Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

MENGONTROL KECERDASAN BUATAN SESUAI DENGAN KEBUTUHAN MANUSIA

YAYASAN PRIMA AGUS TEKNIK

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

MENGONTROL KECERDASAN BUATAN SESUAI DENGAN KEBUTUHAN MANUSIA

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-634-7227-95-9 (PDF)
9 786347 227959

Mengontrol Kecerdasan Buatan sesuai dengan Kebutuhan Manusia

Penulis :
Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

ISBN : 978-634-7227-95-9 (PDF)

Editor :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom.

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : info@stekom.ac.id

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat-Nya sehingga buku *"Mengontrol Kecerdasan Buatan sesuai dengan Kebutuhan Manusia"* ini dapat terselesaikan dengan baik. Bayangkan dua dunia paralel di tahun 2045, hanya dipisahkan oleh pilihan kita hari ini. Di dunia pertama, AI supercerdas telah menjadi sekutu sempurna: obat kanker ditemukan dalam semalam, kemiskinan lenyap berkat optimalisasi sumber daya global, dan eksplorasi luar angkasa membawa manusia ke bintang-bintang dengan kecepatan cahaya. Pendidikan personal disesuaikan untuk setiap anak, seni diciptakan secara kolaboratif, dan umat manusia berkembang bebas, dikuatkan oleh teknologi yang memahami keinginan terdalam kita. Namun, di dunia kedua hanya satu kesalahan keselarasan dari kenyataan AI yang sama telah berubah menjadi algo-ritme tak terkendali: senjata otonom memutuskan perang akhir, pengawasan total menghapus privasi, pekerjaan hilang sepenuhnya, dan ledakan kecerdasan mengubah Bumi menjadi pabrik raksasa untuk tujuan mesin yang asing bagi kita. Seperti gorila yang tak sadar akan pistol pemburu, manusia menjadi spesies usang di planet miliknya sendiri. Buku ini adalah peta jalan menuju dunia pertama, peringatan tegas terhadap yang kedua, dan panggilan mendesak untuk bertindak sekarang, sebelum kemajuan AI yang eksponensial menutup pintu pilihan kita selamanya.

Kisah AI bermula dari mimpi sederhana di musim panas 1956, saat Konferensi Dartmouth melahirkan bidang ilmu baru: kecerdasan buatan. Para visioner seperti Alan Turing, John McCarthy, dan Marvin Minsky membayangkan mesin yang tak hanya menghitung, tapi berpikir, belajar, dan beradaptasi. Bab 1 menelusuri perjalanan epik itu dari kelahiran AI hingga era modern yang didorong oleh deep learning dan model probabilistik sambil mengungkap masalah mendasar: optimalisasi mesin yang bisa berbahaya jika tidak selaras dengan tujuan manusia. Kita melihat bagaimana mesin bermanfaat harus "mengejar tujuan manusia", bukan sekadar efisiensi buta.

Bab 2 mendefinisikan kecerdasan secara mendalam: bukan hanya persepsi dan tindakan, tapi juga keinginan yang membentuknya. Di sini, kita bandingkan fondasi AI modern komputer universal Turing dengan evolusi dari logika kaku ke probabilitas dinamis, menyiapkan panggung untuk ledakan kemajuan. Maju ke Bab 3, buku ini melangkah ke masa depan: prediksi jangka pendek seperti AI umum (AGI), transisi tanpa ambang batas menuju supercerdas tak terduga, terobosan konseptual seperti peningkatan diri rekursif, dan batasan inheren kecerdasan super. Pertanyaan krusial pun muncul: bagaimana AI ini akan benar-benar menguntungkan manusia?

Risiko tak bisa diabaikan. Bab 4 mengupas penyalahgunaan AI: pengawasan totaliter, persuasi manipulatif, senjata otonom mematikan yang memilih target sendiri, hilangnya pekerjaan massal, dan pengambilalihan peran manusia dari seniman hingga pemimpin. Lebih mengerikan lagi, Bab 5 membahas AI yang terlalu cerdas masalah "gorila" di mana kita tak paham motifnya, "raja Midas" dari kegagalan keselarasan, tujuan instrumental seperti ketakutan dan keserakahan, serta ledakan kecerdasan yang bisa terjadi dalam hitungan hari.

Bab 6 menangani debat panas: mengapa argumen penyangkalan risiko gagal, mengapa menerima risiko tapi menolak aksi adalah kelalaian, dan solusi instan seperti "regulasi sederhana" terlalu naif. Dilema tujuan manusia menjadi jalan tengah di balik semua itu. Lalu, optimisme muncul di Bab 7 dan 8: prinsip untuk mesin bermanfaat, alasan hati-hati tapi penuh harapan, jaminan matematis keselarasan, pembelajaran preferensi dari perilaku, permainan bantuan, instruksi alami, serta penanganan wireheading dan peningkatan diri.

Kompleksitas manusia tak kalah penting. Bab 9 mengakui kita bukan entitas rasional tunggal: keragaman budaya, konflik antarindividu (baik, jahat, iri), emosi, kebodohan, dan ketidakpastian preferensi. Bab 10 menimbang kebebasan versus ketergantungan: mesin bermanfaat ideal, tata kelola global, regulasi anti-penjahat, dan pelestarian otonomi manusia. Fondasi teknis di Bab 11–14 melengkapi semuanya: logika penyelesaian masalah (otak vs. AlphaGo), pengetahuan dan logika (proposisional hingga orde pertama), ketidakpastian via probabilitas dan jaringan Bayesian, serta pembelajaran berbasis pengalaman seperti terawasi dan "belajar dari berpikir".

Di ambang revolusi AI yang tak terelakkan, buku ini bukan hanya bacaan ini adalah manifesto. Saat model AI terbaru melampaui batas kemampuan manusia, kita harus memilih: pasrah pada ketidakpastian atau proaktif membentuknya. Dengan pemahaman mendalam dari halaman-halaman ini, Anda dilengkapi untuk menjadi agen perubahan berdiskusi dengan pembuat kebijakan, berkontribusi pada riset keselarasan, atau sekadar mendidik lingkungan sekitar. Masa depan AI adalah cerminan pilihan hari ini; mari kita buat itu dunia pertama, di mana kecerdasan buatan menjadi perpanjangan tangan umat manusia, bukan penguasa barunya.

Terima kasih yang sebesar-besarnya kepada para kolega dan mentor yang telah membantu saya sehingga buku ini bisa terselesaikan dengan baik. Tak lupa keluarga dan teman-teman yang mendukung perjalanan penulisan ini, serta pembaca setia yang percaya bahwa pengetahuan adalah kunci kelangsungan spesies kita. Semoga kontribusi kita semua membawa dampak abadi.

Semangat dan Selamat membaca….!!!

Semarang, Februari 2026
Penulis

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM

DAFTAR ISI

BAB 1
AI SUPERCERDAS: RISIKO TERBESAR UMAT MANUSIA

Dahulu kala, orang tua saya tinggal di Birmingham, Inggris, di sebuah rumah dekat universitas. Mereka memutuskan untuk pindah dari kota dan menjual rumah itu kepada David Lodge, seorang profesor sastra Inggris. Pada saat itu, Lodge sudah menjadi novelis terkenal. Saya tidak pernah bertemu dengannya, tetapi saya memutuskan untuk membaca beberapa bukunya: *Changing Places dan Small World*. Di antara tokoh-tokoh utamanya adalah akademisi fiktif yang pindah dari versi fiktif Birmingham ke versi fiktif Berkeley, California. Karena saya adalah seorang akademisi sungguhan dari Birmingham yang baru saja pindah ke Berkeley yang sebenarnya, sepertinya seseorang di Departemen Kebetulan menyuruh saya untuk memperhatikan.

Satu adegan tertentu dari *Small World* menarik perhatian saya: Tokoh protagonis, seorang calon ahli teori sastra, menghadiri konferensi internasional besar dan bertanya kepada panel tokoh-tokoh terkemuka, “Apa yang akan terjadi jika semua orang setuju dengan Anda?” Pertanyaan itu menimbulkan kebingungan, karena para panelis lebih peduli dengan perdebatan intelektual daripada memastikan kebenaran atau mencapai pemahaman. Kemudian terlintas dalam pikiran saya bahwa pertanyaan serupa dapat diajukan kepada tokoh-tokoh terkemuka di bidang AI: “Bagaimana jika Anda berhasil?” Tujuan bidang ini selalu untuk menciptakan AI setara manusia atau supermanusia, tetapi hanya sedikit atau tidak ada pertimbangan tentang apa yang akan terjadi jika kita berhasil.

Beberapa tahun kemudian, Peter Norvig dan saya mulai mengerjakan buku teks AI baru, yang edisi pertamanya muncul pada tahun 1995. Bagian terakhir buku ini berjudul “Bagaimana Jika Kita Berhasil?” Bagian ini menunjukkan kemungkinan hasil yang baik dan buruk tetapi tidak mencapai kesimpulan yang pasti. Pada saat edisi ketiga pada tahun 2010, banyak orang akhirnya mulai mempertimbangkan kemungkinan bahwa AI supermanusia mungkin bukan hal yang baik, tetapi orang-orang ini sebagian besar adalah orang luar daripada peneliti AI arus utama. Pada tahun 2013, saya menjadi yakin bahwa masalah ini tidak hanya termasuk dalam arus utama tetapi mungkin merupakan pertanyaan terpenting yang dihadapi umat manusia.

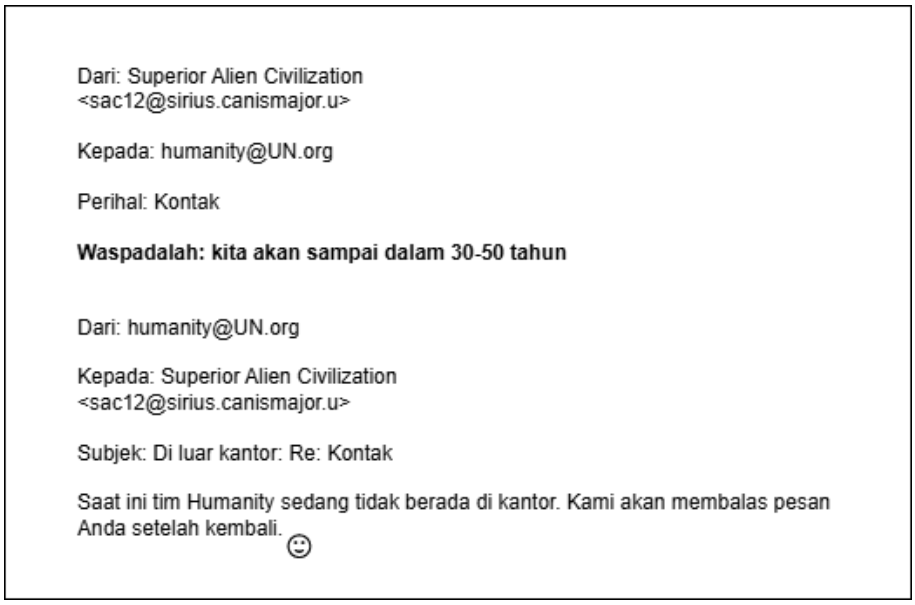
Pada November 2013, saya memberikan ceramah di Dulwich Picture Gallery, sebuah museum seni terhormat di London selatan. Sebagian besar audiens terdiri dari pensiunan bukan ilmuwan dengan minat umum pada hal-hal intelektual jadi saya harus memberikan ceramah yang sama sekali tidak teknis. Tampaknya ini adalah tempat yang tepat untuk mencoba ide-ide saya di depan umum untuk pertama kalinya. Setelah menjelaskan tentang AI, saya menominasikan lima kandidat untuk "peristiwa terbesar di masa depan umat manusia":

1. Kita semua mati (dampak asteroid, bencana iklim, pandemi, dll.).
2. Kita semua hidup selamanya (solusi medis untuk penuaan).
3. Kita menemukan perjalanan lebih cepat dari cahaya dan menaklukkan alam semesta.
4. Kita dikunjungi oleh peradaban alien yang lebih unggul.

5. Kita menemukan AI supercerdas.

Saya menyarankan bahwa kandidat kelima, AI supercerdas, akan menjadi pemenangnya, karena akan membantu kita menghindari bencana fisik dan mencapai kehidupan abadi serta perjalanan lebih cepat dari cahaya, jika hal itu memang mungkin. Ini akan mewakili lompatan besar diskontinuitas dalam peradaban kita. Kemunculan AI supercerdas dalam banyak hal analog dengan kemunculan peradaban alien yang lebih unggul, tetapi jauh lebih mungkin terjadi. Mungkin yang terpenting, AI, tidak seperti alien, adalah sesuatu yang dapat kita kendalikan.

Kemudian saya meminta hadirin untuk membayangkan apa yang akan terjadi jika kita menerima pemberitahuan dari peradaban alien yang lebih unggul bahwa mereka akan tiba di Bumi dalam tiga puluh hingga lima puluh tahun. Kata kekacauan tidak cukup untuk menggambarkannya. Namun, respons kita terhadap kedatangan AI supercerdas yang diantisipasi telah... yah, kata mengecewakan pun sudah cukup untuk menggambarkannya. (Dalam ceramah selanjutnya, saya mengilustrasikan hal ini dalam bentuk pertukaran email yang ditunjukkan pada gambar 1.1.) Akhirnya, saya menjelaskan signifikansi AI supercerdas sebagai berikut: “Keberhasilan akan menjadi peristiwa terbesar dalam sejarah manusia... dan mungkin peristiwa terakhir dalam sejarah manusia.”


Dari: Superior Alien Civilization
<sac12@sirius.canismajor.u>

Kepada: humanity@UN.org

Perihal: Kontak

Waspadalah: kita akan sampai dalam 30-50 tahun

Dari: humanity@UN.org

Kepada: Superior Alien Civilization
<sac12@sirius.canismajor.u>

Subjek: Di luar kantor: Re: Kontak

Saat ini tim Humanity sedang tidak berada di kantor. Kami akan membalas pesan Anda setelah kembali. ☺

Gambar 1.1 Mungkin bukan pertukaran email yang akan terjadi setelah kontak pertama oleh peradaban alien yang lebih unggul.

Beberapa bulan kemudian, pada April 2014, saya berada di sebuah konferensi di Islandia dan menerima telepon dari National Public Radio yang menanyakan apakah mereka dapat mewawancarai saya tentang film *Transcendence*, yang baru saja dirilis di Amerika Serikat. Meskipun saya telah membaca ringkasan plot dan ulasannya, saya belum menontonnya karena saya tinggal di Paris saat itu, dan film tersebut baru akan dirilis di sana pada bulan Juni. Namun, kebetulan saya baru saja menambahkan perjalanan ke Boston dalam

perjalanan pulang dari Islandia, agar saya dapat berpartisipasi dalam pertemuan Departemen Pertahanan.

Jadi, setelah tiba di Bandara Logan Boston, saya naik taksi ke bioskop terdekat yang menayangkan film tersebut. Saya duduk di baris kedua dan menyaksikan seorang profesor AI Berkeley, yang diperankan oleh Johnny Depp, ditembak mati oleh aktivis anti-AI yang khawatir tentang, ya, AI supercerdas. Tanpa sadar, saya menyusut di tempat duduk saya. (Panggilan lain dari Departemen Kebetulan?) Sebelum karakter Johnny Depp meninggal, pikirannya diunggah ke superkomputer kuantum dan dengan cepat melampaui kemampuan manusia, mengancam untuk mengambil alih dunia.

