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of Creation.* Garden City, NY: Anchor.

Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturingand Computation.* New York: Wiley.

Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization.* New York: Public Affairs.

Driscoll, Kevin. 2012. "Code Critique: 'Altair Music of a Sort.' " CriticalCode Studies Working Group Online Conference'taki tebliğ, 6 Şubat.

Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science*131 (3414): 1667–1668.

Dyson, Freeman J. 1979. *Disturbing the Universe.* 1. baskı Sloan Foundation Science Series. New York: Harper & Row.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2): 383–96.

Elga, Adam. 2007. "Reflection and Disagreement." *Noûs* 41 (3): 478–502.

Eliasmith, Chris, Stewart, Terrence C., Choo, Xuan, Bekolay, Trevor, DeWolf, Travis, Tang,Yichuan ve Rasmussen, Daniel. 2012. "A Large-Scale Model of the Functioning Brain."*Science* 338(6111): 1202–5.

Ellis, J. H. 1999. "The History of Non-Secret Encryption." *Cryptologia* 23 (3): 267–73.

Elyasaf, Achiya, Hauptmann, Ami ve Sipper, Moche. 2011. "Ga-Freecell: Evolving Solvers for the Game of Freecell." *Proceedings of the13th Annual Genetic and Evolutionary Computation Conference, 1931–1938*içinde. GECCO '11. New York: ACM.

Eppig, C., Fincher, C. L. ve Thornhill, R. 2010. "Parasite Prevalence and the WorldwideDistribution of Cognitive Ability." *Proceedings of the Royal Society B: Biological Sciences*277 (1701): 3801–8.

Espenshade, T. J., Guzman, J. C. ve Westoff, C. F. 2003. "The Surprising Global Variation in Replacement Fertility." *Population Research and Policy Review* 22 (5–6): 575–83.

Evans, Thomas G. 1964. "A Heuristic Program to Solve Geometric-Analogy Problems." *Proceedings of the April 21–23, 1964, Spring Joint Computer Conference* içinde, 327–338. AFIPS'64. New York: ACM.

Evans, Thomas G. 1968. "A Program for the Solution of a Class of Geometric-Analogy Intelligence-Test Questions." *SemanticInformation Processing* içinde,haz. Marvin Minsky, 271–353. Cambridge, MA: MIT Press.

Faisal, A. A., Selen, L. P. ve Wolpert, D. M. 2008. "Noise in the Nervous System." *Nature Reviews Neuroscience* 9 (4): 292–303.

Faisal, A. A., White, J. A. ve Laughlin, S. B. 2005. "Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15 (12): 1143–9.

Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning."*Nature* 407 (6804): 630–3.

Feldman, J. A. ve Ballard, Dana H. 1982. "Connectionist Models and Their Properties."*Cognitive Science* 6 (3): 205–254.

Foley, J. A., Monfreda, C., Ramankutty, N. ve Zaks, D. 2007. "Our Share of the PlanetaryPie." *Proceedings of the National Academy of Sciences of the United States of America* içinde 104(31): 12585–6.

Forgas, Joseph P., Cooper, Joel ve Crano, William D., haz. 2010. *The Psychology of Attitudesand Attitude Change. Sydney Symposium of Social Psychology.* New York: PsychologyPress.

Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess.* NewYork: Free Press.

Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth – Are TheyCompatible?: Part 6. The Distribution of Wealth.* Teknik rapor, OECD Economics Department Working Papers 929. OECD Publishing.

Freitas, Robert A., Jr. 1980. "A Self-Replicating Interstellar Probe." *Journal of the BritishInterplanetary Society* 33: 251–64.

Freitas, Robert A., Jr. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations."*Foresight Institute.* Nisan. Erişim Tarihi: 28 Temmuz 2013. Bkz. http://www.foresight.org/nano/Ecophagy.html.

Freitas, Robert A., Jr. ve Merkle, Ralph C. 2004. *Kinematic Self-Replicating Machines*.Georgetown, TX: Landes Bioscience.

Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy*. New York: Oxford University Press.

Gammoned.net. 2012. "Snowie." Arşiv versiyonu. Erişim Tarihi: 30 Haziran. Bkz. http://web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.html.

Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* 1 (2): 1.

Georgieff, Michael K. 2007. "Nutrition and the Developing Brain: Nutrient Priorities andMeasurement." *American Journal of Clinical Nutrition* 85 (2): 614S–620S.

Gianaroli, Luca. 2000. "Preimplantation Genetic Diagnosis: Polar Body and EmbryoBiopsy." *Supplement, Human Reproduction* 15 (4): 69–75.

Gilovich, Thomas, Griffin, Dale ve Kahneman, Daniel, haz. 2002. *Heuristics and Biases:The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Gilster, Paul. 2012. "ESO: Habitable Red Dwarf Planets Abundant." Centauri Dreams(blog), 29 Mart.

Goldstone, Jack A. 1980. "Theories of Revolution: The Third Generation." *WorldPolitics*32 (3): 425–53.

Goldstone, Jack A. 2001. "Towards a Fourth Generation of Revolutionary Theory." *Annual Review of Political Science* 4: 139–87.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers* içinde, haz. Franz L. Alt ve Morris Rubinoff, 6: 31–88. New York: Academic Press.

Good, Irving John. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1–4): 67–79.

Good, Irving John. 1976. "Book review of 'The Thinking Computer: Mind Inside Matter'" *International Journal of Man-Machine Studies*içinde 8: 617–20.

Good, Irving John. 1982. "Ethical Machines." *Intelligent Systems: Practice and Perspective* içinde, haz. J. E. Hayes, Donald Michie ve Y.-H. Pao, 555–60. Machine Intelligence10. Chichester: Ellis Horwood.

Goodman, Nelson. 1954. *Fact, Fiction ve Forecast*. 1. Baskı. Londra: Athlone Press.

Gott, J. R., Juric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N. ve Brinkmann, J. 2005. "A Map of the Universe." *Astrophysical Journal* 624 (2): 463–83.

Gottfredson, Linda S. 2002. "G: Highly General and Highly Practical." *The General Factor of Intelligence: How General Is It?içinde*, haz. Robert J. Sternberg and Elena L. Grigorenko,331–80. Mahwah, NJ: Lawrence Erlbaum.

Gould, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: Norton.

Graham, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. New York: Oxford University Press.

Gray, C. M. ve McCormick, D. A. 1996. "Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex."*Science* 274 (5284): 109–13.

Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review*, 21 Eylül.

Guizzo, Erico. 2010. "World Robot Population Reaches 8.6 Million." *IEEE Spectrum*,14 Nisan.

Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction. Science-Fiction Writers.* New York: Oxford University Press.

Haberl, Helmut, Erb, Karl-Heinz ve Krausmann, Fridolin. 2013. "Global Human Appropriation of Net Primary Production (HANPP)." *Encyclopedia of Earth*, 3 Eylül.

Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W. ve Fischer-Kowalski, M. 2007. "Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth's Terrestrial Ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12942–7.

Hajek, Alan. 2009. "Dutch Book Arguments." *Anand, Pattanaik ve Puppe* içinde 2009,173–95.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine.* Amherst, NY:Prometheus Books.

Hampson, R. E., Song, D., Chan, R. H., Sweatt, A. J., Riley, M. R., Gerhardt, G. A., Shin,D. C., Marmarelis, V. Z., Berger, T. W. ve Deadwyler, S. A. 2012. "A Nonlinear Modelfor Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal EnsembleStimulation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20(2): 184–97.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2).

Hanson, Robin. 1995. "Could Gambling Save Science? Encouraging an Honest Consensus."*Social Epistemology* 9 (1): 3–33.

Hanson, Robin. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for InterstellarColonization." Yayımlanmamış elyazması, 1 Temmuz. Erişim Tarihi: 26 Nisan 2012. http://hanson.gmu.edu/filluniv.pdf.

Hanson, Robin. 1998b. "Economic Growth Given Machine Intelligence." Yayımlanmamış elyazması. Erişim Tarihi: 15Mayıs 2013. Bkz. http://hanson.gmu.edu/aigrow.pdf.

Hanson, Robin. 1998c. "Long-Term Growth as a Sequence of Exponential Modes." Yayımlanmamış elyazması. Son ziyaret tarihi Aralık 2000. Bkz. http://hanson.gmu.edu/longgrow.pdf.

Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major EvolutionaryTransitions." Yayımlanmamış elyazması, 23 Eylül. Erişim Tarihi: 12 Ağustos 2012. Bkz. http://hanson.gmu.edu/hardstep.pdf.

Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Yayımlanmamış elyazması, Eylül. Son ziyarettarihi Ekim 2007. Bkz. http://hanson.gmu.edu/futarchy.pdf.

Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision*61 (4): 319–328.

Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45–50.

Hanson, Robin. 2009. "Tiptoe or Dash to Future?" Overcoming Bias (blog), 23 Aralık.

Hanson, Robin. 2012. "Envisioning the Economyand Society, of Whole Brain Emulations."AGI Impacts conference'ta tebliğ, Oxford,8–11 Aralık.

Hart, Oliver. 2008. "Economica Coase Lecture Reference Points and the Theory of the Firm." *Economica* 75 (299): 404–11.

Hay, Nicholas James. 2005. "Optimal Agents." tez, Auckland Üniversitesi.

Hedberg, Sara Reese. 2002. "Dart: Revolutionizing Logistics Planning." *IEEE Intelligent Systems* 17 (3): 81–3.

Helliwell, John, Layard, Richard ve Sachs, Jeffrey. 2012. *World Happiness Report. TheEarth Institute.*

Helmstaedter, M., Briggman, K. L. ve Denk, W. 2011. "High-Accuracy Neurite Reconstructionfor High-Throughput Neuroanatomy." *Nature Neuroscience* 14 (8): 1081–8.

Heyl, Jeremy S. 2005. "The Long-Term Future of Space Travel." *Physical Review D* 72 (10):1–4.

Hibbard, Bill. 2011. "Measuring Agent Intelligence via Hierarchies of Environments." *Artificial General Intelligence: 4th International Conference, AGI 2011* içinde, Mountain View,CA, USA, 3–6 Ağustos 2011. Tutanaklar, haz. Jürgen Schmidhuber, Kristinn R. Tharisson ve Moshe Looks, 303–8. Lecture Notes in Computer Science 6830. Berlin: Springer.

Hinke, R. M., Hu, X., Stillman, A. E., Herkle, H., Salmi, R. ve Ugurbil, K. 1993. "Functional Magnetic Resonance Imaging of Broca's Area During Internal Speech." *Neuroreport*4 (6): 675–8.

Hinxton Group. 2008. *Consensus Statement: Science, Ethics and Policy Challenges of PluripotentStem Cell-Derived Gametes.* Hinxton, Cambridgeshire, Birleşik Krallık, 11 Nisan. Bkz. http://www.hinxtongroup.org/Consensus_HG08_FINAL.pdf.

Hoffman, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race andIts Dangerous Legacy.* New York: Doubleday.

Hofstadter, Douglas R. (1979) 1999. *Gödel, Escher, Bach: An Eternal Golden Braid.* NewYork: Basic Books.

Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3–4).

Horton, Sue, Alderman, Harold ve Rivera, Juan A. 2008. *Copenhagen Consensus 2008*Challenge Paper: Hunger and Malnutrition. Teknik rapor. Copenhagen ConsensusCenter, 11 Mayıs.

Howson, Colin ve Urbach, Peter. 1993. *Scientific Reasoning: The Bayesian Approach.* 2. baskı Chicago: Open Court.

Hsu, Stephen. 2012. "Investigating the Genetic Basis for Intelligence and Other Quantitative Traits." UC Davis Department of Physics Colloquium'da verilen dersten, Davis, CA,13 Şubat.

Huebner, Bryce. 2008. "Do You See What We See? An Investigation of an Argument Against Collective Representation." *PhilosophicalPsychology* 21 (1): 91–112.

Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D. ve Jorde, L. B. 2010. "Mobile Elements Reveal Small Population Size in the Ancient Ancestors of Homo Sapiens." *Proceedingsof the National Academy of Sciences of the United States of America* 107 (5):2147–52.

Huffman, W. Cary ve Pless, Vera. 2003. *Fundamentals of Error-Correcting Codes.* New York: Cambridge University Press.

Hunt, Patrick. 2011. "Late Roman Silk: Smuggling and Espionage in the 6th Century CE." Philolog, Stanford University (blog), Ağustos 2.

Hutter, Marcus. 2001. "Towards a Universal Theory of Artificial Intelligence Based onAlgorithmic Probability and Sequential Decisions." De Raedt and Flach 2001 içinde, 226–38.

Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On AlgorithmicProbability. Texts in Theoretical Computer Science.* Berlin: Springer.

Iliadou, A. N., Janson, P. C. ve Cnattingius, S. 2011. "Epigenetics and Assisted ReproductiveTechnology." *Journal of Internal Medicine* 270 (5): 414–20.

IsakssonAnders. 2007. *Productivity and Aggregate Growth: A Global Picture.* Teknik rapor 05/2007. Viyana, Avusturya: UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch.

Jones, Garret. 2009. "Artificial Intelligence and Economic Growth: A Few Finger-Exercises."Yayımlanmamış elyazması, Ocak. Erişim Tarihi: 5 Kasım 2012. Bkz. http://mason.gmu.edu/~gjonesb/AIandGrowth.

Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb. United States Army in World War II.* Washington, DC: Center of Military History.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory. Cambridge Studies in Probability, Induction and Decision Theory.* New York: Cambridge University Press.

Judd, K. L., Schmedders, K. ve Yeltekin, S. 2012. "Optimal Rules for Patent Races." *International Economic Review* 53 (1): 23–52.

Kalfoglou, A., Suthers, K., Scott, J. ve Hudson, K. 2004. *Reproductive Genetic Testing: What America Thinks.* Genetics and Public Policy Center.

Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities and Permissible Harm.*Oxford Ethics Series. New York: Oxford University Press.

Kandel, Eric R., Schwartz, James H. ve Jessell, Thomas M., haz. 2000. *Principles of Neural Science.* 4. baskı. New York: McGraw-Hill.

Kansa, Eric. 2003. "Social Complexity and Flamboyant Display in Competition: MoreThoughts on the Fermi Paradox." Yayımlanmamış elyazması, arşiv versiyonu.

Karnofsky, Holden. 2012. "Comment on 'Reply to Holden on Tool AI.' " Less Wrong (blog),1 Ağustos.

Kasparov, Garry. 1996. "The Day That I Sensed a New Kind of Intelligence." *Time,* 25 Mart, no. 13.

Kaufman, Jeff. 2011. "Whole Brain Emulation and Nematodes." Jeff Kaufman's Blog (blog), 2 Kasım.

Keim, G. A., Shazeer, N. M., Littman, M. L., Agarwal, S., Cheves, C. M., Fitzgerald, J.,Grosland, J., Jiang, F., Pollard, S. ve Weinmeister, K. 1999. "Proverb: The Probabilistic Cruciverbalist." *Proceedings of the Sixteenth National Conference on Artificial Intelligence* içinde,710–17. Menlo Park, CA: AAAI Press.

Kell, Harrison J., Lubinski, David ve Benbow, Camilla P. 2013. "Who Rises to the Top?Early Indicators." *Psychological Science* 24 (5): 648–59.

Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature*42 (3): 752–82.

KGS Go Server. 2012. "KGS Game Archives: Games of KGS player zen19." Erişim Tarihi: Temmuz 2013. Bkz.http://www.gokgs.com/gameArchives.jsp?user=zen19d&oldAccounts=t&year=2012&month=3.

Knill, Emanuel, Laflamme, Raymond ve Viola, Lorenza. 2000. "Theory of Quantum Error Correction for General Noise." *Physical Review Letters* 84 (11): 2525–8.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V. ve Sterling,P. 2006. "How Much the Eye Tells theBrain." *Current Biology* 16 (14): 1428–34.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson,S. A., Sigurdsson, A., v.d. 2012. "Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488: 471–5.

Koomey, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010.* Teknik rapor, 08/01/2011. Oakland, CA: Analytics Press.