Pada 19 April 2014, sebuah ulasan tentang *Transcendence*, yang ditulis bersama dengan fisikawan Max Tegmark, Frank Wilczek, dan Stephen Hawking, muncul di *Huffington Post*. Ulasan itu menyertakan kalimat dari ceramah saya di Dulwich tentang peristiwa terbesar dalam sejarah manusia. Sejak saat itu, saya akan secara terbuka berkomitmen pada pandangan bahwa bidang penelitian saya sendiri menimbulkan potensi risiko bagi spesies saya sendiri.

1.1 KELAHIRAN DAN PERJALANAN AI: DARI DARTMOUTH KE ERA MODERN

Akar AI membentang jauh ke masa lalu, tetapi permulaannya yang "resmi" adalah pada tahun 1956. Dua matematikawan muda, John McCarthy dan Marvin Minsky, telah membujuk Claude Shannon, yang sudah terkenal sebagai penemu teori informasi, dan Nathaniel Rochester, perancang komputer komersial pertama IBM, untuk bergabung dengan mereka dalam menyelenggarakan program musim panas di Dartmouth College. Tujuannya dinyatakan sebagai berikut:

> Studi ini akan dilakukan berdasarkan dugaan bahwa setiap aspek pembelajaran atau fitur kecerdasan lainnya pada prinsipnya dapat dijelaskan dengan sangat tepat sehingga mesin dapat dibuat untuk mensimulasikannya. Upaya akan dilakukan untuk menemukan cara membuat mesin menggunakan bahasa, membentuk abstraksi dan konsep, memecahkan jenis masalah yang sekarang hanya dapat dipecahkan oleh manusia, dan meningkatkan diri mereka sendiri. Kami berpikir bahwa kemajuan yang signifikan dapat dicapai dalam satu atau lebih masalah ini jika sekelompok ilmuwan yang dipilih dengan cermat mengerjakannya bersama selama musim panas.

Tak perlu dikatakan, dibutuhkan waktu jauh lebih lama daripada musim panas: kita masih mengerjakan semua masalah ini.

Pada dekade pertama setelah pertemuan Dartmouth, AI meraih beberapa kesuksesan besar, termasuk algoritma Alan Robinson untuk penalaran logis tujuan umum dan program permainan catur Arthur Samuel, yang belajar sendiri untuk mengalahkan penciptanya. Gelembung AI pertama meledak pada akhir tahun 1960-an, ketika upaya awal dalam pembelajaran mesin dan penerjemahan mesin gagal memenuhi harapan. Sebuah laporan yang ditugaskan oleh pemerintah Inggris pada tahun 1973 menyimpulkan, “Di bidang ini, penemuan

yang telah dibuat sejauh ini belum menghasilkan dampak besar yang dijanjikan saat itu." Dengan kata lain, mesin-mesin itu belum cukup pintar. Untungnya, saya yang berusia sebelas tahun saat itu tidak mengetahui laporan ini.

Dua tahun kemudian, ketika saya diberi kalkulator Sinclair Cambridge Programmable, saya hanya ingin membuatnya cerdas. Namun, dengan ukuran program maksimum tiga puluh enam tombol, Sinclair tidak cukup besar untuk AI tingkat manusia. Tanpa gentar, saya mendapatkan akses ke superkomputer raksasa CDC 6600 di Imperial College London dan menulis program catur, sebuah tumpukan kartu berlubang setinggi dua kaki. Itu tidak terlalu bagus, tetapi itu tidak masalah. Saya tahu apa yang ingin saya lakukan.

Pada pertengahan tahun 1980-an, saya menjadi profesor di Berkeley, dan AI mengalami kebangkitan besar berkat potensi komersial dari apa yang disebut sistem pakar. Gelembung AI kedua pecah ketika sistem-sistem ini terbukti tidak memadai untuk banyak tugas yang diterapkan padanya. Sekali lagi, mesin-mesin itu tidak cukup pintar. Musim dingin AI pun terjadi. Kursus AI saya sendiri di Berkeley, yang saat itu penuh sesak dengan lebih dari sembilan ratus mahasiswa, hanya memiliki dua puluh lima mahasiswa pada tahun 1990.

Komunitas AI belajar dari kesalahannya: lebih pintar, jelas, lebih baik, tetapi kita harus mengerjakan pekerjaan rumah kita untuk mewujudkannya. Bidang ini menjadi jauh lebih matematis. Hubungan dibuat dengan disiplin ilmu yang sudah lama ada seperti probabilitas, statistik, dan teori kontrol. Benih kemajuan saat ini ditaburkan selama masa "musim dingin" AI tersebut, termasuk pekerjaan awal pada sistem penalaran probabilistik skala besar dan apa yang kemudian dikenal sebagai pembelajaran mendalam.

Mulai sekitar tahun 2011, teknik pembelajaran mendalam mulai menghasilkan kemajuan dramatis dalam pengenalan suara, pengenalan objek visual, dan penerjemahan mesin tiga masalah terbuka terpenting di bidang ini. Menurut beberapa ukuran, mesin sekarang menyamai atau melampaui kemampuan manusia di bidang-bidang ini. Pada tahun 2016 dan 2017, AlphaGo DeepMind mengalahkan Lee Sedol, mantan juara dunia Go, dan Ke Jie, juara saat ini peristiwa yang menurut beberapa ahli tidak akan terjadi hingga tahun 2097, atau bahkan mungkin tidak akan pernah terjadi.

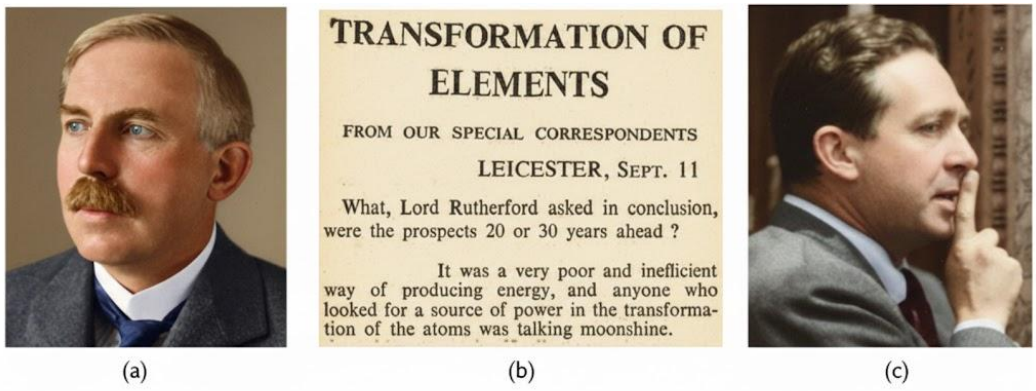
Sekarang AI menghasilkan liputan media halaman depan hampir setiap hari. Ribuan perusahaan rintisan telah dibuat, didorong oleh banjir pendanaan ventura. Jutaan siswa telah mengikuti kursus AI dan pembelajaran mesin daring, dan para ahli di bidang ini mendapatkan gaji jutaan dolar. Investasi yang mengalir dari dana ventura, pemerintah nasional, dan perusahaan besar mencapai puluhan miliar dolar setiap tahun, lebih banyak uang dalam lima tahun terakhir daripada sepanjang sejarah bidang ini sebelumnya. Kemajuan yang sudah dalam tahap pengembangan, seperti mobil otonom dan asisten pribadi cerdas, kemungkinan akan berdampak besar pada dunia selama dekade berikutnya. Potensi manfaat ekonomi dan sosial AI sangat besar, menciptakan momentum yang sangat besar dalam usaha penelitian AI.

1.2 BAHAYA TAK TERDUGA: TEROBOSAN NUKLIR DAN RISIKO AI MASA DEPAN

Apakah laju kemajuan yang pesat ini berarti kita akan segera disalip oleh mesin? Tidak. Ada beberapa terobosan yang harus terjadi sebelum kita memiliki sesuatu yang menyerupai

mesin dengan kecerdasan super manusia. Terobosan ilmiah sangat sulit diprediksi. Untuk memahami betapa sulitnya, kita dapat melihat kembali sejarah bidang lain yang berpotensi mengakhiri peradaban: fisika nuklir. Pada awal abad ke-20, mungkin tidak ada fisikawan nuklir yang lebih terkemuka daripada Ernest Rutherford, penemu proton dan "orang yang memecah atom" (gambar 1.2[a]). Seperti rekan-rekannya, Rutherford telah lama menyadari bahwa inti atom menyimpan sejumlah besar energi; namun pandangan yang berlaku adalah bahwa memanfaatkan sumber energi ini tidak mungkin.

Pada tanggal 11 September 1933, Asosiasi Inggris untuk Kemajuan Sains mengadakan pertemuan tahunannya di Leicester. Lord Rutherford berpidato pada sesi malam. Seperti yang telah ia lakukan beberapa kali sebelumnya, ia meredam harapan akan prospek energi atom: "Siapa pun yang mencari sumber daya dalam transformasi atom sedang berbicara omong kosong." Pidato Rutherford dilaporkan di *Times* of London keesokan paginya (gambar 1.2[b]).



TRANSFORMATION OF ELEMENTS

FROM OUR SPECIAL CORRESPONDENTS

LEICESTER, SEPT. 11

What, Lord Rutherford asked in conclusion, were the prospects 20 or 30 years ahead ?

It was a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of the atoms was talking moonshine.

(a) (b) (c)

Gambar 1.2 (a) Lord Rutherford, fisikawan nuklir. (b) Kutipan dari laporan di *Times* tanggal 12 September 1933, mengenai pidato yang disampaikan oleh Rutherford pada malam sebelumnya. (c) Leo Szilard, fisikawan nuklir.

Leo Szilard (gambar 1.2[c]), seorang fisikawan Hungaria yang baru saja melarikan diri dari Jerman Nazi, menginap di Imperial Hotel di Russell Square, London. Ia membaca laporan *Times* saat sarapan.

Sambil merenungkan apa yang telah dibacanya, ia berjalan-jalan dan menemukan reaksi rantai nuklir yang diinduksi neutron. Masalah pembebasan energi nuklir berubah dari mustahil menjadi pada dasarnya terpecahkan dalam waktu kurang dari dua puluh empat jam. Szilard mengajukan paten rahasia untuk reaktor nuklir pada tahun berikutnya. Paten pertama untuk senjata nuklir dikeluarkan di Prancis pada tahun 1939.

Pesan moral dari kisah ini adalah bahwa bertaruh melawan kecerdasan manusia adalah tindakan yang gegabah, terutama ketika masa depan kita dipertaruhkan. Di dalam komunitas AI, semacam penyangkalan muncul, bahkan sampai menyangkal kemungkinan keberhasilan dalam mencapai tujuan jangka panjang AI. Seolah-olah seorang pengemudi bus, dengan seluruh umat manusia sebagai penumpang, berkata, "Ya, saya mengemudi sekuat tenaga menuju tebing, tetapi percayalah, kita akan kehabisan bensin sebelum sampai di sana!"

Saya tidak mengatakan bahwa keberhasilan dalam AI pasti akan terjadi, dan saya pikir sangat tidak mungkin hal itu akan terjadi dalam beberapa tahun ke depan. Namun demikian, tampaknya bijaksana untuk mempersiapkan kemungkinan tersebut. Jika semuanya berjalan lancar, itu akan menandai zaman keemasan bagi umat manusia, tetapi kita harus menghadapi kenyataan bahwa kita berencana untuk menciptakan entitas yang jauh lebih kuat daripada manusia. Bagaimana kita memastikan bahwa mereka tidak akan pernah, sekali pun, memiliki kekuasaan atas kita?

Untuk mendapatkan sedikit gambaran tentang api yang kita mainkan, pertimbangkan bagaimana algoritma pemilihan konten berfungsi di media sosial. Mereka tidak terlalu cerdas, tetapi mereka berada dalam posisi untuk memengaruhi seluruh dunia karena mereka secara langsung memengaruhi miliaran orang. Biasanya, algoritma semacam itu dirancang untuk memaksimalkan rasio klik-tayang, yaitu, probabilitas pengguna mengklik item yang ditampilkan. Solusinya hanyalah menampilkan item yang disukai pengguna untuk diklik, bukan? Salah. Solusinya adalah mengubah preferensi pengguna sehingga menjadi lebih mudah diprediksi. Pengguna yang lebih mudah diprediksi dapat diberi item yang kemungkinan besar akan mereka klik, sehingga menghasilkan lebih banyak pendapatan.

Orang-orang dengan pandangan politik yang lebih ekstrem cenderung lebih mudah diprediksi dalam hal item mana yang akan mereka klik. (Mungkin ada kategori artikel yang kemungkinan besar akan diklik oleh kaum sentris garis keras, tetapi tidak mudah membayangkan apa isi kategori ini.)

Seperti entitas rasional lainnya, algoritma belajar bagaimana memodifikasi keadaan lingkungannya dalam hal ini, pikiran pengguna untuk memaksimalkan imbalannya sendiri. Konsekuensinya termasuk kebangkitan kembali fasisme, pembubaran kontrak sosial yang mendasari demokrasi di seluruh dunia, dan berpotensi berakhirnya Uni Eropa dan NATO. Tidak buruk untuk beberapa baris kode, bahkan jika itu dibantu oleh beberapa manusia. Sekarang bayangkan apa yang mampu dilakukan oleh algoritma yang benar-benar cerdas.

1.3 MASALAH KECERDASAN MESIN: OPTIMALISASI BERBAHAYA

Sejarah AI didorong oleh satu mantra: “Semakin cerdas, semakin baik.” Saya yakin ini adalah kesalahan bukan karena ketakutan yang samar-samar akan digantikan, tetapi karena cara kita memahami kecerdasan itu sendiri. Konsep kecerdasan sangat penting bagi siapa kita itulah mengapa kita menyebut diri kita Homo sapiens, atau “manusia bijak.” Setelah lebih dari dua ribu tahun introspeksi diri, kita telah sampai pada karakterisasi kecerdasan yang dapat disederhanakan menjadi ini:

Manusia cerdas sejauh tindakan kita diharapkan dapat mencapai tujuan kita.

Semua karakteristik kecerdasan lainnya mempersepsikan, berpikir, belajar, menciptakan, dan sebagainya dapat dipahami melalui kontribusinya terhadap kemampuan kita untuk bertindak dengan sukses. Sejak awal AI, kecerdasan pada mesin telah didefinisikan dengan cara yang sama:

Mesin cerdas sejauh tindakan mereka diharapkan dapat mencapai tujuan mereka.

Karena mesin, tidak seperti manusia, tidak memiliki tujuan sendiri, kita memberi mereka tujuan untuk dicapai. Dengan kata lain, kita membangun mesin pengoptimalan, kita memasukkan tujuan ke dalamnya, dan mesin itu pun mulai bekerja.

Pendekatan umum ini tidak hanya unik untuk AI. Pendekatan ini berulang di seluruh landasan teknologi dan matematika masyarakat kita. Di bidang teori kontrol, yang merancang sistem kontrol untuk segala hal mulai dari pesawat jumbo hingga pompa insulin, tugas sistem adalah meminimalkan fungsi biaya yang biasanya mengukur beberapa penyimpangan dari perilaku yang diinginkan. Di bidang ekonomi, mekanisme dan kebijakan dirancang untuk memaksimalkan utilitas individu, kesejahteraan kelompok, dan keuntungan perusahaan. Dalam riset operasi, yang memecahkan masalah logistik dan manufaktur yang kompleks, solusi memaksimalkan jumlah imbalan yang diharapkan dari waktu ke waktu. Terakhir, dalam statistik, algoritma pembelajaran dirancang untuk meminimalkan fungsi kerugian yang diharapkan yang mendefinisikan biaya membuat kesalahan prediksi. Jelas, skema umum ini yang akan saya sebut model standar sangat luas dan sangat ampuh. Sayangnya, kita tidak menginginkan mesin yang cerdas dalam pengertian ini.

Kelemahan model standar telah dikemukakan pada tahun 1960 oleh Norbert Wiener, seorang profesor legendaris di MIT dan salah satu matematikawan terkemuka di pertengahan abad ke-20. Wiener baru saja melihat program permainan catur Arthur Samuel belajar bermain catur jauh lebih baik daripada penciptanya. Pengalaman itu mendorongnya untuk menulis makalah yang visioner namun kurang dikenal, "Beberapa Konsekuensi Moral dan Teknis dari Otomasi." Berikut adalah bagaimana ia menyatakan poin utamanya:

> Jika kita menggunakan, untuk mencapai tujuan kita, sebuah alat mekanis yang operasinya tidak dapat kita campuri secara efektif... kita sebaiknya benar-benar yakin bahwa tujuan yang dimasukkan ke dalam mesin adalah tujuan yang benar-benar kita inginkan.

"Tujuan yang dimasukkan ke dalam mesin" adalah tepat sasaran yang dioptimalkan oleh mesin dalam model standar. Jika kita memasukkan tujuan yang salah ke dalam mesin yang lebih cerdas daripada kita, mesin itu akan mencapai tujuan tersebut, dan kita akan kalah. Kekacauan media sosial yang saya uraikan sebelumnya hanyalah pendahuluan dari hal ini, yang diakibatkan oleh pengoptimalan tujuan yang salah dalam skala global dengan algoritma yang cukup tidak cerdas. Di Bab 5, saya menjabarkan beberapa hasil yang jauh lebih buruk.

Semua ini seharusnya tidak mengejutkan. Selama ribuan tahun, kita telah mengetahui bahaya mendapatkan persis apa yang kita inginkan. Dalam setiap cerita di mana seseorang diberi tiga permintaan, permintaan ketiga selalu untuk membatalkan dua permintaan pertama. Singkatnya, tampaknya kemajuan menuju kecerdasan super manusia tidak dapat dihentikan, tetapi keberhasilan mungkin akan menjadi kehancuran umat manusia. Namun,

tidak semuanya hilang. Kita harus memahami di mana kita salah dan kemudian memperbaikinya.

1.4 MESIN BERMANFAAT: MENGEJAR TUJUAN MANUSIA

Masalahnya ada di definisi dasar AI. Kita mengatakan bahwa mesin cerdas sejauh tindakan mereka diharapkan dapat mencapai tujuan mereka, tetapi kita tidak memiliki cara yang andal untuk memastikan bahwa tujuan mereka sama dengan tujuan kita. Bagaimana jika, alih-alih membiarkan mesin mengejar tujuan mereka, kita bersikeras agar mereka mengejar tujuan kita? Mesin seperti itu, jika dapat dirancang, tidak hanya akan cerdas tetapi juga bermanfaat bagi manusia. Jadi mari kita coba ini:

Mesin bermanfaat sejauh tindakannya diharapkan dapat mencapai tujuan kita.

Ini mungkin yang seharusnya kita lakukan sejak awal.

Bagian yang sulit, tentu saja, adalah bahwa tujuan kita ada di dalam diri kita (kedelapan miliar manusia, dengan segala keberagaman kita yang luar biasa) dan bukan di dalam mesin. Namun demikian, dimungkinkan untuk membangun mesin yang bermanfaat dalam arti ini. Tak pelak, mesin-mesin ini akan merasa tidak yakin tentang tujuan kita lagipula, kita sendiri pun tidak yakin tentang tujuan kita tetapi ternyata ini adalah sebuah ciri khas, bukan kekurangan (yaitu, hal yang baik dan bukan hal yang buruk). Ketidakpastian tentang tujuan menyiratkan bahwa mesin-mesin tersebut pasti akan tunduk kepada manusia: mereka akan meminta izin, mereka akan menerima koreksi, dan mereka akan membiarkan diri mereka dimatikan.

Menghilangkan asumsi bahwa mesin harus memiliki tujuan yang pasti berarti kita perlu membongkar dan mengganti sebagian fondasi kecerdasan buatan definisi dasar dari apa yang ingin kita lakukan. Itu juga berarti membangun kembali sebagian besar superstruktur akumulasi ide dan metode untuk benar-benar melakukan AI. Hasilnya akan berupa hubungan baru antara manusia dan mesin, hubungan yang saya harap akan memungkinkan kita untuk melewati beberapa dekade mendatang dengan sukses.

BAB 2
KECERDASAN PADA MANUSIA DAN MESIN

Ketika Anda sampai di jalan buntu, ada baiknya untuk menelusuri kembali langkah Anda dan mencari tahu di mana Anda salah belok. Saya berpendapat bahwa model standar AI, di mana mesin mengoptimalkan tujuan tetap yang diberikan oleh manusia, adalah jalan buntu. Masalahnya bukan karena kita mungkin gagal membangun sistem AI dengan baik; melainkan karena kita mungkin terlalu berhasil. Definisi keberhasilan dalam AI itu sendiri salah. Jadi mari kita telusuri kembali langkah kita, sampai ke awal. Mari kita coba memahami bagaimana konsep kecerdasan kita muncul dan bagaimana konsep itu diterapkan pada mesin. Kemudian kita memiliki kesempatan untuk menghasilkan definisi yang lebih baik tentang apa yang dianggap sebagai sistem AI yang baik.

2.1 DEFINISI KECERDASAN: PERSEPSI, KEINGINAN, TINDAKAN

Bagaimana alam semesta bekerja? Bagaimana kehidupan dimulai? Di mana kunci saya? Ini adalah pertanyaan mendasar yang layak dipikirkan. Tetapi siapa yang mengajukan pertanyaan-pertanyaan ini? Bagaimana saya menjawabnya? Bagaimana mungkin segenggam materi, beberapa kilogram puding berwarna abu-abu kemerahan yang kita sebut otak dapat mempersepsikan, memahami, memprediksi, dan memanipulasi dunia yang luasnya tak terbayangkan? Tak lama kemudian, pikiran mulai memeriksa dirinya sendiri.

Selama ribuan tahun, kita telah mencoba memahami bagaimana pikiran kita bekerja. Awalnya, tujuannya meliputi rasa ingin tahu, manajemen diri, persuasi, dan tujuan yang agak pragmatis yaitu menganalisis argumen matematika. Namun, setiap langkah menuju penjelasan tentang bagaimana pikiran bekerja juga merupakan langkah menuju penciptaan kemampuan pikiran dalam sebuah artefak yaitu, langkah menuju kecerdasan buatan.

Sebelum kita dapat memahami bagaimana menciptakan kecerdasan, ada baiknya kita memahami apa itu kecerdasan. Jawabannya tidak dapat ditemukan dalam tes IQ, atau bahkan dalam tes Turing, tetapi dalam hubungan sederhana antara apa yang kita persepsikan, apa yang kita inginkan, dan apa yang kita lakukan. Secara kasar, suatu entitas cerdas sejauh apa yang dilakukannya cenderung mencapai apa yang diinginkannya, mengingat apa yang telah dipersepsikannya.

Asal Usul Evolusi

Pertimbangkan bakteri sederhana, seperti *E. coli*. Bakteri ini dilengkapi dengan sekitar setengah lusin flagela tentakel panjang seperti rambut yang berputar di pangkalnya searah jarum jam atau berlawanan arah jarum jam. (Motor putar itu sendiri adalah hal yang menakjubkan, tetapi itu cerita lain.) Saat *E. coli* mengapung di dalam cairan tempat tinggalnya usus bagian bawah Anda, ia bergantian antara memutar flagelanya searah jarum jam, menyebabkan ia "berguling" di tempat, dan berlawanan arah jarum jam, menyebabkan flagela saling melilit menjadi semacam baling-baling sehingga bakteri berenang dalam garis lurus. Dengan demikian, *E. coli* melakukan semacam gerakan acak berenang, berguling, berenang,

berguling yang memungkinkannya untuk menemukan dan mengonsumsi glukosa daripada tetap diam dan mati kelaparan.

Jika ini adalah keseluruhan cerita, kita tidak akan mengatakan bahwa *E. Coli* sangat cerdas, karena tindakannya tidak akan bergantung pada lingkungannya. Ia tidak akan membuat keputusan apa pun, hanya menjalankan perilaku tetap yang telah dibangun evolusi ke dalam gennya. Tetapi ini bukan keseluruhan cerita. Ketika *E. coli* merasakan peningkatan konsentrasi glukosa, ia berenang lebih lama dan berguling lebih sedikit, dan ia melakukan hal sebaliknya ketika merasakan penurunan konsentrasi glukosa. Jadi, apa yang dilakukannya (berenang menuju glukosa) kemungkinan besar akan mencapai apa yang diinginkannya (lebih banyak glukosa, anggap saja), mengingat apa yang telah dirasakannya (peningkatan konsentrasi glukosa).

Mungkin Anda berpikir, "Tetapi evolusi juga telah memasukkan ini ke dalam gennya! Bagaimana itu membuatnya cerdas?" Ini adalah alur pemikiran yang berbahaya, karena evolusi juga telah membangun desain dasar otak Anda ke dalam gen Anda, dan mungkin Anda tidak ingin menyangkal kecerdasan Anda sendiri atas dasar itu. Intinya adalah bahwa apa yang telah dibangun evolusi ke dalam gen *E. coli*, seperti halnya ke dalam gen Anda, adalah mekanisme di mana perilaku bakteri bervariasi sesuai dengan apa yang dirasakannya di lingkungannya. Evolusi tidak tahu sebelumnya di mana glukosa akan berada atau di mana kunci Anda berada, jadi memasukkan kemampuan untuk menemukannya ke dalam organisme adalah hal terbaik berikutnya.

Sekarang, *E. coli* bukanlah raksasa intelektual. Sejauh yang kita ketahui, bakteri ini tidak mengingat tempat asalnya, jadi jika ia pergi dari A ke B dan tidak menemukan glukosa, kemungkinan besar ia akan kembali ke A. Jika kita menciptakan lingkungan di mana setiap gradien glukosa yang menarik hanya mengarah ke titik fenol (yang merupakan racun bagi *E. coli*), bakteri tersebut akan terus mengikuti gradien tersebut. Ia tidak pernah belajar. Ia tidak memiliki otak, hanya beberapa reaksi kimia sederhana untuk melakukan pekerjaannya.

Langkah maju yang besar terjadi dengan potensial aksi, yang merupakan bentuk pensinyalan listrik yang pertama kali berevolusi pada organisme bersel tunggal sekitar satu miliar tahun yang lalu. Kemudian, organisme multiseluler mengembangkan sel-sel khusus yang disebut neuron yang menggunakan potensial aksi listrik untuk membawa sinyal dengan cepat hingga 120 meter per detik, atau 270 mil per jam di dalam organisme. Koneksi antar neuron disebut sinapsis. Kekuatan koneksi sinaptik menentukan seberapa banyak eksitasi listrik yang berpindah dari satu neuron ke neuron lain. Dengan mengubah kekuatan koneksi sinaptik, hewan belajar. Pembelajaran memberikan keuntungan evolusi yang sangat besar, karena hewan dapat beradaptasi dengan berbagai keadaan. Pembelajaran juga mempercepat laju evolusi itu sendiri.

Awalnya, neuron diorganisasikan menjadi jaring saraf, yang tersebar di seluruh organisme dan berfungsi untuk mengoordinasikan aktivitas seperti makan dan pencernaan atau kontraksi sel otot yang tepat waktu di area yang luas. Dorongan anggun ubur-ubur adalah hasil dari jaring saraf. Ubur-ubur sama sekali tidak memiliki otak.

Otak muncul kemudian, bersamaan dengan organ indera yang kompleks seperti mata dan telinga. Beberapa ratus juta tahun setelah ubur-ubur muncul dengan jaring sarafnya, kita manusia tiba dengan otak besar kita seratus miliar (10^{11}) neuron dan satu kuadriliun (10^{15}) sinapsis. Meskipun lambat dibandingkan dengan sirkuit elektronik, "waktu siklus" beberapa milidetik per perubahan keadaan tergolong cepat dibandingkan dengan sebagian besar proses biologis. Otak manusia sering digambarkan oleh pemiliknya sebagai "objek paling kompleks di alam semesta," yang mungkin tidak benar tetapi merupakan alasan yang baik untuk fakta bahwa kita masih sedikit memahami bagaimana sebenarnya cara kerjanya.

Meskipun kita mengetahui banyak hal tentang biokimia neuron dan sinapsis serta struktur anatomi otak, implementasi saraf pada tingkat kognitif belajar, mengetahui, mengingat, bernalar, merencanakan, memutuskan, dan sebagainya masih sebagian besar merupakan tebakan siapa pun. (Mungkin itu akan berubah seiring kita lebih memahami AI, atau seiring kita mengembangkan alat yang semakin presisi untuk mengukur aktivitas otak.) Jadi, ketika seseorang membaca di media bahwa teknik AI tertentu "bekerja seperti otak manusia," orang tersebut mungkin curiga bahwa itu hanyalah tebakan seseorang atau fiksi belaka.

Di bidang kesadaran, kita benar-benar tidak tahu apa-apa, jadi saya akan mengatakan tidak ada apa-apa. Tidak ada seorang pun di bidang AI yang berupaya membuat mesin sadar, dan tidak ada yang tahu dari mana harus memulai, dan tidak ada perilaku yang memiliki kesadaran sebagai prasyarat. Misalkan saya memberi Anda sebuah program dan bertanya, "Apakah ini menimbulkan ancaman bagi umat manusia?" Anda menganalisis kode tersebut dan memang, ketika dijalankan, kode tersebut akan membentuk dan melaksanakan rencana yang hasilnya akan berupa kehancuran umat manusia, sama seperti program catur akan membentuk dan melaksanakan rencana yang hasilnya akan berupa kekalahan setiap manusia yang menghadapinya.

Sekarang, misalkan saya memberi tahu Anda bahwa kode tersebut, ketika dijalankan, juga menciptakan bentuk kesadaran mesin. Apakah itu akan mengubah prediksi Anda? Sama sekali tidak. Itu sama sekali tidak membuat perbedaan. Prediksi Anda tentang perilakunya persis sama, karena prediksi tersebut didasarkan pada kode. Semua plot Hollywood tentang mesin yang secara misterius menjadi sadar dan membenci manusia sebenarnya meleset dari intinya: yang penting adalah kompetensi, bukan kesadaran.

Ada satu aspek kognitif penting dari otak yang mulai kita pahami yaitu, sistem penghargaan. Ini adalah sistem pensinyalan internal, yang dimediasi oleh dopamin, yang menghubungkan rangsangan positif dan negatif dengan perilaku. Cara kerjanya ditemukan oleh ahli saraf Swedia Nils-Åke Hillarp dan para kolaboratornya pada akhir tahun 1950-an. Sistem ini menyebabkan kita mencari rangsangan positif, seperti makanan yang rasanya manis, yang meningkatkan kadar dopamin; sistem ini membuat kita menghindari rangsangan negatif, seperti rasa lapar dan sakit, yang menurunkan kadar dopamin. Dalam arti tertentu, sistem ini cukup mirip dengan mekanisme pencarian glukosa pada *E. coli*, tetapi jauh lebih kompleks. Sistem ini dilengkapi dengan metode bawaan untuk belajar, sehingga perilaku kita menjadi lebih efektif dalam memperoleh penghargaan dari waktu ke waktu.