Koubi, Vally. 1999. "Military Technology Races." *International Organization* 53 (3): 537–65.

Koubi, Vally ve Lalman, David. 2007. "Distribution of Power and Military R&D." *Journal of Theoretical Politics* 19 (2): 133–52.

Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J. ve Lanza, G. 2003. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence.* 2. baskı Genetic Programming. Norwell, MA: Kluwer Academic.

Kremer, Michael. 1993. "Population Growth and Technological Change: One Million B.C.to 1990." *Quarterly Journal of Economics* 108 (3): 681–716.

Kruel, Alexander. 2011. "Interview Series on Risks from AI." Less Wrong Wiki (blog).Erişim Tarihi: 26 Ekim 2013. Bkz. http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI.

Kruel, Alexander. 2012. "Q&A with Experts on Risks From AI #2." Less Wrong (blog), 9 Ocak.

Krusienski, D. J. ve Shih, J. J. 2011. "Control of a Visual Keyboard Using an ElectrocorticographicBrain–Computer Interface." *Neurorehabilitation and Neural Repair* 25 (4):323–31.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions.* 1. baskı Chicago: Universityof Chicago Press.

Kuipers, Benjamin. 2012. "An Existing, Ecologically-Successful Genus of CollectivelyIntelligent Artificial Creatures." 4th International Conference, ICCCI 2012'deki tebliğ, Ho Chi Minh City, Vietnam, 28–30 Kasım.

Kurzweil, Ray. 2001. "Response to Stephen Hawking." Kurzweil Accelerating Intelligence.5 Eylül. Erişim Tarihi: 31 Aralık 2012. Bkz. http://www.kurzweilai.net/response-to-stephen-hawking.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York:Viking.

Laffont, Jean-Jacques ve Martimort, David. 2002. *The Theory of Incentives: The Principal-Agent Model.* Princeton, NJ: Princeton University Press.

Lancet, The. 2008. "Iodine Deficiency—Way to Go Yet." *The Lancet* 372 (9633): 88.

Landauer, Thomas K. 1986. "How Much Do People Remember? Some Estimates of theQuantity of Learned Information in Long-Term Memory." *Cognitive Science* 10 (4):477–93.

Lebedev, Anastasiya. 2004. "The Man Who Saved the World Finally Recognized." *Mos-News,* 21 Mayıs.

Lebedev, M. A. ve Nicolelis, M. A. 2006. "Brain–Machine Interfaces: Past, Present andFuture." *Trends in Neuroscience* 29 (9): 536–46.

Legg, Shane. 2008. "Machine Super Intelligence." Doktora tezi, Lugano Üniversitesi.

Leigh, E. G., Jr. 2010. "The Group Selection Controversy." *Journal of Evolutionary Biology*23(1): 6–19.

Lenat, Douglas B. 1982. "Learning Program Helps Win National Fleet Wargame Tournament."*SIGART Newsletter*79: 16–17.

Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and DomainConcepts." *Artificial Intelligence* 21 (1–2): 61–98.

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs*29 (4): 342–70.

Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND Journal ofEconomics* 28 (2): 228–47.

Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction.* Londra:Routledge.

Lewis, David. 1988. "Desire as Belief." *Mind: A Quarterly Review of Philosophy* 97 (387):323–32.

Li, Ming ve Vitányi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and ItsApplications. Texts in Computer Science.* New York: Springer.

Lin, Thomas, Mausam ve Etzioni, Oren. 2012. "Entity Linking at Web Scale." *Proceedingsof the Joint Workshop on Automatic Knowledge Base Construction and Web-scaleKnowledge Extraction (AKBC-WEKEX '12)*içinde, haz. James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek ve Partha Pratim Talukdar,84–88. Madison, WI: Omnipress.

Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047–54.

Louis Harris & Associates. 1969. "Science, Sexand Morality Survey, study no. 1927." *LifeMagazine* (New York) 4.

Lynch, Michael. 2010. "Rate, Molecular Spectrum ve Consequences of Human Mutation."*Proceedings of the National Academy of Sciences of the United States of America* 107(3): 961–8.

Lyons, Mark K. 2011. "Deep Brain Stimulation: Current and Future Clinical Applications."*Mayo Clinic Proceedings* 86 (7): 662–72.

MacAskill, William. 2010. "Moral Uncertainty and Intertheoretic Comparisons of Value." Tez, Oxford Üniversitesi.

McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18):1174–82.

McCorduck, Pamela. 1979. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence.* San Francisco: W. H. Freeman.

Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing*24 (2): 202–7.

MacKay, David J. C. 2003. *Information Theory, Inferenceand Learning Algorithms.* NewYork: Cambridge University Press.

McLean, George ve Stewart, Brian. 1979. "Norad False Alarm Causes Uproar." *The National.* 10 Kasım. Ottawa, ON: CBC, 2012. Haber bülteni.

Maddison, Angus. 1999. "Economic Progress: The Last Half Century in Historical Perspective."*Facts and Fancies of Human Development: Annual Symposium and CunninghamLecture*içinde, 1999, haz. Ian Castles. Occasional Paper Series, 1/2000. Acton,ACT: Academy of the Social Sciences in Australia.

Maddison, Angus. 2001. *The World Economy: A Millennial Perspective.* Development CentreStudies. Paris: Development Centre of the Organisation for Economic Co-operation and Development.

Maddison, Angus. 2005. *Growth and Interaction in the World Economy: The Roots ofModernity.* Washington, DC: AEI Press.

Maddison, Angus. 2007. *Contours of the World Economy, 1–2030 AD: Essays in Macro-Economic History.* New York: Oxford University Press.

Maddison, Angus. 2010. "Statistics of World Population, GDP and Per Capita GDP 1–2008 AD." Erişim Tarihi: 26 Ekim 2013. Bkz. http://www.ggdc.net/maddison/Historical_Statistics/vertical-file_02-2010.xls.

Mai, Q., Yu, Y., Li, T., Wang, L., Chen, M. J., Huang, S. Z., Zhou, C. ve Zhou, Q. 2007. "Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts." *Cell Research* 17 (12): 1008–19.

Mak, J. N. ve Wolpaw, J. R. 2009. "Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects." *IEEE Reviews in Biomedical Engineering* 2: 187–99.

Mankiw, N. Gregory. 2009. *Macroeconomics*. 7. baskı New York, NY: Worth.

Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333): 198–203.

Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times*, 16 Şubat.

Markram, Henry. 2006. "The Blue Brain Project." *Nature Reviews Neuroscience* 7 (2):153–160.

Mason, Heather. 2003. "Gallup Brain: The Birth of In Vitro Fertilization." *Gallup*, 5 Ağustos.

Menzel, Randolf ve Giurfa, Martin. 2001. "Cognitive Architecture of a Mini-Brain: The Honeybee." *Trends in Cognitive Sciences* 5 (2): 62–71.

Metzinger, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Mijic, Roko. 2010. "Bootstrapping Safe AGI Goal Systems." Roadmapsto AGI and the Future of AGI Workshop'taki tebliğ, Lugano, İsviçre, 8 Mart.

Mike, Mike. 2013. "Face of Tomorrow." Erişim Tarihi: 30 Haziran2012 .Bkz. http://faceoftomorrow.org.

Milgrom, Paul ve Roberts, John. 1990. "Bargaining Costs, Influence Costsand the Organization of Economic Activity." *Perspectives on Positive Political Economy* içinde, haz. James E. Alt ve Kenneth A. Shepsle, 57–89. New York: Cambridge UniversityPress.

Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits onOur Capacity for Processing Information." *Psychological Review* 63 (2): 81–97.

Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York: Doubleday.

Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer and More Dangerous World*. Dallas, TX: BenBella Books.

Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall.

Minsky, Marvin, ed. 1968. *Semantic Information Processing*. Cambridge, MA: MIT Press.

Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.' "Yayımlanmamış elyazması, 1 Ekim. Erişim Tarihi: 31 Aralık 2012. Bkz. http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html.

Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster.

Minsky, Marvin ve Papert, Seymour. 1969. *Perceptrons: An Introduction to ComputationalGeometry*. 1. baskı Cambridge, MA: MIT Press.