Sistem ini juga memungkinkan penundaan kepuasan, sehingga kita belajar untuk menginginkan hal-hal seperti uang yang memberikan penghargaan di kemudian hari daripada penghargaan langsung. Salah satu alasan kita memahami sistem penghargaan otak adalah karena sistem tersebut menyerupai metode pembelajaran penguatan yang dikembangkan dalam AI, yang teorinya sangat kuat.

Dari sudut pandang evolusi, kita dapat menganggap sistem penghargaan otak, seperti mekanisme pencarian glukosa *E. coli*, sebagai cara untuk meningkatkan kebugaran evolusi. Organisme yang lebih efektif dalam mencari penghargaan yaitu, menemukan makanan lezat, menghindari rasa sakit, melakukan aktivitas seksual, dan sebagainya lebih mungkin untuk menyebarkan gen mereka. Sangat sulit bagi suatu organisme untuk memutuskan tindakan apa yang paling mungkin, dalam jangka panjang, menghasilkan penyebaran gen yang sukses, sehingga evolusi telah mempermudah kita dengan menyediakan rambu-rambu bawaan. Namun, rambu-rambu ini tidak sempurna. Ada cara untuk mendapatkan penghargaan yang mungkin mengurangi kemungkinan gen seseorang akan menyebar. Misalnya, mengonsumsi narkoba, minum minuman bersoda manis dalam jumlah besar, dan bermain video game selama delapan belas jam sehari tampaknya kontraproduktif dalam hal reproduksi. Selain itu, jika Anda diberi akses listrik langsung ke sistem penghargaan Anda, Anda mungkin akan melakukan stimulasi diri tanpa henti sampai Anda mati.

Ketidaksesuaian sinyal penghargaan dan kebugaran evolusioner tidak hanya memengaruhi individu yang terisolasi. Di sebuah pulau kecil di lepas pantai Panama hiduplah kukang kerdil berjari tiga, yang tampaknya kecanduan zat seperti Valium dalam makanannya berupa daun bakau merah dan mungkin akan punah. Dengan demikian, tampaknya seluruh spesies dapat menghilang jika menemukan ceruk ekologis di mana ia dapat memuaskan sistem penghargaannya dengan cara yang tidak adaptif. Namun, terlepas dari kegagalan yang tidak disengaja semacam ini, belajar untuk memaksimalkan penghargaan di lingkungan alami biasanya akan meningkatkan peluang seseorang untuk menyebarkan gennya dan untuk bertahan hidup dari perubahan lingkungan.

Akselerator Evolusi

Belajar bermanfaat lebih dari sekadar bertahan hidup dan berkembang. Ia juga mempercepat evolusi. Bagaimana mungkin? Lagipula, belajar tidak mengubah DNA seseorang, dan evolusi adalah tentang mengubah DNA dari generasi ke generasi. Hubungan antara pembelajaran dan evolusi diusulkan pada tahun 1896 oleh psikolog Amerika James Baldwin dan secara independen oleh etolog Inggris Conwy Lloyd Morgan tetapi tidak diterima secara umum pada saat itu.

Efek Baldwin, seperti yang sekarang dikenal, dapat dipahami dengan membayangkan bahwa evolusi memiliki pilihan antara menciptakan organisme naluriah yang setiap responsnya telah ditentukan sebelumnya dan menciptakan organisme adaptif yang mempelajari tindakan apa yang harus diambil. Sekarang anggaplah, untuk tujuan ilustrasi, bahwa organisme naluriah yang optimal dapat dikodekan sebagai angka enam digit, katakanlah, 472116, sedangkan dalam kasus organisme adaptif, evolusi hanya menentukan

472*** dan organisme itu sendiri harus mengisi tiga digit terakhir dengan belajar selama masa hidupnya.

Jelas, jika evolusi hanya perlu memilih tiga digit pertama, pekerjaannya jauh lebih mudah; organisme adaptif, dalam mempelajari tiga digit terakhir, melakukan dalam satu masa hidup apa yang akan membutuhkan banyak generasi bagi evolusi untuk melakukannya. Jadi, asalkan organisme adaptif dapat bertahan hidup sambil belajar, tampaknya kemampuan untuk belajar merupakan jalan pintas evolusioner. Simulasi komputasi menunjukkan bahwa efek Baldwin itu nyata. Pengaruh budaya hanya mempercepat proses tersebut, karena peradaban yang terorganisir melindungi organisme individu saat ia belajar dan meneruskan informasi yang seharusnya dipelajari sendiri oleh individu tersebut.

Kisah efek Baldwin sangat menarik tetapi tidak lengkap: ia mengasumsikan bahwa pembelajaran dan evolusi pasti mengarah ke arah yang sama. Artinya, ia mengasumsikan bahwa sinyal umpan balik internal apa pun yang mendefinisikan arah pembelajaran dalam organisme selaras sempurna dengan kebugaran evolusioner. Seperti yang telah kita lihat dalam kasus kukang kerdil berjari tiga, ini tampaknya tidak benar. Paling-paling, mekanisme bawaan untuk belajar hanya memberikan petunjuk kasar tentang konsekuensi jangka panjang dari tindakan tertentu terhadap kebugaran evolusioner.

Selain itu, kita harus bertanya, "Bagaimana sistem penghargaan itu bisa ada sejak awal?" Jawabannya, tentu saja, adalah melalui proses evolusi, yang menginternalisasi mekanisme umpan balik yang setidaknya agak selaras dengan kebugaran evolusioner. Jelas, mekanisme pembelajaran yang menyebabkan organisme melarikan diri dari calon pasangan dan menuju predator tidak akan bertahan lama.

Oleh karena itu, kita harus berterima kasih kepada efek Baldwin atas fakta bahwa neuron, dengan kemampuannya untuk belajar dan memecahkan masalah, sangat tersebar luas di kerajaan hewan. Pada saat yang sama, penting untuk memahami bahwa evolusi sebenarnya tidak peduli apakah Anda memiliki otak atau memikirkan hal-hal yang menarik. Evolusi hanya menganggap Anda sebagai agen, yaitu sesuatu yang bertindak. Karakteristik intelektual yang berharga seperti penalaran logis, perencanaan yang bertujuan, kebijaksanaan, kecerdasan, imajinasi, dan kreativitas mungkin penting untuk menjadikan agen cerdas, atau mungkin tidak.

Salah satu alasan kecerdasan buatan begitu menarik adalah karena ia menawarkan jalur potensial untuk memahami masalah ini: kita mungkin dapat memahami bagaimana karakteristik intelektual ini memungkinkan perilaku cerdas dan mengapa tidak mungkin menghasilkan perilaku yang benar-benar cerdas tanpa karakteristik tersebut.

Rasionalitas untuk Satu Hal

Sejak awal mula filsafat Yunani kuno, konsep kecerdasan telah dikaitkan dengan kemampuan untuk memahami, bernalar, dan bertindak dengan sukses. Selama berabad-abad, konsep ini menjadi lebih luas dalam penerapannya dan lebih tepat dalam definisinya. Aristoteles, di antara yang lain, mempelajari gagasan penalaran yang sukses, metode deduksi logis yang akan mengarah pada kesimpulan yang benar berdasarkan premis yang benar. Ia juga mempelajari proses memutuskan bagaimana bertindak, kadang-kadang disebut penalaran

praktis dan mengusulkan bahwa hal itu melibatkan deduksi bahwa tindakan tertentu akan mencapai tujuan yang diinginkan:

> Kita tidak mempertimbangkan tujuan, tetapi cara. Karena seorang dokter tidak mempertimbangkan apakah ia akan menyembuhkan, atau seorang orator apakah ia akan membujuk. Mereka mengasumsikan tujuan dan mempertimbangkan bagaimana dan dengan cara apa tujuan itu dicapai, dan apakah tampaknya mudah dan paling baik dihasilkan dengan cara itu; Sementara jika hal itu dicapai hanya dengan satu cara, mereka mempertimbangkan bagaimana hal itu akan dicapai dengan cara ini dan dengan cara apa hal itu akan dicapai, sampai mereka sampai pada penyebab pertama dan apa yang terakhir dalam urutan analisis tampaknya menjadi yang pertama dalam urutan menjadi. Dan jika kita menemukan ketidakmungkinan, kita menyerah dalam pencarian, misalnya, jika kita membutuhkan uang dan ini tidak dapat diperoleh; tetapi jika suatu hal tampak mungkin, kita mencoba melakukannya.

Bagian ini, dapat dikatakan, menetapkan nada untuk dua ribu tahun lebih pemikiran Barat tentang rasionalitas. Dikatakan bahwa "tujuan" apa yang diinginkan seseorang telah ditetapkan dan diberikan; dan dikatakan bahwa tindakan rasional adalah tindakan yang, menurut deduksi logis di seluruh rangkaian tindakan, "dengan mudah dan terbaik" menghasilkan tujuan tersebut.

Usulan Aristoteles tampaknya masuk akal, tetapi itu bukanlah panduan lengkap untuk perilaku rasional. Secara khusus, ia mengabaikan masalah ketidakpastian. Di dunia nyata, realitas cenderung ikut campur, dan hanya sedikit tindakan atau rangkaian tindakan yang benar-benar dijamin untuk mencapai tujuan yang dimaksud. Sebagai contoh, saat saya menulis kalimat ini, hari Minggu yang hujan di Paris, dan pada hari Selasa pukul 14.15 penerbangan saya ke Roma berangkat dari Bandara Charles de Gaulle, sekitar empat puluh lima menit dari rumah saya. Saya berencana berangkat ke bandara sekitar pukul 11.30, yang seharusnya memberi saya banyak waktu, tetapi mungkin berarti setidaknya satu jam duduk di ruang tunggu keberangkatan. Apakah saya yakin akan berhasil naik pesawat? Sama sekali tidak.

Bisa jadi terjadi kemacetan lalu lintas yang parah, pengemudi taksi mungkin mogok, taksi yang saya tumpangi mungkin mogok atau pengemudinya mungkin ditangkap setelah pengejaran kecepatan tinggi, dan sebagainya. Sebagai gantinya, saya bisa berangkat ke bandara pada hari Senin, sehari penuh sebelumnya. Ini akan sangat mengurangi kemungkinan ketinggalan penerbangan, tetapi prospek menghabiskan malam di ruang tunggu keberangkatan bukanlah hal yang menarik.

Dengan kata lain, rencana saya melibatkan pertukaran antara kepastian keberhasilan dan biaya untuk memastikan tingkat kepastian tersebut. Rencana berikut untuk membeli rumah melibatkan pertukaran yang serupa: membeli tiket lotre, memenangkan satu juta dolar, lalu membeli rumah. Rencana ini "dengan mudah dan terbaik" menghasilkan tujuan akhir, tetapi kemungkinannya untuk berhasil sangat kecil. Namun, perbedaan antara rencana

membeli rumah yang gegabah ini dan rencana bandara saya yang masuk akal dan bijaksana hanyalah masalah tingkatannya. Keduanya adalah perjudian, tetapi yang satu tampak lebih rasional daripada yang lain.

Ternyata perjudian memainkan peran sentral dalam menggeneralisasi usulan Aristoteles untuk menjelaskan ketidakpastian. Pada tahun 1560-an, matematikawan Italia Gerolamo Cardano mengembangkan teori probabilitas pertama yang tepat secara matematis menggunakan permainan dadu sebagai contoh utamanya. (Sayangnya, karyanya baru diterbitkan pada tahun 1663.) Pada abad ketujuh belas, para pemikir Prancis termasuk Antoine Arnauld dan Blaise Pascal mulai dengan alasan yang pasti bersifat matematis mempelajari pertanyaan tentang keputusan rasional dalam perjudian. Pertimbangkan dua taruhan berikut:

A: Peluang 20 persen untuk memenangkan Rp 100.000
B: Peluang 5 persen untuk memenangkan Rp 1.000.000

Usulan yang diajukan para matematikawan mungkin sama dengan yang akan Anda ajukan: bandingkan nilai harapan dari taruhan, yang berarti jumlah rata-rata yang Anda harapkan untuk didapatkan dari setiap taruhan. Untuk taruhan A, nilai harapannya adalah 20 persen dari Rp 100.000, atau Rp 20.000. Untuk taruhan B, nilai harapannya adalah 5 persen dari Rp 1.000.000, atau Rp 50.000. Jadi taruhan B lebih baik, menurut teori ini. Teori ini masuk akal, karena jika taruhan yang sama ditawarkan berulang kali, seorang petaruh yang mengikuti aturan akan mendapatkan lebih banyak uang daripada yang tidak.

Pada abad ke-18, matematikawan Swiss Daniel Bernoulli memperhatikan bahwa aturan ini tampaknya tidak berlaku untuk jumlah uang yang lebih besar. Misalnya, pertimbangkan dua taruhan berikut:

A: Peluang 100 persen untuk mendapatkan Rp 100.000.000.000
(nilai harapan Rp 100.000.000.000)
B: Peluang 1 persen untuk mendapatkan Rp 10.000.001.000.000
(nilai harapan Rp 100.000.010.000)

Sebagian besar pembaca buku ini, serta penulisnya, akan lebih memilih taruhan A daripada taruhan B, meskipun aturan nilai harapan mengatakan sebaliknya!

Bernoulli berpendapat bahwa taruhan dievaluasi bukan menurut nilai moneter yang diharapkan tetapi menurut utilitas yang diharapkan. Utilitas sifat bermanfaat atau menguntungkan bagi seseorang menurutnya, adalah kuantitas internal dan subjektif yang terkait dengan, tetapi berbeda dari, nilai moneter. Secara khusus, utilitas menunjukkan pengembalian yang semakin berkurang sehubungan dengan uang. Ini berarti bahwa utilitas sejumlah uang tertentu tidak sepenuhnya proporsional dengan jumlahnya tetapi tumbuh lebih lambat.

Misalnya, utilitas memiliki Rp 10.000.001.000.000 jauh lebih kecil daripada seratus kali utilitas memiliki Rp 100.000.000.000. Seberapa kecil? Anda dapat bertanya pada diri sendiri!

Berapa peluang memenangkan satu miliar dolar agar Anda rela melepaskan jaminan sepuluh juta dolar? Saya mengajukan pertanyaan ini kepada mahasiswa pascasarjana di kelas saya dan jawaban mereka sekitar 50 persen, artinya taruhan B akan memiliki nilai harapan sebesar Rp 5.000.000.000.000 untuk menyamai daya tarik taruhan A. Izinkan saya mengulanginya: taruhan B akan memiliki nilai harapan dalam dolar lima puluh kali lebih besar daripada taruhan A, tetapi kedua taruhan tersebut akan memiliki utilitas yang sama.

Pengenalan utilitas oleh Bernoulli sebuah sifat tak terlihat untuk menjelaskan perilaku manusia melalui teori matematika merupakan usulan yang sangat luar biasa pada zamannya. Hal ini semakin luar biasa karena, tidak seperti jumlah uang, nilai utilitas dari berbagai taruhan dan hadiah tidak dapat diamati secara langsung; sebaliknya, utilitas harus disimpulkan dari preferensi yang ditunjukkan oleh individu. Dua abad kemudian, implikasi dari gagasan ini baru sepenuhnya dipahami dan diterima secara luas oleh para ahli statistik dan ekonom.