Moore Andrew. 2011. "Hedonism." *The Stanford Encyclopedia of Philosophy* içinde, Kış 2011, haz. Edward N. Zalta. Stanford, CA: Stanford University.

Moravec, Hans P. 1976. "The Role of Raw Power in Intelligence." Yayımlanmamış elyazması,12 Mayıs. Erişim Tarihi: 12 Ağustos 2012. Bkz. http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html.

Moravec, Hans P. 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover." Doktora tezi, Stanford Üniversitesi.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1.

Moravec, Hans P. 1999. "Rise of the Robots." *Scientific American*, Aralık, 124–35.

Muehlhauser, Luke ve Helm, Louie. 2012. "The Singularity and Machine Ethics." *Singularity Hypotheses: A Scientific and Philosophical Assessment* içinde, haz. Amnon Eden, Johnny Saraker, James H. Moor ve Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Muehlhauser, Luke ve Salamon, Anna. 2012. "Intelligence Explosion: Evidence and Import." *Singularity Hypotheses: A Scientific and Philosophical Assessment içinde*, haz. Amnon Eden, Johnny Saraker, James H. Moor ve Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Mu¨ller, Vincent C. ve Bostrom, Nick. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." *Fundamental Issues of Artificial Intelligence* içinde, haz. Vincent C. Mu¨ller. Synthese Library; Berlin: Springer.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press.

Nachman, Michael W. ve Crowell, Susan L. 2000. "Estimate of the Mutation Rate per Nucleotide in Humans." *Genetics* 156 (1): 297–304.

Nagy, Z. P. ve Chang, C. C. 2007. "Artificial Gametes." *Theriogenology* 67 (1): 99–104.

Nagy, Z. P., Kerkis, I. ve Chang, C. C. 2008. "Development of Artificial Gametes." *Reproductive BioMedicine Online* 16 (4): 539–44.

NASA. 2013. "International Space Station: Facts and Figures." Bkz. http://www.nasa.gov/worldbook/intspacestation_worldbook.html.

Newborn, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere*. New York: Springer.

Newell, Allen, Shaw, J. C. ve Simon, Herbert A. 1958. "Chess-Playing Programs and the Problem of Complexity." *IBM Journal of Research and Development* 2 (4): 320–35.

Newell, Allen, Shaw, J. C. ve Simon, Herbert A. 1959. "Report on a General Problem-Solving Program: Proceedings of the International Conference on Information Processing." *Information Processing* içinde, 256–64. Paris: UNESCO.

Nicolelis, Miguel A. L. ve Lebedev, Mikhail A. 2009. "Principles of Neural Ensemble Physiology Underlying the Operation of Brain–Machine Interfaces." *Nature Reviews Neuroscience* 10 (7): 530–40.

Nilsson, Nils J. 1984. *Shakey the Robot, Technical Note* 323. Menlo Park, CA: AI Center, SRI International, Nisan.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York: Cambridge University Press.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F. ve Turkheimer, E. 2012. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67 (2): 130–59.

Niven, Larry. 1973. "The Defenseless Dead." *Ten Tomorrows* içinde, haz. Roger Elwood, 91–142. New York: Fawcett.

Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *Journal of Economic History* 67 (1): 128–59.

Norton, John D. 2011. "Waiting for Landauer." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42 (3): 184–98.

Olds, James ve Milner, Peter. 1954. "Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain." *Journal of Comparative and Physiological Psychology* 47 (6): 419–27.

Olum, Ken D. 2002. "The Doomsday Argument and the Number of Possible Observers." *Philosophical Quarterly* 52 (207): 164–84.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Singularity Summit 2007'deki tebliğ, San Francisco, CA, 8–9 Eylül.

Omohundro, Stephen M. 2008. "The Basic AI Drives." *Artificial General Intelligence 2008: Proceedings of the First AGI Conference içinde*, haz. Pei Wang, Ben Goertzel ve Stan Franklin, 483–92. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." *Singularity Hypotheses: A Scientific and Philosophical Assessment* içinde, haz. Amnon Eden, Johnny Soraker, James H. Moor ve Eric Steinhart. The Frontiers Collection. Berlin: Springer.

O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9): 32–40.

Oshima, Hideki ve Katayama, Yoichi. 2010. "Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans." *Neurologia medico-chirurgica* 50 (9): 845–52.

Parfit, Derek. 1986. *Reasons and Persons*. New York: Oxford University Press.

Parfit, Derek. 2011. *On What Matters*. 2 cilt. The Berkeley Tanner Lectures. New York: Oxford University Press.

Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." *Airpower Journal* 11 (4).

Pasqualotto, Emanuele, Federici, Stefano ve Belardinelli, Marta Olivetti. 2012. "Toward Functioning and Usable Brain–Computer Interfaces (BCIs): A Literature Review." *Disability and Rehabilitation: Assistive Technology* 7 (2): 89–103.

Pearl, Judea. 2009. *Causality: Models, Reasoning ve Inference*. 2. baskı. New York: Cambridge University Press.

Perlmutter, J. S. ve Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29: 229–57.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2 ve Davis, O. S. 2013. "Common DNA Markers Can Account for More Than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2): 562–8.

Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times*, 1 Ağustos.

Pourret, Olivier, Naim, Patrick ve Marcot, Bruce, haz. 2008. *Bayesian Networks: A Practical Guide to Applications*. Chichester, West Sussex, Birleşik Krallık: Wiley.

Powell, A., Shennan, S. ve Thomas, M. G. 2009. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." *Science* 324 (5932): 1298–1301.

Price, Huw. 1991. "Agency and Probabilistic Causality." *British Journal for the Philosophy of Science* 42 (2): 157–76.

Qian, M., Wang, D., Watkins, W. E., Gebski, V., Yan, Y. Q., Li, M. ve Chen, Z. P. 2005. "The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China." *Asia Pacific Journal of Clinical Nutrition* 14 (1): 32–42.

Quine, Willard Van Orman ve Ullian, Joseph Silbert. 1978. *The Web of Belief*, haz. Richard Malin Ohmann, cilt 2. New York: Random House.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5–31.

Rajab, Moheeb Abu, Zarfoss, Jay, Monrose, Fabian ve Terzis Andreas. 2006. "A Multifaceted Approach to Understanding the Botnet Phenomenon." *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement* içinde, 41–52. New York: ACM.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.

Read, J. I. ve Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837): 2693–710.

Repantis, D., Schlattmann, P., Laisney, O. ve Heuser, I. 2010. "Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review." *Pharmacological Research* 62 (3): 187–206.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York: Simon & Schuster.

Rhodes, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. New York: Vintage.

Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tonu, Martin, Nicolas W., Westra, Harm-Jan, Shakhbazov, Konstantin, Abdellaoui, Abdel, v.d. 2013. "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340 (6139): 1467–71.

Ring, Mark ve Orseau, Laurent. 2011. "Delusion, Survival ve Intelligent Agents." *Artificial General Intelligence: 4th International Conference, AGI 2011*, Mountain View, CA, ABD, 3–6 Ağustos 2011. Tutanaklar, haz. Jürgen Schmidhuber, Kristinn R. Tharisson ve Moshe Looks, 11–20. Lecture Notes in Computer Science 6830. Berlin: Springer.

Ritchie, Graeme, Manurung, Ruli ve Waller, Annalu. 2007. "A Practical Application of Computational Humour." *Proceedings of the 4th International Joint Workshop on Computational Creativity*, haz. Amilcar Cardoso and Geraint A. Wiggins, 91–8. Londra: Goldsmiths, Londra Üniversitesi.

Roache, Rebecca. 2008. "Ethics, Speculation and Values." *NanoEthics* 2 (3): 317–27.

Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J. ve Gardner, E. 2008. "A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects." *Astrophysical Journal* 684 (1): 691–706.

Roe, Anne. 1953. *The Making of a Scientist*. New York: Dodd, Mead.

Roy, Deb. 2012. "About." Erişim Tarihi: 14Ekim. Bkz. http://web.media.mit.edu/~dkroy/.

Rubin, Jonathan ve Watson, Ian. 2011. "Computer Poker: A Review." *Artificial Intelligence* 175 (5–6): 958–87.

Rumelhart, D. E., Hinton, G. E. ve Williams, R. J. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–6.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." *The Philosophy of Logical Atomism and Other Essays 1914–1919* içinde, haz. John G. Slater, 8: 157–244. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin.

Russell, Bertrand ve Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914–1970.* New York: Routledge.

Russell, Stuart J. ve Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach.* 3. baskı Upper Saddle River, NJ: Prentice-Hall.

Sabrosky, Curtis W. 1952. "How Many Insects Are There?" *Insects* içinde, haz. United States Department of Agriculture, 1–7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

Salamon, Anna. 2009. "When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth." Taslak yazı, Aralık 27.

Salem, D. J. ve Rowan, A. N. 2001. *The State of the Animals: 2001.* Public Policy Series. Washington, DC: Humane Society Press.

Salverda, W., Nolan, B. ve Smeeding, T. M. 2009. *The Oxford Handbook of Economic Inequality.* Oxford: Oxford University Press.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–19.

Sandberg Anders. 1999. "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains." *Journal of Evolution and Technology* 5.

Sandberg Anders. 2010. "An Overview of Models of Technological Singularity." Roadmaps to AGI and the Future of AGI Workshop'taki tebliğ, Lugano, İsviçre, 8 Mart.

Sandberg Anders. 2013. "Feasibility of Whole Brain Emulation." *Philosophy and Theory of Artificial Intelligence* içinde, haz. Vincent C. Müller, 5: 251–64. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York: Springer.

Sandberg Anders ve Bostrom, Nick. 2006. "Converging Cognitive Enhancements." *Annals of the New York Academy of Sciences 1093*: 201–27.

Sandberg Anders ve Bostrom, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, Oxford Üniversitesi.

Sandberg Anders ve Bostrom, Nick. 2011. *Machine Intelligence Survey*. Technical Report 2011-1. Future of Humanity Institute, Oxford Üniversitesi.

Sandberg Anders ve Savulescu, Julian. 2011. "The Social and Economic Impacts of Cognitive Enhancement." *Enhancing Human Capacities* içinde, haz. Julian Savulescu, Ruud ter Meulen ve Guy Kahane, 92–112. Malden, MA: Wiley-Blackwell.

Schaeffer, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers.* New York: Springer.

Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P. ve Sutphen, S. 2007. "Checkers Is Solved." *Science* 317 (5844): 1518–22.

Schalk, Gerwin. 2008. "Brain–Computer Symbiosis." *Journal of Neural Engineering* 5 (1): P1–P15.

Schelling, Thomas C. 1980. *The Strategy of Conflict.* 2. baskı Cambridge, MA: Harvard University Press.

Schultz, T. R. 2000. "In Search of Ant Ancestors." *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14028–9.

Schultz, W., Dayan, P. ve Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593–9.

Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." *Encyclopedia of Artificial Intelligence* içinde, haz. Stuart C. Shapiro ve David Eckroth, 1: 488–503. New York: Wiley.

Schwitzgebel, Eric. 2013. "If Materialism is True, the United States is Probably Conscious." Taslak Yazı, 8 Şubat.

Sen, Amartya ve Williams, Bernard, haz. 1982. *Utilitarianism and Beyond.* New York: Cambridge University Press.

Shanahan, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds.* New York: Oxford University Press.

Shannon, Robert V. 2012. "Advances in Auditory Prostheses." *Current Opinion in Neurology* 25 (1): 61–6.

Shapiro, Stuart C. 1992. "Artificial Intelligence." *Encyclopedia of Artificial Intelligence*, 2. baskı, 1: 54–7. New York: Wiley.

Sheppard, Brian. 2002. "World-Championship-Caliber Scrabble." *Artificial Intelligence* 134 (1–2): 241–75.

Shoemaker, Sydney. 1969. "Time Without Change." *Journal of Philosophy* 66 (12): 363–81.

Shulman, Carl. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks.* Berkeley, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms.* Berkeley, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2012. "Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones?" AGI Impacts conference'taki tebliğ, Oxford, 8–11 Aralık.

Shulman, Carl ve Bostrom, Nick. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7–8): 103–30.

Shulman, Carl ve Bostrom, Nick. 2014. "Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer?" *Global Policy* 5 (1): 85–92.

Shulman, Carl, Jonsson, Henrik ve Tarleton, Nick. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan* içinde. Tutanaklar, haz. Carson Reynolds ve Alvaro Cassinelli, 23–25. AP-CAP 2009.

Sidgwick, Henry ve Jones, Emily Elizabeth Constance. 2010. *The Methods of Ethics.* Charleston, SC: Nabu Press.

Silver, Albert. 2006. "How Strong Is GNU Backgammon?" Backgammon Galore! 16 Eylül. Erişim Tarihi: Ekim 26, 2013. Bkz. http://www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu.

Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P. ve Hochberg, L. R. 2011. "Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array." *Journal of Neural Engineering* 8 (2): 025027.

Simester, Duncan ve Knez, Marc. 2002. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business* 75 (2): 283–304.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. New York: Harper & Row.

Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review* 118 (4): 465–500.

Slagle, James R. 1963. "A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus." *Journal of the ACM* 10 (4): 507–20.

Smeding, H. M., Speelman, J. D., Koning-Haanstra, M., Schuurman, P. R., Nijssen, P., van Laar, T. ve Schmand, B. 2006. "Neuropsychological Effects of Bilateral STN Stimulation in Parkinson Disease: A Controlled Study." *Neurology* 66 (12): 1830–6.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind: A Quarterly Review of Philosophy* 96 (381): 36–61.

Smith, Michael, Lewis, David ve Johnston, Mark. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* 63: 89–174.

Sparrow, Robert. 2013. "In Vitro Eugenics." *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101200. İnternet baskısı 4 Nisan 2013. Bkz. http://jme.bmj.com/content/early/2013/02/13/medethics-2012-101200.full.

Stansberry, Matt ve Kudritzki, Julian. 2012. *Uptime Institute 2012 Data Center Industry Survey*. Uptime Institute.

Stapledon, Olaf. 1937. *Star Maker*. London: Methuen.

Steriade, M., Timofeev, I., Durmuller, N. ve Grenier, F. 1998. "Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30–40 Hz) Spike Bursts." *Journal of Neurophysiology* 79 (1): 483–90.

Stewart, P. W., Lonky, E., Reihman, J., Pagano, J., Gump, B. B. ve Darvill, T. 2008. "The Relationship Between Prenatal PCB Exposure and Intelligence (IQ) in 9-Year-Old Children." *Environmental Health Perspectives* 116 (10): 1416–22.

Sun, W., Yu, H., Shen, Y., Banno, Y., Xiang, Z. ve Zhang, Z. 2012. "Phylogeny and Evolutionary History of the Silkworm." *Science China Life Sciences* 55 (6): 483–96.

Sundet, J., Barlaug, D. ve Torjussen, T. 2004. "The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century." *Intelligence* 32 (4): 349–62.

Sutton, Richard S. ve Barto verew G. 1998. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Talukdar, D., Sudhir, K. ve Ainslie, A. 2002. "Investigating New Product Diffusion Across Products and Countries." *Marketing Science* 21 (1): 97–114.

Teasdale, Thomas W. ve Owen, David R. 2008. "Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect." *Intelligence* 36 (2): 121–6.

Tegmark, Max ve Bostrom, Nick. 2005. "Is a Doomsday Catastrophe Likely?" *Nature* 438: 754.

Teitelman, Warren. 1966. "Pilot: A Step Towards Man–Computer Symbiosis." Doktora tezi, MIT.

Temple, Robert K. G. 1986. *The Genius of China: 3000 Years of Science, Discovery and Invention*. 1. baskı New York: Simon & Schuster.

Tesauro, Gerald. 1995. "Temporal Difference Learning and TD-Gammon." *Communications of the ACM* 38 (3): 58–68.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, Philip E. ve Belkin, Aaron. 1996. "Counterfactual Thought Experiments in World Politics: Logical, Methodological ve Psychological Perspectives." *Counterfactual Thought Experiments in World Politics: Logical, Methodological ve Psychological Perspectives* içinde, haz. Philip E. Tetlock and Aaron Belkin, 1–38. Princeton, NJ: Princeton University Press.