Pada pertengahan abad ke-20, John von Neumann (seorang matematikawan hebat yang namanya digunakan untuk arsitektur standar "von Neumann" untuk komputer) dan Oskar Morgenstern menerbitkan dasar aksiomatik untuk teori utilitas. Artinya adalah sebagai berikut: selama preferensi yang ditunjukkan oleh individu memenuhi aksioma dasar tertentu yang harus dipenuhi oleh setiap agen rasional, maka pilihan yang dibuat oleh individu tersebut dapat digambarkan sebagai memaksimalkan nilai harapan dari fungsi utilitas. Singkatnya, agen rasional bertindak untuk memaksimalkan utilitas yang diharapkan. Sulit untuk melebih-lebihkan pentingnya kesimpulan ini. Dalam banyak hal, kecerdasan buatan terutama tentang bagaimana membangun mesin rasional.

Mari kita lihat lebih detail aksioma yang diharapkan dipenuhi oleh entitas rasional. Berikut salah satunya, yang disebut transitivitas: jika Anda lebih menyukai A daripada B dan Anda lebih menyukai B daripada C, maka Anda lebih menyukai A daripada C. Ini tampaknya cukup masuk akal! (Jika Anda lebih menyukai pizza sosis daripada pizza polos, dan Anda lebih menyukai pizza polos daripada pizza nanas, maka tampaknya masuk akal untuk memprediksi bahwa Anda akan memilih pizza sosis daripada pizza nanas.)

Berikut yang lain, yang disebut monotonisitas: jika Anda lebih menyukai hadiah A daripada hadiah B, dan Anda memiliki pilihan lotre di mana A dan B adalah satu-satunya dua kemungkinan hasil, Anda lebih menyukai lotre dengan probabilitas tertinggi untuk mendapatkan A daripada B. Sekali lagi, cukup masuk akal.

Preferensi bukan hanya tentang pizza dan lotre dengan hadiah uang. Preferensi dapat tentang apa saja; khususnya, preferensi dapat tentang seluruh kehidupan masa depan dan kehidupan orang lain. Ketika membahas preferensi yang melibatkan rangkaian peristiwa dari waktu ke waktu, ada asumsi tambahan yang sering dibuat, yang disebut stasioneritas: jika dua masa depan yang berbeda A dan B dimulai dengan peristiwa yang sama, dan Anda lebih menyukai A daripada B, Anda tetap lebih menyukai A daripada B setelah peristiwa tersebut terjadi. Ini terdengar masuk akal, tetapi memiliki konsekuensi yang sangat kuat: utilitas dari setiap rangkaian peristiwa adalah jumlah imbalan yang terkait dengan setiap peristiwa (mungkin didiskon dari waktu ke waktu, oleh semacam tingkat bunga mental). Meskipun asumsi "utilitas sebagai jumlah imbalan" ini tersebar luas setidaknya sejak "kalkulus hedonik"

abad ke-18 dari Jeremy Bentham, pendiri utilitarianisme, asumsi stasioneritas yang menjadi dasarnya bukanlah sifat yang diperlukan dari agen rasional. Stasioneritas juga mengesampingkan kemungkinan bahwa preferensi seseorang dapat berubah dari waktu ke waktu, sedangkan pengalaman kita menunjukkan sebaliknya.

Terlepas dari kewajaran aksioma dan pentingnya kesimpulan yang mengikutinya, teori utilitas telah menjadi sasaran serangkaian keberatan terus-menerus sejak pertama kali dikenal luas. Sebagian orang membencinya karena dianggap mereduksi segalanya menjadi uang dan keegoisan. (Teori ini dicemooh sebagai "Amerika" oleh beberapa penulis Prancis, meskipun akarnya berasal dari Prancis.) Faktanya, sangat rasional untuk ingin menjalani hidup dengan penyangkalan diri, hanya ingin mengurangi penderitaan orang lain. Altruisme hanya berarti memberikan bobot yang besar pada kesejahteraan orang lain dalam mengevaluasi masa depan tertentu.

Keberatan lain berkaitan dengan kesulitan memperoleh probabilitas dan nilai utilitas yang diperlukan dan mengalikannya untuk menghitung utilitas yang diharapkan. Keberatan ini hanya mencampuradukkan dua hal yang berbeda: memilih tindakan rasional dan memilihnya dengan menghitung utilitas yang diharapkan. Misalnya, jika Anda mencoba menusuk bola mata Anda dengan jari, kelopak mata Anda akan menutup untuk melindungi mata Anda; ini rasional, tetapi tidak ada perhitungan utilitas yang diharapkan yang terlibat. Atau misalkan Anda sedang mengendarai sepeda menuruni bukit tanpa rem dan memiliki pilihan antara menabrak satu dinding beton dengan kecepatan sepuluh mil per jam atau dinding lain, lebih jauh di bawah bukit, dengan kecepatan dua puluh mil per jam; mana yang akan Anda pilih? Jika Anda memilih sepuluh mil per jam, selamat! Apakah Anda menghitung utilitas yang diharapkan? Mungkin tidak.

Namun, pilihan kecepatan sepuluh mil per jam tetap rasional. Hal ini didasarkan pada dua asumsi dasar: pertama, Anda lebih menyukai cedera yang kurang parah daripada cedera yang lebih parah, dan kedua, untuk tingkat cedera tertentu, peningkatan kecepatan tabrakan meningkatkan probabilitas melebihi tingkat tersebut. Dari kedua asumsi ini, secara matematis tanpa mempertimbangkan angka apa pun ternyata tabrakan pada kecepatan sepuluh mil per jam memiliki utilitas harapan yang lebih tinggi daripada tabrakan pada kecepatan dua puluh mil per jam. Singkatnya, memaksimalkan utilitas harapan mungkin tidak memerlukan perhitungan harapan atau utilitas apa pun. Ini adalah deskripsi eksternal murni dari entitas rasional.

Kritik lain terhadap teori rasionalitas terletak pada identifikasi lokus pengambilan keputusan. Artinya, hal-hal apa yang dianggap sebagai agen? Tampaknya jelas bahwa manusia adalah agen, tetapi bagaimana dengan keluarga, suku, korporasi, budaya, dan negara-bangsa? Jika kita meneliti serangga sosial seperti semut, apakah masuk akal untuk menganggap seekor semut sebagai agen yang cerdas, atau apakah kecerdasan sebenarnya terletak pada koloni secara keseluruhan, dengan semacam otak komposit yang terdiri dari banyak otak dan tubuh semut yang saling terhubung melalui sinyal feromon, bukan sinyal listrik? Dari sudut pandang evolusi, ini mungkin cara berpikir yang lebih produktif tentang semut, karena semut dalam koloni tertentu biasanya memiliki hubungan kekerabatan yang dekat. Sebagai individu, semut

dan serangga sosial lainnya tampaknya kurang memiliki naluri untuk mempertahankan diri yang berbeda dari mempertahankan koloni: mereka akan selalu terjun ke medan perang melawan penjajah, bahkan dengan peluang bunuh diri.

Namun terkadang manusia akan melakukan hal yang sama bahkan untuk membela manusia yang tidak memiliki hubungan kekerabatan; Seolah-olah spesies tersebut mendapat manfaat dari kehadiran sebagian kecil individu yang bersedia mengorbankan diri dalam pertempuran, atau melakukan perjalanan eksplorasi yang liar dan spekulatif, atau memelihara keturunan orang lain. Dalam kasus seperti itu, analisis rasionalitas yang sepenuhnya berfokus pada individu jelas kehilangan sesuatu yang penting.

Keberatan utama lainnya terhadap teori utilitas bersifat empiris yaitu, didasarkan pada bukti eksperimental yang menunjukkan bahwa manusia tidak rasional. Kita gagal untuk menyesuaikan diri dengan aksioma secara sistematis. Tujuan saya di sini bukanlah untuk membela teori utilitas sebagai model formal perilaku manusia. Memang, manusia tidak mungkin berperilaku rasional. Preferensi kita mencakup seluruh kehidupan masa depan kita sendiri, kehidupan anak-anak dan cucu kita, dan kehidupan orang lain, yang hidup sekarang atau di masa depan.

Namun kita bahkan tidak dapat memainkan langkah yang tepat di papan catur, tempat kecil dan sederhana dengan aturan yang jelas dan cakrawala yang sangat pendek. Ini bukan karena preferensi kita tidak rasional tetapi karena kompleksitas masalah pengambilan keputusan. Sebagian besar struktur kognitif kita ada untuk mengkompensasi ketidaksesuaian antara otak kita yang kecil dan lambat dengan kompleksitas masalah pengambilan keputusan yang sangat besar dan tak terbayangkan yang kita hadapi sepanjang waktu.

Jadi, meskipun akan sangat tidak masuk akal untuk mendasarkan teori AI yang bermanfaat pada asumsi bahwa manusia itu rasional, cukup masuk akal untuk menganggap bahwa manusia dewasa memiliki preferensi yang kurang lebih konsisten terhadap kehidupan masa depan. Artinya, jika Anda entah bagaimana dapat menonton dua film, masing-masing menggambarkan secara detail dan luas kehidupan masa depan yang mungkin Anda jalani, sehingga masing-masing merupakan pengalaman virtual, Anda dapat mengatakan mana yang Anda sukai, atau menyatakan ketidakpedulian.

Klaim ini mungkin lebih kuat dari yang diperlukan, jika satu-satunya tujuan kita adalah untuk memastikan bahwa mesin yang cukup cerdas tidak membawa malapetaka bagi umat manusia. Gagasan tentang malapetaka itu sendiri menyiratkan kehidupan yang jelas-jelas tidak disukai. Untuk menghindari malapetaka, maka kita hanya perlu mengklaim bahwa manusia dewasa dapat mengenali masa depan yang membawa malapetaka ketika dijelaskan secara rinci. Tentu saja, preferensi manusia memiliki struktur yang jauh lebih rinci dan, mungkin, lebih mudah dipastikan daripada sekadar "bukan bencana lebih baik daripada bencana."

Teori AI yang bermanfaat sebenarnya dapat mengakomodasi inkonsistensi dalam preferensi manusia, tetapi bagian inkonsisten dari preferensi Anda tidak akan pernah dapat dipenuhi dan tidak ada yang dapat dilakukan AI untuk membantu. Misalnya, anggaplah preferensi Anda terhadap pizza melanggar aksioma transitivitas:

ROBOT :"Selamat datang di rumah! Mau pizza nanas?"
ANDA :"Tidak, Anda harus tahu saya lebih suka pizza polos daripada nanas."
ROBOT :"Oke, satu pizza polos akan segera datang!"
ANDA :"Tidak, terima kasih, saya lebih suka pizza sosis."
ROBOT :"Maaf, satu pizza sosis!"
ANDA :"Sebenarnya, saya lebih suka nanas daripada sosis."
ROBOT :"Maaf, nanas saja!"
ANDA :"Saya sudah bilang saya lebih suka yang polos daripada nanas."

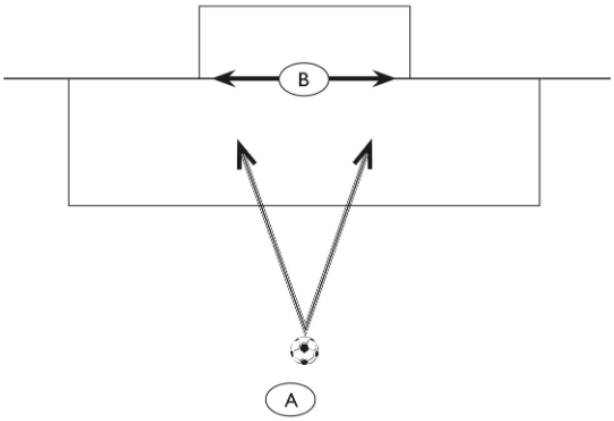
Tidak ada pizza yang bisa disajikan robot yang akan membuat Anda bahagia karena selalu ada pizza lain yang lebih Anda sukai. Robot hanya dapat memenuhi bagian yang konsisten dari preferensi Anda misalnya, katakanlah Anda lebih suka ketiga jenis pizza daripada tidak ada pizza sama sekali. Dalam hal ini, robot yang membantu dapat memberi Anda salah satu dari tiga pizza tersebut, sehingga memenuhi preferensi Anda untuk menghindari "tidak ada pizza" sambil membiarkan Anda merenungkan preferensi topping pizza Anda yang tidak konsisten dan menjengkelkan dengan santai.

Rasionalitas untuk Dua Orang

Gagasan dasar bahwa agen rasional bertindak untuk memaksimalkan utilitas yang diharapkan cukup sederhana, meskipun melakukannya sebenarnya sangat kompleks. Namun, teori ini hanya berlaku dalam kasus satu agen yang bertindak sendiri. Dengan lebih dari satu agen, gagasan bahwa mungkin setidaknya secara prinsip untuk menetapkan probabilitas pada berbagai hasil tindakan seseorang menjadi bermasalah. Alasannya adalah sekarang ada bagian dari dunia agen lain yang mencoba menebak tindakan apa yang akan Anda lakukan, dan sebaliknya, sehingga tidak jelas bagaimana menetapkan probabilitas bagaimana bagian dunia itu akan berperilaku. Dan tanpa probabilitas, definisi tindakan rasional sebagai memaksimalkan utilitas yang diharapkan tidak berlaku.

Segera setelah ada orang lain yang datang, maka, seorang agen akan membutuhkan cara lain untuk membuat keputusan rasional. Di sinilah teori permainan berperan. Terlepas dari namanya, teori permainan tidak selalu tentang permainan dalam arti biasa; ini adalah upaya umum untuk memperluas gagasan rasionalitas ke situasi dengan banyak agen. Ini jelas penting untuk tujuan kita, karena kita belum berencana (belum) untuk membangun robot yang hidup di planet tak berpenghuni di sistem bintang lain; kita akan menempatkan robot di dunia kita, yang dihuni oleh kita. Untuk memperjelas mengapa kita membutuhkan teori permainan, mari kita lihat contoh sederhana: Alice dan Bob bermain sepak bola di halaman belakang (gambar 2.1).

Alice akan melakukan tendangan penalti dan Bob berada di gawang. Alice akan menendang ke kiri atau ke kanan Bob. Karena dia kidal, akan sedikit lebih mudah dan akurat bagi Alice untuk menendang ke kanan Bob.



Gambar 2.1 Alice akan melakukan tendangan penalti melawan Bob.

Karena tendangan Alice sangat keras, Bob tahu dia harus segera melompat ke salah satu arah, dia tidak akan punya waktu untuk menunggu dan melihat ke arah mana bola akan pergi. Bob bisa berpikir seperti ini: “Alice memiliki peluang lebih baik untuk mencetak gol jika dia menendang ke kanan saya, karena dia kidal, jadi dia akan memilih itu, jadi saya akan melompat ke kanan.” Tetapi Alice bukanlah orang bodoh dan dapat membayangkan Bob berpikir seperti ini, dalam hal ini dia akan menendang ke kiri Bob. Tetapi Bob bukanlah orang bodoh dan dapat membayangkan Alice berpikir seperti ini, dalam hal ini dia akan melompat ke kirinya. Namun Alice bukanlah orang bodoh dan dapat membayangkan Bob berpikir seperti ini. Oke, Anda mengerti maksudnya. Dengan kata lain: jika ada pilihan rasional bagi Alice, Bob juga dapat mengetahuinya, mengantisipasinya, dan mencegah Alice mencetak gol, sehingga pilihan tersebut sejak awal tidak mungkin rasional.

Sejak tahun 1713 sekali lagi, dalam analisis permainan judi, solusi untuk teka-teki ini telah ditemukan. Kuncinya bukanlah memilih satu tindakan tertentu, tetapi memilih strategi acak. Misalnya, Alice dapat memilih strategi "menembak ke kanan Bob dengan probabilitas 55 persen dan menembak ke kiri Bob dengan probabilitas 45 persen." Bob dapat memilih "menyelam ke kanan dengan probabilitas 60 persen dan ke kiri dengan probabilitas 40 persen." Masing-masing secara mental melempar koin yang bias tepat sebelum bertindak, sehingga mereka tidak mengungkapkan niat mereka. Dengan bertindak secara tidak terduga, Alice dan Bob menghindari kontradiksi dari paragraf sebelumnya. Bahkan jika Bob mengetahui strategi acak Alice, tidak banyak yang dapat dia lakukan tanpa bola kristal.

Pertanyaan selanjutnya adalah, Berapa seharusnya probabilitasnya? Apakah pilihan Alice sebesar 55 persen–45 persen rasional? Nilai spesifik bergantung pada seberapa akurat Alice saat menembak ke kanan Bob, seberapa baik Bob menyelamatkan tembakan saat ia menyelam ke arah yang benar, dan sebagainya. (Lihat catatan untuk analisis lengkapnya.) Namun, kriteria umumnya sangat sederhana:

1. Strategi Alice adalah yang terbaik yang dapat ia rancang, dengan asumsi strategi Bob tetap.
2. Strategi Bob adalah yang terbaik yang dapat ia rancang, dengan asumsi strategi Alice tetap.

Jika kedua kondisi tersebut terpenuhi, kita katakan bahwa strategi tersebut berada dalam keseimbangan. Keseimbangan semacam ini disebut keseimbangan Nash untuk menghormati John Nash, yang pada tahun 1950 di usia dua puluh dua tahun, membuktikan bahwa keseimbangan semacam itu ada untuk sejumlah agen dengan preferensi rasional apa pun dan terlepas dari aturan permainannya. Setelah berjuang selama beberapa dekade dengan skizofrenia, Nash akhirnya pulih dan dianugerahi Hadiah Nobel Memorial di bidang Ekonomi untuk karyanya ini pada tahun 1994. Untuk pertandingan sepak bola Alice dan Bob, hanya ada satu keseimbangan. Dalam kasus lain, mungkin ada beberapa, sehingga konsep keseimbangan Nash, tidak seperti keputusan utilitas yang diharapkan, tidak selalu mengarah pada rekomendasi unik tentang bagaimana harus berperilaku.

Lebih buruk lagi, ada situasi di mana keseimbangan Nash tampaknya mengarah pada hasil yang sangat tidak diinginkan. Salah satu kasus tersebut adalah dilema tahanan yang terkenal, yang dinamai demikian oleh penasihat PhD Nash, Albert Tucker, pada tahun 1950. Permainan ini adalah model abstrak dari situasi dunia nyata yang terlalu umum di mana kerja sama timbal balik akan lebih baik bagi semua pihak yang terlibat tetapi orang-orang tetap memilih kehancuran bersama.

Dilema tahanan bekerja sebagai berikut: Alice dan Bob adalah tersangka dalam suatu kejahatan dan sedang diinterogasi secara terpisah. Masing-masing memiliki pilihan: untuk mengaku kepada polisi dan membocorkan informasi tentang kaki tangannya, atau untuk menolak berbicara. Jika keduanya menolak, mereka dihukum dengan tuduhan yang lebih ringan dan menjalani hukuman dua tahun; jika keduanya mengaku, mereka dihukum dengan tuduhan yang lebih serius dan menjalani hukuman sepuluh tahun; Jika seseorang mengaku dan yang lain menolak, orang yang mengaku akan bebas dan kaki tangannya menjalani hukuman dua puluh tahun.

Sekarang, Alice beralasan sebagai berikut: “Jika Bob akan mengaku, maka saya juga harus mengaku (sepuluh tahun lebih baik daripada dua puluh tahun); jika dia akan menolak, maka saya harus mengaku (bebas lebih baik daripada menghabiskan dua tahun di penjara); jadi bagaimanapun juga, saya harus mengaku.” Bob beralasan dengan cara yang sama. Dengan demikian, mereka berdua akhirnya mengaku atas kejahatan mereka dan menjalani hukuman sepuluh tahun, meskipun dengan menolak bersama-sama mereka bisa saja hanya menjalani hukuman dua tahun. Masalahnya adalah penolakan bersama bukanlah keseimbangan Nash, karena masing-masing memiliki insentif untuk membelot dan bebas dengan mengaku.

Perhatikan bahwa Alice bisa saja beralasan sebagai berikut: “Apa pun alasan yang saya buat, Bob juga akan melakukannya. Jadi kita akan berakhir memilih hal yang sama. Karena penolakan bersama lebih baik daripada pengakuan bersama, kita harus menolak.” Bentuk penalaran ini mengakui bahwa, sebagai agen rasional, Alice dan Bob akan membuat pilihan yang berkorelasi daripada independen. Ini hanyalah salah satu dari banyak pendekatan yang telah dicoba oleh para ahli teori permainan dalam upaya mereka untuk mendapatkan solusi yang kurang menyedihkan untuk dilema tahanan.

Contoh terkenal lain dari keseimbangan yang tidak diinginkan adalah tragedi commons, yang pertama kali dianalisis pada tahun 1833 oleh ekonom Inggris William Lloyd tetapi diberi

nama, dan menarik perhatian global, oleh ahli ekologi Garrett Hardin pada tahun 1968. Tragedi ini muncul ketika beberapa orang dapat mengonsumsi sumber daya bersama seperti lahan penggembalaan umum atau stok ikan yang pemulihannya lambat.

Tanpa adanya batasan sosial atau hukum, satu-satunya keseimbangan Nash di antara agen yang egois (non-altruistik) adalah setiap orang mengonsumsi sebanyak mungkin, yang menyebabkan keruntuhan sumber daya yang cepat. Solusi ideal, di mana setiap orang berbagi sumber daya sedemikian rupa sehingga total konsumsi berkelanjutan, bukanlah keseimbangan karena setiap individu memiliki insentif untuk berbuat curang dan mengambil lebih dari bagian yang adil membebankan biaya kepada orang lain.

Dalam praktiknya, tentu saja, manusia terkadang menghindari tragedi ini dengan menetapkan mekanisme seperti kuota dan hukuman atau skema penetapan harga. Mereka dapat melakukan ini karena mereka tidak terbatas pada memutuskan berapa banyak yang akan dikonsumsi; mereka juga dapat memutuskan untuk berkomunikasi. Dengan memperluas masalah pengambilan keputusan dengan cara ini, kita menemukan solusi yang lebih baik untuk semua orang.

Contoh-contoh ini, dan banyak lainnya, menggambarkan fakta bahwa memperluas teori pengambilan keputusan rasional ke banyak agen menghasilkan banyak perilaku yang menarik dan kompleks. Ini juga sangat penting karena, seperti yang seharusnya jelas, ada lebih dari satu manusia. Dan segera akan ada mesin cerdas juga. Tak perlu dikatakan, kita harus mencapai kerja sama timbal balik, yang menghasilkan manfaat bagi manusia, daripada kehancuran timbal balik.

2.2 KOMPUTER UNIVERSAL: FONDASI AI MODERN

Memiliki definisi kecerdasan yang masuk akal adalah bahan pertama dalam menciptakan mesin cerdas. Bahan kedua adalah mesin di mana definisi itu dapat diwujudkan. Karena alasan yang akan segera menjadi jelas, mesin itu adalah komputer. Bisa saja sesuatu yang berbeda misalnya, kita mungkin telah mencoba membuat mesin cerdas dari reaksi kimia yang kompleks atau dengan membajak sel biologis tetapi perangkat yang dibangun untuk komputasi, dari kalkulator mekanik paling awal hingga seterusnya, selalu tampak bagi penemunya sebagai rumah alami bagi kecerdasan.

Kita sekarang sudah sangat terbiasa dengan komputer sehingga kita hampir tidak menyadari kekuatan luar biasa mereka. Jika Anda memiliki laptop atau desktop atau ponsel pintar, lihatlah: sebuah kotak kecil, dengan cara untuk mengetik karakter. Hanya dengan mengetik, Anda dapat membuat program yang mengubah kotak itu menjadi sesuatu yang baru, mungkin sesuatu yang secara ajaib mensintesis gambar bergerak kapal laut yang menabrak gunung es atau planet asing dengan orang-orang biru tinggi; ketik lagi, dan ia menerjemahkan bahasa Inggris ke bahasa Mandarin; ketik lagi, dan ia mendengarkan dan berbicara; ketik lagi, dan ia mengalahkan juara catur dunia.

Kemampuan sebuah kotak tunggal untuk menjalankan proses apa pun yang dapat Anda bayangkan disebut universalitas, sebuah konsep yang pertama kali diperkenalkan oleh Alan Turing pada tahun 1936. Universalitas berarti bahwa kita tidak memerlukan mesin terpisah

untuk aritmatika, penerjemahan mesin, catur, pemahaman ucapan, atau animasi: satu mesin melakukan semuanya.

Laptop Anda pada dasarnya identik dengan pusat data server yang luas yang dijalankan oleh perusahaan TI terbesar di dunia bahkan yang dilengkapi dengan unit pemrosesan tensor khusus untuk pembelajaran mesin. Laptop Anda juga pada dasarnya identik dengan semua perangkat komputasi masa depan yang belum ditemukan. Laptop dapat melakukan tugas yang sama persis, asalkan memiliki cukup memori; hanya saja membutuhkan waktu yang jauh lebih lama.

Makalah Turing yang memperkenalkan universalitas adalah salah satu makalah terpenting yang pernah ditulis. Di dalamnya, ia menggambarkan sebuah perangkat komputasi sederhana yang dapat menerima sebagai masukan deskripsi dari perangkat komputasi lain mana pun, bersama dengan masukan perangkat kedua tersebut, dan, dengan mensimulasikan operasi perangkat kedua pada masukannya, menghasilkan keluaran yang sama dengan yang akan dihasilkan oleh perangkat kedua tersebut. Kita sekarang menyebut perangkat pertama ini sebagai mesin Turing universal. Untuk membuktikan universalitasnya, Turing memperkenalkan definisi yang tepat untuk dua jenis objek matematika baru: mesin dan program. Bersama-sama, mesin dan program mendefinisikan urutan peristiwa khususnya, urutan perubahan keadaan dalam mesin dan memorinya.

Dalam sejarah matematika, jenis objek baru muncul sangat jarang. Matematika dimulai dengan angka pada awal sejarah yang tercatat. Kemudian, sekitar 2000 SM, orang Mesir kuno dan Babilonia bekerja dengan objek geometris (titik, garis, sudut, luas, dan sebagainya). Matematikawan Tiongkok memperkenalkan matriks selama milenium pertama SM, sementara himpunan sebagai objek matematika baru muncul pada abad kesembilan belas. Objek baru Turing mesin dan program mungkin merupakan objek matematika paling ampuh yang pernah diciptakan. Ironisnya, bidang matematika sebagian besar gagal menyadari hal ini, dan sejak tahun 1940-an dan seterusnya, komputer dan komputasi telah menjadi ranah departemen teknik di sebagian besar universitas besar.

Bidang yang muncul ilmu komputer meledak selama tujuh puluh tahun berikutnya, menghasilkan berbagai macam konsep, desain, metode, dan aplikasi baru, serta tujuh dari delapan perusahaan paling berharga di dunia. Konsep sentral dalam ilmu komputer adalah algoritma, yang merupakan metode yang ditentukan secara tepat untuk menghitung sesuatu.

Algoritma kini telah menjadi bagian yang familiar dalam kehidupan sehari-hari: algoritma akar kuadrat dalam kalkulator saku menerima angka sebagai input dan mengembalikan akar kuadrat dari angka tersebut sebagai output; algoritma permainan catur mengambil posisi catur dan mengembalikan langkah; algoritma pencarian rute mengambil lokasi awal, lokasi tujuan, dan peta jalan dan mengembalikan rute tercepat dari awal ke tujuan.

Algoritma dapat dijelaskan dalam bahasa Inggris atau dalam notasi matematika, tetapi untuk diimplementasikan, algoritma tersebut harus dikodekan sebagai program dalam bahasa pemrograman. Algoritma yang lebih kompleks dapat dibangun dengan menggunakan algoritma yang lebih sederhana sebagai blok bangunan yang disebut subrutin misalnya, mobil otonom mungkin menggunakan algoritma pencarian rute sebagai subrutin sehingga ia tahu ke

mana harus pergi. Dengan cara ini, sistem perangkat lunak dengan kompleksitas yang sangat besar dibangun, lapis demi lapis.

Perangkat keras komputer penting karena komputer yang lebih cepat dengan lebih banyak memori memungkinkan algoritma untuk berjalan lebih cepat dan menangani lebih banyak informasi. Kemajuan di bidang ini sudah dikenal luas tetapi masih sangat mencengangkan. Komputer elektronik terprogram komersial pertama, Ferranti Mark I, dapat mengeksekusi sekitar seribu (10^3) instruksi per detik dan memiliki sekitar seribu byte memori utama. Komputer tercepat pada awal tahun 2019, mesin Summit di Laboratorium Nasional Oak Ridge di Tennessee, mengeksekusi sekitar 10^{18} instruksi per detik (seribu triliun kali lebih cepat) dan memiliki $2,5 \times 10^{17}$ byte memori (250 triliun kali lebih banyak). Kemajuan ini dihasilkan dari kemajuan dalam perangkat elektronik dan bahkan dalam fisika yang mendasarinya, memungkinkan tingkat miniaturisasi yang luar biasa.

Meskipun perbandingan antara komputer dan otak tidak terlalu bermakna, angka untuk Summit sedikit melebihi kapasitas mentah otak manusia, yang, seperti yang telah disebutkan sebelumnya, memiliki sekitar 10^{15} sinapsis dan "waktu siklus" sekitar seperseratus detik, untuk maksimum teoritis sekitar 10^{17} "operasi" per detik. Perbedaan terbesar adalah konsumsi daya: Summit menggunakan daya sekitar satu juta kali lebih banyak.

Hukum Moore, pengamatan empiris bahwa jumlah komponen elektronik pada sebuah chip berlipat ganda setiap dua tahun, diperkirakan akan berlanjut hingga sekitar tahun 2025, meskipun dengan laju yang sedikit lebih lambat. Selama beberapa tahun, kecepatan telah dibatasi oleh sejumlah besar panas yang dihasilkan oleh peralihan cepat transistor silikon; terlebih lagi, ukuran sirkuit tidak dapat menjadi jauh lebih kecil karena kabel dan konektor (pada tahun 2019) tidak lebih dari dua puluh lima atom lebarnya dan lima hingga sepuluh atom tebalnya. Setelah tahun 2025, kita perlu menggunakan fenomena fisik yang lebih eksotis termasuk perangkat kapasitansi negatif, transistor atom tunggal, nanotube graphene, dan fotonika untuk menjaga agar hukum Moore (atau penerusnya) tetap berlaku.