Thompson, Adrian. 1997. "Artificial Evolution in the Physical World." *Evolutionary Robotics: From Intelligent Robots to Artificial Life* içinde, haz. Takashi Gomi, 101–25. ER'97. Carp, ON: Applied AI Systems.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., v.d. 2006. "Stanley: The Robot That Won the DARPA Grand Challenge." *Journal of Field Robotics* 23 (9): 661–92.

Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E. ve Svoboda, K. 2002. "Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex." *Nature* 420 (6917): 788–94.

Traub, Wesley A. 2012. "Terrestrial, Habitable-Zone Exoplanet Frequency from Kepler." *Astrophysical Journal* 745 (1): 1–10.

Truman, James W., Taylor, Barbara J. ve Awad, Timothy A. 1993. "Formation of the Adult Nervous System." The *Development of Drosophila Melanogaster* içinde, haz. Michael Bate ve Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory.

Tuomi, Ilkka. 2002. "The Lives and the Death of Moore's Law." *First Monday* 7 (11).

Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60.

Turkheimer, Eric, Haley Andreana, Waldron, Mary, D'Onofrio, Brian ve Gottesman, Irving I. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14 (6): 623–8.

Uauy, Ricardo ve Dangour, Alan D. 2006. "Nutrition in Brain Development and Aging: Role of Essential Fatty Acids." Ek, *Nutrition Reviews* 64 (5): S24–S33.

Ulam, Stanislaw M. 1958. "John von Neumann." *Bulletin of the American Mathematical Society* 64 (3): 1–49.

Uncertain Future, The. 2012. "Frequently Asked Questions." *The Uncertain Future*. Erişim Tarihi: 25 Mart 2012. Bkz. http://www.theuncertainfuture.com/faq.html.

ABD Kongresi, Teknoloji Değerlendirme Ofisi. 1995. *U.S.–Russian Cooperation in Space ISS-618*. Washington, DC: U.S. Government Printing Office, Nisan.

Van Zanden, Jan Luiten. 2003. *On Global Economic History: A Personal View on an Agenda for Future Research. Amsterdam: International Institute of Social History*, 23 Temmuz.

Vardi, Moshe Y. 2012. "Artificial Intelligence: Past and Future." *Communications of the ACM* 55 (1): 5.

Vassar, Michael ve Freitas, Robert A., Jr. 2006. "Lifeboat Foundation Nanoshield." *Lifeboat Foundation*. Erişim Tarihi: 12 Mayıs 2012. Bkz. http://lifeboat.com/ex/nanoshield.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace* içinde, 11–22. NASA Conference Publication 10129. NASA Lewis Research Center.

Visscher, P. M., Hill, W. G. ve Wray, N. R. 2008. "Heritability in the Genomics Era: Concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–66.

Vollenweider, Franz, Gamma, Alex, Liechti, Matthias ve Huber, Theo. 1998. "Psychological and Cardiovascular Effects and Short-Term Sequelae of MDMA ('Ecstasy') in MDMA-Na. ve Healthy Volunteers." *Neuropsychopharmachology* 19 (4): 241–51.

Wade, Michael J. 1976. "Group Selections Among Laboratory Populations of Tribolium." *Proceedings of the National Academy of Sciences of the United States of America* 73 (12): 4604–7.

Wainwright, Martin J. ve Jordan, Michael I. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1 (1–2): 1–305.

Walker, Mark. 2002. "Prolegomena to Any Future Philosophy." *Journal of Evolution and Technology* 10 (1).

Walsh, Nick Paton. 2001. "Alter our DNA or robots will take over, warns Hawking." *The Observer*, 1 Eylül. http://www.theguardian.com/uk/2001/sep/02/medicalscience.genetics.

Warwick, Kevin. 2002. *I, Cyborg*. Londra: Century.

Wehner, M., Oliker, L. ve Shalf, J. 2008. "Towards Ultra-High Resolution Models of Climate and Weather." *International Journal of High Performance Computing Applications* 22 (2): 149–65.

Weizenbaum, Joseph. 1966. "Eliza: A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1): 36–45.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, CA: W. H. Freeman.

Werbos, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. New York: Wiley.

White, J. G., Southgate, E., Thomson, J. N. ve Brenner, S. 1986. "The Structure of the Nervous System of the Nematode Caenorhabditis Elegans." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314 (1165): 1–340.

Whitehead, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean*. Chicago: University of Chicago Press.

Whitman, William B., Coleman, David C. ve Wiebe, William J. 1998. "Prokaryotes: The Unseen Majority." *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578–83.

Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131 (3410): 1355–8.

Wikipedia. 2012a, s.v. "Computer Bridge." Erişim Tarihi: 30 Haziran 2013. Bkz. http://en.wikipedia.org/wiki/Computer_bridge.

Wikipedia. 2012b, s.v. "Supercomputer." Erişim Tarihi: 30 Haziran 2013. Bkz. http://et.wikipedia.org/wiki/Superarvuti.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton Science Library. Princeton, NJ: Princeton University Press.

Winograd, Terry. 1972. *Understanding Natural Language*. New York: Academic Press.

Wood, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation*. Londra: A&C Black.

World Bank. 2008. *Global Economic Prospects: Technology Diffusion in the Developing World*, 42097. Washington, DC.

World Robotics. 2011. *Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots*. Erişim Tarihi: 30 Haziran 2012. Bkz. http://www.bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf.

World Values Survey. 2008. WVS 2005-2008. Erişim Tarihi: 29 Ekim 2013. Bkz. http://www.wvsevsdb.com/wvs/WVSAnalizeStudy.jsp.

Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny*. New York: Vintage.

Yaeger, Larry. 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision ve Behavior or PolyWorld: Life in a New Context." *Proceedings of the Artificial Life III Conference* içinde, haz. C. G. Langton, 263–98. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Berkeley, CA: Machine Intelligence Research Institute, 15 Haziran.

Yudkowsky, Eliezer. 2002. "The AI-Box Experiment." Erişim Tarihi: 15 Ocak 2012. Bkz. http://yudkowsky.net/singularity/aibox.

Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition*. Berkeley, CA: Machine Intelligence Research Institute, Mayıs.

Yudkowsky, Eliezer. 2007. "Levels of Organization in General Intelligence." *Artificial General Intelligence* içinde, haz. Ben Goertzel ve Cassio Pennachin, 389–501. Cognitive Technologies. Berlin: Springer.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." *Global Catastrophic Risks* içinde, haz. Nick Bostrom ve Milan M. Ćirković, 308–45. New York: Oxford University Press.

Yudkowsky, Eliezer. 2008b. "Sustained Strong Recursion." Less Wrong (blog), 5 Aralık.

Yudkowsky, Eliezer. 2010. *Timeless Decision Theory*. Berkeley, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2011. *Complex Value Systems are Required to Realize Valuable Futures*. Berkeley, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2013. *Intelligence Explosion Microeconomics*, Technical Report 2013–1. Berkeley, CA: Machine Intelligence Research Institute.

Zahavi, Amotz ve Zahavi, Avishag. 1997. The *Handicap Principle: A Missing Piece of Darwin's Puzzle*. İng. Çev. N. Zahavi-Ely ve M. P. Ely. New York: Oxford University Press.

Zalasiewicz, J., Williams, M., Smith, A., Barry, T. L., Coe, A. L., Bown, P. R., Brenchley, P., v.d. 2008. "Are We Now Living in the Anthropocene?" *GSA Today* 18 (2): 4–8.

Zeira, Joseph. 2011. "Innovations, Patent Races and Endogenous Growth." *Journal of Economic Growth* 16 (2): 135–56.

Zuleta, Hernando. 2008. "An Empirical Note on Factor Shares." *Journal of International Trade and Economic Development* 17 (3): 379–90.