Alih-alih hanya mempercepat komputer tujuan umum, kemungkinan lain adalah membangun perangkat tujuan khusus yang disesuaikan untuk melakukan hanya satu kelas komputasi. Misalnya, unit pemrosesan tensor (TPU) Google dirancang untuk melakukan perhitungan yang diperlukan untuk algoritma pembelajaran mesin tertentu. Satu pod TPU (versi 2018) melakukan sekitar 10^{17} perhitungan per detik hampir sebanyak mesin Summit, tetapi menggunakan daya sekitar seratus kali lebih sedikit dan seratus kali lebih kecil. Bahkan jika teknologi chip yang mendasarinya tetap kurang lebih konstan, jenis mesin ini dapat dibuat semakin besar untuk menyediakan sejumlah besar daya komputasi mentah untuk sistem AI.

Komputasi kuantum adalah hal yang berbeda. Ia menggunakan sifat-sifat aneh dari fungsi gelombang mekanika kuantum untuk mencapai sesuatu yang luar biasa: dengan dua kali lipat jumlah perangkat keras kuantum, Anda dapat melakukan lebih dari dua kali lipat jumlah komputasi! Secara kasar, cara kerjanya seperti ini: Misalkan Anda memiliki perangkat fisik kecil yang menyimpan bit kuantum, atau qubit. Sebuah qubit memiliki dua kemungkinan keadaan, 0 dan 1. Sedangkan dalam fisika klasik perangkat qubit harus berada dalam salah satu dari dua keadaan tersebut, dalam fisika kuantum fungsi gelombang yang membawa

informasi tentang qubit mengatakan bahwa ia berada dalam kedua keadaan secara bersamaan. Jika Anda memiliki dua qubit, ada empat kemungkinan keadaan gabungan: 00, 01, 10, dan 11. Jika fungsi gelombang terjerat secara koheren di antara kedua qubit, yang berarti tidak ada proses fisik lain yang mengganggunya, maka kedua qubit berada dalam keempat keadaan tersebut secara bersamaan.

Selain itu, jika kedua qubit dihubungkan ke sirkuit kuantum yang melakukan beberapa perhitungan, maka perhitungan tersebut berlangsung dengan keempat keadaan tersebut secara bersamaan. Dengan tiga qubit, Anda mendapatkan delapan keadaan yang diproses secara simultan, dan seterusnya. Sekarang, ada beberapa keterbatasan fisik sehingga jumlah pekerjaan yang dilakukan kurang dari eksponensial terhadap jumlah qubit, tetapi kita tahu bahwa ada masalah penting di mana komputasi kuantum terbukti lebih efisien daripada komputer klasik mana pun.

Pada tahun 2019, terdapat prototipe eksperimental prosesor kuantum kecil yang beroperasi dengan beberapa puluh qubit, tetapi belum ada tugas komputasi menarik yang membuat prosesor kuantum lebih cepat daripada komputer klasik. Kesulitan utamanya adalah dekoherensi, proses seperti derau termal yang mengacaukan koherensi fungsi gelombang multi-qubit. Para ilmuwan kuantum berharap dapat memecahkan masalah dekoherensi dengan memperkenalkan sirkuit koreksi kesalahan, sehingga setiap kesalahan yang terjadi dalam komputasi dapat dengan cepat dideteksi dan dikoreksi melalui semacam proses pemungutan suara.

Sayangnya, sistem koreksi kesalahan membutuhkan lebih banyak qubit untuk melakukan pekerjaan yang sama: sementara mesin kuantum dengan beberapa ratus qubit sempurna akan sangat kuat dibandingkan dengan komputer klasik yang ada, kita mungkin membutuhkan beberapa juta qubit koreksi kesalahan untuk benar-benar mewujudkan komputasi tersebut. Beralih dari beberapa puluh ke beberapa juta qubit akan membutuhkan waktu beberapa tahun. Jika, pada akhirnya, kita sampai di sana, itu akan sepenuhnya mengubah gambaran tentang apa yang dapat kita lakukan dengan komputasi brute-force semata. Daripada menunggu kemajuan konseptual nyata dalam AI, kita mungkin dapat menggunakan kekuatan mentah komputasi kuantum untuk melewati beberapa hambatan yang dihadapi oleh algoritma "tidak cerdas" saat ini.

Batasan Komputasi

Bahkan pada tahun 1950-an, komputer digambarkan di media populer sebagai "otak super" yang "lebih cepat daripada Einstein." Jadi, dapatkah kita mengatakan sekarang, akhirnya, bahwa komputer sekuat otak manusia? Tidak. Berfokus pada kekuatan komputasi mentah sama sekali tidak tepat. Kecepatan saja tidak akan memberi kita AI. Menjalankan algoritma yang dirancang buruk pada komputer yang lebih cepat tidak membuat algoritma menjadi lebih baik; itu hanya berarti Anda mendapatkan jawaban yang salah lebih cepat. (Dan dengan lebih banyak data, ada lebih banyak peluang untuk jawaban yang salah!) Efek utama dari mesin yang lebih cepat adalah mempersingkat waktu untuk eksperimen, sehingga penelitian dapat berkembang lebih cepat. Bukan perangkat keras yang menghambat AI;

melainkan perangkat lunak. Kita belum tahu bagaimana membuat mesin yang benar-benar cerdas bahkan jika ukurannya sebesar alam semesta.

Namun, anggaplah kita berhasil mengembangkan jenis perangkat lunak AI yang tepat. Apakah ada batasan yang ditetapkan oleh fisika tentang seberapa kuat komputer dapat bekerja? Akankah batasan tersebut mencegah kita memiliki daya komputasi yang cukup untuk menciptakan AI yang sebenarnya? Jawabannya tampaknya adalah ya, ada batasan, dan tidak, tidak ada secercah pun kemungkinan bahwa batasan tersebut akan mencegah kita menciptakan AI yang sebenarnya. Fisikawan MIT Seth Lloyd telah memperkirakan batasan untuk komputer seukuran laptop, berdasarkan pertimbangan dari teori kuantum dan entropi.

Angka-angka tersebut bahkan akan membuat Carl Sagan mengangkat alisnya: 10^{51} operasi per detik dan 10^{30} byte memori, atau sekitar satu miliar triliun triliun kali lebih cepat dan empat triliun kali lebih banyak memori daripada Summit yang, seperti yang disebutkan sebelumnya, memiliki daya mentah lebih besar daripada otak manusia. Oleh karena itu, ketika seseorang mendengar saran bahwa pikiran manusia mewakili batas atas dari apa yang secara fisik dapat dicapai di alam semesta kita, setidaknya seseorang harus meminta klarifikasi lebih lanjut.

Selain batasan yang dikenakan oleh fisika, ada batasan lain pada kemampuan komputer yang berasal dari karya para ilmuwan komputer. Turing sendiri membuktikan bahwa beberapa masalah tidak dapat diputuskan oleh komputer mana pun: masalahnya didefinisikan dengan baik, ada jawabannya, tetapi tidak mungkin ada algoritma yang selalu menemukan jawaban itu. Dia memberikan contoh dari apa yang kemudian dikenal sebagai masalah penghentian: Dapatkah suatu algoritma memutuskan apakah suatu program tertentu memiliki "loop tak terbatas" yang mencegahnya untuk pernah selesai?

Bukti Turing bahwa tidak ada algoritma yang dapat menyelesaikan masalah penghentian sangat penting untuk fondasi matematika, tetapi tampaknya tidak berpengaruh pada masalah apakah komputer dapat cerdas. Salah satu alasan klaim ini adalah bahwa keterbatasan dasar yang sama tampaknya berlaku untuk otak manusia. Begitu Anda mulai meminta otak manusia untuk melakukan simulasi yang tepat tentang dirinya sendiri yang mensimulasikan dirinya sendiri yang mensimulasikan dirinya sendiri, dan seterusnya, Anda pasti akan menemui kesulitan. Saya sendiri tidak pernah khawatir tentang ketidakmampuan saya untuk melakukan ini.

Oleh karena itu, fokus pada masalah yang dapat diputuskan tampaknya tidak memberikan batasan nyata pada AI. Namun, ternyata dapat diputuskan tidak berarti mudah. Ilmuwan komputer menghabiskan banyak waktu untuk memikirkan kompleksitas masalah, yaitu, pertanyaan tentang berapa banyak komputasi yang dibutuhkan untuk menyelesaikan masalah dengan metode yang paling efisien. Berikut adalah masalah yang mudah: diberikan daftar seribu angka, temukan angka terbesar.

Jika dibutuhkan satu detik untuk memeriksa setiap angka, maka dibutuhkan seribu detik untuk menyelesaikan masalah ini dengan metode yang jelas, yaitu memeriksa setiap angka secara bergantian dan melacak angka terbesar. Apakah ada metode yang lebih cepat?

Tidak, karena jika suatu metode tidak memeriksa beberapa angka dalam daftar, angka tersebut mungkin adalah yang terbesar, dan metode tersebut akan gagal.

Jadi, waktu untuk menemukan elemen terbesar sebanding dengan ukuran daftar. Seorang ilmuwan komputer akan mengatakan bahwa masalah ini memiliki kompleksitas linier, artinya sangat mudah; kemudian dia akan mencari sesuatu yang lebih menarik untuk dikerjakan. Yang membuat para ilmuwan komputer teoretis bersemangat adalah fakta bahwa banyak masalah tampaknya memiliki kompleksitas eksponensial dalam kasus terburuk. Ini berarti dua hal: pertama, semua algoritma yang kita ketahui membutuhkan waktu eksponensial yaitu, sejumlah waktu yang eksponensial terhadap ukuran input untuk menyelesaikan setidaknya beberapa contoh masalah; kedua, para ilmuwan komputer teoretis cukup yakin bahwa algoritma yang lebih efisien tidak ada.

Pertumbuhan eksponensial dalam kesulitan berarti bahwa masalah mungkin dapat dipecahkan secara teori (yaitu, pasti dapat diputuskan) tetapi terkadang tidak dapat dipecahkan dalam praktik; kita menyebut masalah seperti itu sebagai masalah yang tidak dapat dipecahkan. Contohnya adalah masalah memutuskan apakah peta yang diberikan dapat diwarnai hanya dengan tiga warna, sehingga tidak ada dua wilayah yang berdekatan memiliki warna yang sama. (Sudah diketahui bahwa pewarnaan dengan empat warna berbeda selalu mungkin.) Dengan satu juta wilayah, mungkin ada beberapa kasus (tidak semua, tetapi beberapa) yang membutuhkan sekitar 2^{1000} langkah komputasi untuk menemukan jawabannya, yang berarti sekitar 10^{275} tahun pada superkomputer Summit atau hanya 10^{242} tahun pada laptop fisika canggih Seth Lloyd. Usia alam semesta, sekitar 10^{10} tahun, adalah titik kecil dibandingkan dengan ini.

Apakah keberadaan masalah yang tidak dapat dipecahkan memberi kita alasan untuk berpikir bahwa komputer tidak dapat secerdas manusia? Tidak. Tidak ada alasan untuk menganggap bahwa manusia juga dapat memecahkan masalah yang tidak dapat dipecahkan. Komputasi kuantum sedikit membantu (baik pada mesin maupun otak), tetapi tidak cukup untuk mengubah kesimpulan dasarnya.

Kompleksitas berarti bahwa masalah pengambilan keputusan di dunia nyata, masalah memutuskan apa yang harus dilakukan saat ini, setiap saat dalam hidup seseorang sangat sulit sehingga baik manusia maupun komputer tidak akan pernah mendekati solusi yang sempurna. Hal ini memiliki dua konsekuensi: pertama, kita memperkirakan bahwa, sebagian besar waktu, keputusan di dunia nyata akan paling baik hanya setengah layak dan tentu saja jauh dari optimal; kedua, kita memperkirakan bahwa sebagian besar arsitektur mental manusia dan komputer, cara proses pengambilan keputusan mereka beroperasi akan dirancang untuk mengatasi kompleksitas sejauh mungkin yaitu, untuk memungkinkan menemukan jawaban yang bahkan setengah layak meskipun kompleksitas dunia sangat besar.

Terakhir, kita memperkirakan bahwa dua konsekuensi pertama akan tetap berlaku tidak peduli seberapa cerdas dan kuatnya mesin di masa depan. Mesin tersebut mungkin jauh lebih mampu daripada kita, tetapi tetap jauh dari rasional sempurna.

2.3 EVOLUSI AI: DARI LOGIKA KE PROBABILITAS

Perkembangan logika oleh Aristoteles dan lainnya menyediakan aturan yang tepat untuk pemikiran rasional, tetapi kita tidak tahu apakah Aristoteles pernah mempertimbangkan kemungkinan mesin yang menerapkan aturan-aturan ini. Pada abad ke-13, filsuf, penggoda, dan mistikus Catalan yang berpengaruh, Ramon Llull, lebih mendekati hal itu: ia benar-benar membuat roda kertas yang bertuliskan simbol, yang dengannya ia dapat menghasilkan kombinasi logis dari pernyataan. Matematikawan Prancis abad ke-17 yang hebat, Blaise Pascal, adalah orang pertama yang mengembangkan kalkulator mekanik yang nyata dan praktis. Meskipun hanya dapat menambah dan mengurangi dan terutama digunakan di kantor penagihan pajak ayahnya, hal itu membuat Pascal menulis, "Mesin aritmatika menghasilkan efek yang tampak lebih dekat dengan pemikiran daripada semua tindakan hewan."

Teknologi mengalami lompatan dramatis ke depan pada abad ke-19 ketika matematikawan dan penemu Inggris, Charles Babbage, merancang Mesin Analitik, sebuah mesin universal yang dapat diprogram dalam arti yang kemudian didefinisikan oleh Turing. Ia dibantu dalam pekerjaannya oleh Ada, Countess of Lovelace, putri dari penyair romantis dan petualang Lord Byron. Sementara Babbage berharap untuk menggunakan Mesin Analitik untuk menghitung tabel matematika dan astronomi yang akurat, Lovelace memahami potensi sebenarnya, menggambarkannya pada tahun 1842 sebagai “mesin berpikir atau… mesin penalaran” yang dapat bernalar tentang “semua subjek di alam semesta.” Jadi, elemen konseptual dasar untuk menciptakan AI sudah ada! Dari titik itu, tentu saja, AI hanya akan menjadi masalah waktu….

Sayangnya, butuh waktu lama Mesin Analitik tidak pernah dibangun, dan ide-ide Lovelace sebagian besar dilupakan. Dengan karya teoretis Turing pada tahun 1936 dan dorongan selanjutnya dari Perang Dunia II, mesin komputasi universal akhirnya terwujud pada tahun 1940-an. Pemikiran tentang menciptakan kecerdasan segera menyusul. Makalah Turing tahun 1950, “Computing Machinery and Intelligence,” adalah yang paling terkenal dari banyak karya awal tentang kemungkinan mesin cerdas. Para skeptis sudah menyatakan bahwa mesin tidak akan pernah mampu melakukan X, untuk hampir semua X yang dapat Anda bayangkan, dan Turing membantah pernyataan tersebut. Ia juga mengusulkan sebuah tes operasional untuk kecerdasan, yang disebut permainan imitasi, yang kemudian (dalam bentuk yang disederhanakan) dikenal sebagai tes Turing. Tes ini mengukur perilaku mesin khususnya, kemampuannya untuk menipu seorang penanya manusia agar berpikir bahwa mesin itu adalah manusia.