Dizin

A

adım riskleri 275
ahlaki büyüme 252
ahlaki doğruluk 256, 257, 258, 259, 353
ahlaki statü 199, 345
alet yapay zekâ 181
altyapı bolluğu 150, 152, 183, 265
antropik 261, 264, 338, 355
Arendt, Hannah 129
Armstrong, Stuart 15, 123, 289, 313, 316, 317, 332, 334, 340, 348, 349
asil-vekil problemi 155, 156
Asimov, Isaac 168, 339

B

Baruch planı 109
Bayesçi ağlar 26
Bayesçi aracı 27
belirleyici stratejik üstünlük 148, 188, 189, 210, 295
bellek ağları 305
bellek kapasitesi 91
Berliner, Hans 29, 315
Beşinci Nesil Bilgisayar 23
beyin-bilgisayar arayüzleri 104, 171
biçimleri 61, 63, 68, 80, 127, 135, 165, 183, 209, 217, 261, 272, 278, 303
bilge-tekel sürdürülebilirlik eşiği 126, 127
bilgiişlem gücü 79, 88, 92, 93, 184, 327, 342, 344
bilinç 207, 231, 345
bilişsel gelişim 276, 277, 280, 320
Birleşmiş Milletler 109, 110, 295, 296
biyolojik biliş 39, 63, 273
biyolojik geliştirme 171
Brown, Louise 62, 357
büyüme 17, 18, 19, 23, 46, 68, 69, 74, 94, 96, 97, 101, 113, 194, 252, 266, 273, 313, 323, 328, 336, 343
büyüme hızı 18, 19, 336

C

C. elegans 319
CEV 249, 250-257, 260-267, 351-354
Chalmers, David 15, 40, 289, 318, 334, 338, 350
CHINOOK 29
Christiano, Paul 15, 237, 238, 246, 306
Church-Turing tezi 264

D

DART 32
Deep Blue 30
Deep Fritz 39
değer öğrenici 237
değer öğrenme 226, 228, 230, 231, 232, 234, 237, 246
değer yükleme 219, 221, 222, 223, 224, 226, 227, 230, 232, 234, 237, 238, 245, 246
derin öğrenme 305
Dewey, Daniel 15, 228, 229, 230
dikkat temelli üretici modeller 305
diklik tezi 333
DNA sentezi 58, 121

doğal dil 31, 305
doğrudan saptama 167, 168, 170, 263
dolaylı normatiflik 167, 180, 249, 260, 314
donanım sarkması 282
Drexler, Eric 15, 124, 280, 324, 330, 332, 357
durum riskleri 275
dünya ekonomisi 18, 196
Dyson, Freeman 124, 325

E

Einstein, Albert 75, 89, 106
ekonomik büyüme 195, 196
elektronik kafalama 148, 150, 223, 228, 229, 245
ELIZA 22
embriyo seçilimi 58, 59, 62, 70, 86, 320, 321
emülasyon 51-54, 72, 84, 88, 91, 92, 125, 200-204, 210-213, 236, 238, 239, 243, 246, 278, 279, 284, 285, 286, 324, 344, 345, 346, 356, 357, 358
epistemoloji 116, 263, 264, 265, 354
Eurisko 29, 315
evcimenlik 167, 170, 176, 180, 221, 225
evrensel hızlandırıcı 274
evrim 25, 40, 41, 42, 44, 130, 187, 191, 207, 208, 224, 279, 319, 332

F

fayda işlevi 109, 145, 152, 161, 219, 220, 229, 230, 231, 232, 238, 246, 347
felaket 141, 144, 146, 155, 156, 184, 188, 207, 281, 283, 289, 329, 343, 359
Fields Madalyası 299
FreeCell 30

G

Genel Problem Çözücü 22
genetik algoritmalar 24, 25, 41, 279, 282
genetik seçilim 55, 62, 273, 280
Ginsberg, Matt 30
Go 30
Good, I.J. 20, 313, 314, 315, 318
Gorbaçov, Mihail 107, 108
gözetleme 68, 83, 105, 160, 296, 330, 349, 356
güçlü yapay zekâ 34
güdülenme seçilimi 157, 159, 171, 172, 176, 180, 185, 200, 212, 215, 221, 225, 240, 338, 349

H

Hanson, Robin 15, 18, 19, 192, 313, 323, 324, 328, 343, 345, 355, 358
hedef içeriği 262, 264
hedonizm 169
hedonyum 169
hızlandırıcı 274
hızlı süper zekâ 71, 72, 73, 77, 324
Hill, Benny 129
hiperparametre 305
Hodgkin-Huxley modeli 43
hükümdar yapay zekâ 266, 341

I-İ

IBM 30, 90
ikinci geçiş 210
insan düzeyinde 20, 31-42, 54, 65, 73, 81, 82, 87-94, 113, 144, 226, 227, 241, 244, 245, 284, 306
İnsan Genomu Projesi 107, 296, 330
insan zekâsı 114
internet 33, 34, 93, 106, 119, 121, 158, 166, 175, 225, 318, 325
işbirliği 29, 107, 108, 130, 161, 164, 205, 211, 213, 288, 291, 292, 294, 295, 296, 330, 339, 346, 359, 360, 361
işsizlik 84, 192, 342

K

kâhin yapay zekâ 265

karakter tanıma 32
karar teorisi 263, 267, 354
Kasparov, Garry 15, 30, 315
kaynak edinimi 137, 138, 227
Kepler, Johannes 31
Knuth, Donald 30, 316
kolektif süper zekâ 67, 71, 73, 74, 75, 77
kolektif zekâ 68, 70, 76, 171, 324, 325, 332
kombinasyonel patlama 25
komputronyum 124, 150, 151, 169, 258
kontrol yöntemleri 157, 165, 225, 244, 246, 338
konuşma tanıma 31, 32, 65, 305
Kopernik, Nicolaus 31
kötücül başarısızlık 146, 153
kurum tasarımı 240, 244
Kurzweil, Ray 19, 313, 322
kutulama 157, 158, 176, 179

L

Lenat, Douglas 29, 315
Logistello 30

M

mahremiyet 306
makine öğrenimi 25, 26, 32, 34, 222, 347
makine zekâsı 19, 20, 34, 42, 47, 144, 146, 192, 207, 211, 215, 244, 246, 269, 272, 273, 274, 279, 282, 283, 286, 288, 295, 300, 303, 305, 329, 346, 348, 359
Manhattan Projesi 104, 106, 108, 330
McCarthy, John 21, 26, 33, 34, 316, 331
McCulloch-Pitts nöronu 279
MegaEarth 75
Mill, John Stuart 248
Minsky, Marvin 34, 313, 314, 315, 316, 337
Moore yasası 41, 46, 96
Moravec, Hans 40, 41, 314, 318, 345

N

nanoteknoloji 116, 119, 121, 122, 127, 137, 214, 272, 332, 356
Newton, Isaac 75
Nilsson, Nils 34, 35, 314, 316, 317
nitelikli süper zekâ 71, 75, 76, 77
nootropik 319
Norvig, Peter 28, 34, 316, 335, 337
nöral ağlar 21, 24, 167, 305
nöromorfik yapay zekâ 52, 286, 287, 319

O-Ö

Oliphant, Mark 106
O'Neill, Gerard 124
optimallik fikri 28
optimizasyon 25, 26, 41, 81, 82, 84, 85, 89, 93, 94, 95, 96, 97, 114, 314, 333
optimizasyon gücü 81, 82, 84, 85, 93, 94, 95, 96, 114
Othello 30
otomasyon 121, 143
otomat teorisi 21
ödül sinyali 222, 229
öjeni 55, 61, 62, 321, 332

P

Parfit, Derek 15, 333, 334
poker 30

R

Rawls, John 180, 341, 350, 361
Reagan, Ronald 108
Riemann hipotezi 150, 170
Roosevelt, Franklin D. 106
Russell, Bertrand 15, 22, 28, 109, 168, 331, 335, 337, 339

S

Samuel, Arthur 29, 315
sanal gerçeklik 238, 242, 357

Sanayi Devrimi 17, 18, 84, 102, 193, 196
Sandberg, Anders 15, 43, 44, 53, 123, 124, 313, 317, 318, 319, 321, 322, 326, 328, 329, 332, 334
satranç 26, 29, 30, 37, 71, 115, 163, 315, 316
Schaeffer, Jonathan 29, 315
Schelling noktası 217, 340
Scrabble 30
sermaye 87, 105, 193, 194, 195, 197, 199, 202, 316, 344
SHRDLU 22
Shulman, Carl 15, 42, 163, 211, 212, 289, 318, 320, 321, 334, 339, 343, 346, 347, 349, 350, 356, 360, 362
siber suçlar 306
siborg 66
Stratejik Savunma İnisiyatifi 108
süper güçler 103, 113, 128
süper organizmalar 212
süper zekâ 13, 14, 20, 39, 58, 67-77, 82, 113-117, 171, 175, 176, 178, 180, 181, 210, 225, 226, 233, 235, 236, 239, 240, 247, 248, 249, 251, 253, 256, 257, 261, 262, 265, 266, 267, 271, 272, 275, 289, 291, 295, 296, 307, 324, 329, 332, 337, 338, 339, 351, 354, 357
sürücüsüz arabalar 306

T

takviyeli öğrenme 45, 222, 223, 228, 239, 245, 279, 305, 332, 337, 347, 348
tanım 35, 39, 74, 229, 233
tanımı 22, 72, 232, 234, 295, 317
tarama 48, 50, 51, 54, 57, 87, 88, 94, 287, 319
Tarım Devrimi 18, 19, 84, 102
tavla 29, 222
tekel 99, 100, 108, 109, 110, 111, 113, 126, 127, 128, 137, 141, 145, 191, 209, 210, 215, 217, 272, 283, 286, 295, 329, 330, 332, 333, 335, 343, 358, 360
tekillik 17, 19, 69, 313
teknoloji yarışı 243
Terminatör 305, 306
tersten örnekleme 149, 150, 183, 233
Tesauro, Gerry 29, 315
teşvik yöntemleri 157
tetikleyici 37, 157, 166, 167
TextRunner 90
Thrun, Sebastian 34, 314
tohum yapay zekâ 95, 114, 118, 225, 226, 234, 253, 282, 318
toplumsal bütünleşme 160, 185, 189
Tribolium castaneum 187
Truman, Harry 43, 107
Turing, Alan 20, 27, 40, 47, 63, 264, 305, 318, 325
tüm beyin 39, 45, 48, 50, 51, 52, 54, 65, 69, 70, 72, 79, 80, 87, 88, 94, 105, 106, 125, 132, 204, 224, 237, 238, 239, 277, 278, 279, 281, 284, 285, 286, 287, 288, 295, 319, 323, 329, 346, 356, 357, 358
tüm beyin emülasyonu 39, 45, 48, 51, 54, 69, 70, 72, 87, 88, 105, 106, 204, 238, 278, 279, 281, 284, 285, 286, 287, 288, 295, 323, 329, 356, 357

U-Ü

uygunluk işlevi 25, 43
uzlaşmazlık 84, 85, 86, 87, 88, 89, 91, 93, 94, 96, 97
ücretler 159, 192, 193, 194
üssel büyüme 197

V

varoluşsal risk 126, 127, 183, 271, 272, 285, 302, 305, 359
veri madenciliği 32, 33, 358

Vinge, Vernor 19, 69, 313, 323, 324
von Neumann, John 63, 109, 122, 123, 137, 138, 313, 331, 336

W

Watson 30, 90, 316
Whitehead, Alfred N. 22, 326
Wigner, Eugene 106
Winston, Patrick 34, 316

Y

yapay aracı 109, 204
yapay zekâ 21-24, 39, 81, 87, 113, 181, 182, 250-267, 351-354, 358, 360, 361
yarış dinamiği 241, 289, 291, 295
yeterlilik kontrolü 155, 160, 349
yinelemeli kendini geliştirme 47, 171, 303
yolları 85, 87, 92, 120, 147, 148, 152, 156, 171, 187, 243, 267
Yudkowsky, Eliezer 15, 89, 96, 114, 120, 130, 187, 228, 234, 235, 249, 250, 251, 252, 253, 255, 314, 318, 324, 328, 333, 337, 338, 344, 345, 347, 350, 351, 354, 355
Yüce Meryem yaklaşımı 236, 237
yükleme 47, 48, 132, 147, 171, 199, 219, 221, 222, 223, 224, 225, 226, 227, 230, 232, 234, 237, 238, 245, 246
yüz tanıma 32, 305, 347

Z

zekâ patlaması 19, 20, 83, 118, 162, 164, 266, 277, 283, 285, 288, 291, 294, 300, 303, 328, 343, 359, 360
zihin suçu 152, 183, 251, 265, 353

SÜPER ZEKÂ

YAPAY ZEKÂ UYGULAMALARI, TEHLİKELER VE STRATEJİLER

NICK BOSTROM

BİLİM VE TEKNOLOJİ

DİĞER HAYVANLARIN GÜÇLÜ KASLARI VE KESKİN PENÇELERİ VAR, BİZİMSE DAHA AKILLI BEYİNLERİMİZ. GENEL ZEKÂ BAKIMINDAN SAHİP OLDUĞUMUZ MÜTEVAZI ÜSTÜNLÜK SAYESİNDE DİL, TEKNOLOJİ VE KARMAŞIK BİR TOPLUMSAL ÖRGÜTLENME GELİŞTİRMEYİ BAŞARDIK. BU ÜSTÜNLÜK, HER KUŞAĞIN ÖNCEKİLERİN BAŞARILARINA YENİLERİNİ EKLEMESİYLE ZAMAN İÇİNDE DAHA DA İLERLEDİ. BİR GÜN İNSAN BEYNİNİ AŞAN MAKİNE BEYİNLER YARATMAYI BAŞARIRSAK, O ZAMAN BU YENİ SÜPER ZEKÂ ÇOK GÜÇLÜ HALE GELEBİLİR. NASIL Kİ ŞU AN GORİLLERİN KADERİ BİZ İNSANLARA BAĞIMLIYSA, BİZİM TÜRÜMÜZÜN KADERİ DE MAKİNE SÜPER ZEKÂSININ EYLEMLERİNE BAĞLI OLABİLİR. ANCAK BİZİM DE BİR ÜSTÜNLÜĞÜMÜZ VAR: YAPAY ZEKÂYI İNSANLAR YARATACAK. İNSANİ DEĞERLERİ KORUYACAK BİR TÜR SÜPER ZEKÂ YARATABİLİRİZ. BUNU YAPMAK İÇİN KESİNLİKLE GÜÇLÜ BİR NEDENİMİZ OLACAKTIR.

NICK BOSTROM, SÜPER ZEKÂ OLASILIĞININ SUNDUĞU ZORLUĞU VE BUNA EN İYİ ŞEKİLDE NASIL YANIT VEREBİLECEĞİMİZİ ANLAMAYA ÇALIŞIYOR. İNSANLIĞIN GEÇMİŞTEN BU YANA KARŞILAŞTIĞI EN BÜYÜK ZORLUK BUDUR DEMEK YANLIŞ OLMAZ. BU İŞİ BAŞARALIM YA DA BAŞARAMAYALIM, MUHTEMELEN KARŞILAŞIP KARŞILAŞACAĞIMIZ SON ZORLUK OLACAK.

ZEKÂNIN DOĞASINI DETAYLICA İNCELEYEN BU KİTAP, DOĞAL VE YAPAY ZEKÂNIN İÇERİĞİNİ VE SINIRLARINI KAPSAMLI BİR ŞEKİLDE SUNUYOR. GELECEK BELKİ DOĞAL, BELKİ DE YAPAY ZEKÂNIN ELİNDE OLACAK AMA KESİN OLAN BİR ŞEY VARSA, ZEKÂ EVRENDEKİ VARLIĞINI SÜRDÜRECEK.

"BU KİTABI HARARETLE ÖNERİYORUM."
–BILL GATES

NICK BOSTROM, OXFORD ÜNİVERSİTESİ'NDE FELSEFE PROFESÖRÜ VE İNSANLIĞIN GELECEĞİ ENSTİTÜSÜ VE GELECEK TEKNOLOJİLERİ PROGRAMI BAŞKANI.

İNGİLİZCEDEN ÇEVİREN: FERİT BURAK AYDAR

TFG 0494411 RNR
2019 TEMMUZ TH 0494411 RNR

34 TL

ISBN 978-605-2116-78-4
9 786052 116784

KÜY