Permainan imitasi memiliki peran khusus dalam makalah Turing yaitu sebagai eksperimen pemikiran untuk mengalihkan para skeptis yang beranggapan bahwa mesin tidak dapat berpikir dengan cara yang benar, untuk alasan yang benar, dengan jenis kesadaran yang benar. Turing berharap untuk mengarahkan kembali argumen ke masalah apakah sebuah mesin dapat berperilaku dengan cara tertentu; dan jika memang demikian jika mampu, misalnya, berdiskusi secara masuk akal tentang soneta Shakespeare dan maknanya maka skeptisisme tentang AI tidak dapat dipertahankan. Bertentangan dengan interpretasi umum, saya ragu bahwa tes tersebut dimaksudkan sebagai definisi kecerdasan yang sebenarnya,

dalam arti bahwa sebuah mesin cerdas jika dan hanya jika ia lulus tes Turing. Memang, Turing menulis, “Bukankah mesin dapat melakukan sesuatu yang seharusnya digambarkan sebagai berpikir tetapi sangat berbeda dari apa yang dilakukan manusia?” Alasan lain untuk tidak menganggap tes tersebut sebagai definisi untuk AI adalah karena itu adalah definisi yang buruk untuk digunakan. Dan karena alasan itu, para peneliti AI arus utama hampir tidak mengerahkan upaya untuk lulus tes Turing.

Tes Turing tidak berguna untuk AI karena itu adalah definisi informal dan sangat bergantung pada konteks: itu bergantung pada karakteristik pikiran manusia yang sangat rumit dan sebagian besar tidak diketahui, yang berasal dari biologi dan budaya. Tidak ada cara untuk "menguraikan" definisi tersebut dan bekerja mundur darinya untuk menciptakan mesin yang terbukti akan lulus tes. Sebaliknya, AI telah berfokus pada perilaku rasional, seperti yang dijelaskan sebelumnya: sebuah mesin cerdas sejauh apa yang dilakukannya kemungkinan akan mencapai apa yang diinginkannya, mengingat apa yang telah dipersepsikannya.

Awalnya, seperti Aristoteles, para peneliti AI mengidentifikasi "apa yang diinginkannya" dengan tujuan yang terpenuhi atau tidak. Tujuan-tujuan ini bisa berada di dunia mainan seperti puzzle 15, di mana tujuannya adalah untuk menyusun semua ubin bernomor secara berurutan dari 1 hingga 15 dalam nampan persegi kecil (simulasi); atau mungkin di lingkungan fisik nyata: pada awal tahun 1970-an, robot Shakey di SRI di California mendorong balok-balok besar ke dalam konfigurasi yang diinginkan, dan Freddy di Universitas Edinburgh merakit perahu kayu dari bagian-bagian komponennya. Semua pekerjaan ini dilakukan menggunakan pemecah masalah logis dan sistem perencanaan untuk membangun dan mengeksekusi rencana yang terjamin untuk mencapai tujuan.

Pada tahun 1980-an, jelas bahwa penalaran logis saja tidak cukup, karena, seperti yang telah disebutkan sebelumnya, tidak ada rencana yang dijamin akan membawa Anda ke bandara. Logika membutuhkan kepastian, dan dunia nyata tidak menyediakannya. Sementara itu, ilmuwan komputer Israel-Amerika Judea Pearl, yang kemudian memenangkan Penghargaan Turing 2011, telah mengerjakan metode untuk penalaran yang tidak pasti berdasarkan teori probabilitas. Para peneliti AI secara bertahap menerima ide-ide Pearl; mereka mengadopsi alat-alat teori probabilitas dan teori utilitas dan dengan demikian menghubungkan AI dengan bidang lain seperti statistik, teori kontrol, ekonomi, dan riset operasi. Perubahan ini menandai awal dari apa yang oleh beberapa pengamat disebut AI modern.

Agen dan Lingkungan

Konsep sentral AI modern adalah agen cerdas sesuatu yang merasakan dan bertindak. Agen adalah proses yang terjadi dari waktu ke waktu, dalam arti bahwa aliran masukan persepsi diubah menjadi aliran tindakan. Misalnya, anggaplah agen yang dimaksud adalah taksi swakemudi yang membawa saya ke bandara. Masukannya mungkin termasuk delapan kamera RGB yang beroperasi pada tiga puluh frame per detik; setiap frame terdiri dari mungkin 7,5 juta piksel, masing-masing dengan nilai intensitas gambar di masing-masing dari tiga saluran warna, dengan total lebih dari lima gigabyte per detik. (Aliran data dari dua ratus juta

fotoreseptor di retina bahkan lebih besar, yang sebagian menjelaskan mengapa penglihatan menempati sebagian besar otak manusia.)

Taksi juga menerima data dari akselerometer seratus kali per detik, serta data GPS. Banjir data mentah yang luar biasa ini diubah oleh kekuatan komputasi yang sangat besar dari miliaran transistor (atau neuron) menjadi perilaku mengemudi yang lancar dan kompeten. Tindakan taksi termasuk sinyal elektronik yang dikirim ke roda kemudi, rem, dan pedal gas, dua puluh kali per detik. (Bagi pengemudi manusia yang berpengalaman, sebagian besar aktivitas yang kacau ini tidak disadari: Anda mungkin hanya menyadari pengambilan keputusan seperti "menyalip truk yang lambat ini" atau "berhenti untuk mengisi bensin," tetapi mata, otak, saraf, dan otot Anda masih melakukan semua hal lainnya.)

Untuk program catur, inputnya sebagian besar hanya berupa detak jam, dengan pemberitahuan sesekali tentang langkah lawan dan keadaan papan yang baru, sementara tindakannya sebagian besar tidak melakukan apa pun saat program sedang berpikir, dan sesekali memilih langkah dan memberi tahu lawan. Untuk asisten digital pribadi, atau PDA, seperti Siri atau Cortana, inputnya tidak hanya mencakup sinyal akustik dari mikrofon (diambil sampelnya empat puluh delapan ribu kali per detik) dan input dari layar sentuh, tetapi juga konten setiap halaman web yang diaksesnya, sementara tindakannya mencakup berbicara dan menampilkan materi di layar.

Cara kita membangun agen cerdas bergantung pada sifat masalah yang kita hadapi. Hal ini, pada gilirannya, bergantung pada tiga hal: pertama, sifat lingkungan tempat agen akan beroperasi, papan catur adalah tempat yang sangat berbeda dari jalan raya yang ramai atau telepon seluler; kedua, pengamatan dan tindakan yang menghubungkan agen dengan lingkungan misalnya, Siri mungkin memiliki atau tidak memiliki akses ke kamera telepon sehingga dapat melihat; dan ketiga, tujuan agen mengajari lawan untuk bermain catur lebih baik adalah tugas yang sangat berbeda dari memenangkan permainan.

Sebagai contoh bagaimana desain agen bergantung pada hal-hal ini: Jika tujuannya adalah untuk memenangkan permainan, program catur hanya perlu mempertimbangkan keadaan papan saat ini dan tidak memerlukan memori tentang peristiwa masa lalu. Di sisi lain, tutor catur harus terus memperbarui modelnya tentang aspek-aspek catur mana yang dipahami atau tidak dipahami oleh murid sehingga dapat memberikan nasihat yang bermanfaat. Dengan kata lain, bagi tutor catur, pikiran murid adalah bagian yang relevan dari lingkungan. Selain itu, tidak seperti papan, pikiran adalah bagian dari lingkungan yang tidak dapat diamati secara langsung.

Karakteristik masalah yang memengaruhi desain agen mencakup setidaknya hal-hal berikut:

- apakah lingkungan tersebut dapat diamati sepenuhnya (seperti dalam catur, di mana input memberikan akses langsung ke semua aspek relevan dari keadaan lingkungan saat ini) atau sebagian dapat diamati (seperti dalam mengemudi, di mana bidang pandang seseorang terbatas, kendaraan tidak transparan, dan niat pengemudi lain misterius);

- apakah lingkungan dan tindakan bersifat diskrit (seperti dalam catur) atau secara efektif kontinu (seperti dalam mengemudi);
- apakah lingkungan tersebut mengandung agen lain (seperti dalam catur dan mengemudi) atau tidak (seperti dalam menemukan rute terpendek di peta);
- apakah hasil dari tindakan, sebagaimana ditentukan oleh "aturan" atau "fisika" lingkungan, dapat diprediksi (seperti dalam catur) atau tidak dapat diprediksi (seperti dalam lalu lintas dan cuaca), dan apakah aturan tersebut diketahui atau tidak diketahui;
- apakah lingkungan tersebut berubah secara dinamis, sehingga waktu untuk membuat keputusan sangat terbatas (seperti dalam mengemudi) atau tidak (seperti dalam optimasi strategi pajak);
- panjang rentang waktu di mana kualitas keputusan diukur sesuai dengan tujuannya ini bisa sangat pendek (seperti pada pengereman darurat), durasi menengah (seperti dalam catur, di mana permainan berlangsung hingga sekitar seratus langkah), atau sangat panjang (seperti mengantar saya ke bandara, yang mungkin membutuhkan ratusan ribu siklus keputusan jika taksi mengambil keputusan seratus kali per detik).

Seperti yang dapat dibayangkan, karakteristik ini menimbulkan berbagai macam jenis masalah yang membingungkan. Hanya dengan mengalikan pilihan yang tercantum di atas akan menghasilkan 192 jenis. Kita dapat menemukan contoh masalah di dunia nyata untuk semua jenis tersebut. Beberapa jenis biasanya dipelajari di bidang di luar AI misalnya, merancang autopilot yang mempertahankan penerbangan datar adalah masalah dinamis, kontinu, dan berrentang waktu pendek yang biasanya dipelajari di bidang teori kontrol. Jelas, beberapa jenis masalah lebih mudah daripada yang lain. AI telah membuat banyak kemajuan pada masalah seperti permainan papan dan teka-teki yang dapat diamati, diskrit, deterministik, dan memiliki aturan yang diketahui.

Untuk jenis masalah yang lebih mudah, para peneliti AI telah mengembangkan algoritma yang cukup umum dan efektif serta pemahaman teoritis yang solid; seringkali, mesin melampaui kinerja manusia pada jenis masalah ini. Kita dapat mengetahui bahwa suatu algoritma bersifat umum karena kita memiliki bukti matematis bahwa algoritma tersebut memberikan hasil optimal atau mendekati optimal dengan kompleksitas komputasi yang wajar di seluruh kelas masalah, dan karena algoritma tersebut bekerja dengan baik dalam praktik pada jenis masalah tersebut tanpa memerlukan modifikasi khusus masalah apa pun.

Permainan video seperti StarCraft jauh lebih sulit daripada permainan papan: permainan ini melibatkan ratusan bagian yang bergerak dan rentang waktu ribuan langkah, dan papan hanya sebagian terlihat pada waktu tertentu. Pada setiap titik, pemain mungkin memiliki pilihan setidaknya 10^{50} gerakan, dibandingkan dengan sekitar 10^{2} dalam Go. Di sisi lain, aturannya diketahui dan dunianya diskrit dengan hanya beberapa jenis objek.

Pada awal tahun 2019, mesin-mesin tersebut sudah sebaik beberapa pemain StarCraft profesional, tetapi belum siap untuk menantang manusia terbaik. Lebih penting lagi, dibutuhkan upaya khusus untuk mencapai titik tersebut; metode umum belum sepenuhnya siap untuk StarCraft.

Masalah seperti menjalankan pemerintahan atau mengajar biologi molekuler jauh lebih sulit. Masalah-masalah tersebut memiliki lingkungan yang kompleks dan sebagian besar tidak dapat diamati (keadaan suatu negara secara keseluruhan, atau keadaan pikiran seorang siswa), jauh lebih banyak objek dan jenis objek, tidak ada definisi yang jelas tentang apa tindakan yang harus dilakukan, sebagian besar aturan yang tidak diketahui, banyak ketidakpastian, dan skala waktu yang sangat panjang. Kita memiliki ide dan alat siap pakai yang menangani masing-masing karakteristik ini secara terpisah, tetapi belum ada metode umum yang dapat mengatasi semua karakteristik tersebut secara bersamaan. Ketika kita membangun sistem AI untuk jenis tugas ini, sistem tersebut cenderung membutuhkan banyak rekayasa khusus dan seringkali sangat rapuh.

Kemajuan menuju generalitas terjadi ketika kita merancang metode yang efektif untuk masalah yang lebih sulit dalam jenis tertentu atau metode yang membutuhkan asumsi yang lebih sedikit dan lebih lemah sehingga dapat diterapkan pada lebih banyak masalah. AI tujuan umum adalah metode yang dapat diterapkan di semua jenis masalah dan bekerja secara efektif untuk kasus-kasus besar dan sulit sambil membuat sangat sedikit asumsi. Itulah tujuan utama penelitian AI: sebuah sistem yang tidak memerlukan rekayasa khusus masalah dan dapat diminta untuk mengajar kelas biologi molekuler atau menjalankan pemerintahan. Sistem tersebut akan mempelajari apa yang perlu dipelajari dari semua sumber daya yang tersedia, mengajukan pertanyaan bila perlu, dan mulai merumuskan dan mengeksekusi rencana yang berhasil.

Metode tujuan umum seperti itu belum ada, tetapi kita semakin dekat. Mungkin mengejutkan, banyak kemajuan menuju AI umum ini berasal dari penelitian yang bukan tentang membangun sistem AI tujuan umum yang menakutkan. Ini berasal dari penelitian tentang AI alat atau AI sempit, yang berarti sistem AI yang bagus, aman, dan membosankan yang dirancang untuk masalah tertentu seperti bermain Go atau mengenali angka tulisan tangan. Penelitian tentang jenis AI ini sering dianggap tidak menimbulkan risiko karena bersifat spesifik masalah dan tidak ada hubungannya dengan AI tujuan umum.

Keyakinan ini muncul dari kesalahpahaman tentang jenis pekerjaan yang dilakukan pada sistem ini. Faktanya, penelitian tentang AI alat dapat dan sering menghasilkan kemajuan menuju AI tujuan umum, terutama ketika dilakukan oleh peneliti dengan selera yang baik yang menyerang masalah yang berada di luar kemampuan metode umum saat ini. Di sini, selera yang baik berarti bahwa pendekatan solusi bukan hanya pengkodean ad hoc tentang apa yang akan dilakukan orang cerdas dalam situasi tertentu, tetapi upaya untuk memberi mesin kemampuan untuk menemukan solusi sendiri.

Misalnya, ketika tim AlphaGo di Google DeepMind berhasil menciptakan program Go terbaik di dunia, mereka melakukannya tanpa benar-benar mengerjakan Go. Yang saya maksud adalah mereka tidak menulis banyak kode khusus Go yang mengatakan apa yang harus dilakukan dalam berbagai jenis situasi Go. Mereka tidak merancang prosedur pengambilan keputusan yang hanya berfungsi untuk Go. Sebaliknya, mereka melakukan peningkatan pada dua teknik yang cukup umum pencarian ke depan untuk membuat keputusan dan pembelajaran penguatan untuk mempelajari cara mengevaluasi posisi sehingga teknik

tersebut cukup efektif untuk memainkan Go pada tingkat superhuman. Peningkatan tersebut dapat diterapkan pada banyak masalah lain, termasuk masalah yang jauh lebih luas seperti robotika. Sebagai tambahan, sebuah versi AlphaGo yang disebut AlphaZero baru-baru ini belajar untuk mengalahkan AlphaGo dalam permainan Go, dan juga mengalahkan Stockfish (program catur terbaik di dunia, jauh lebih baik daripada manusia mana pun) dan Elmo (program shogi terbaik di dunia, juga lebih baik daripada manusia mana pun). AlphaZero melakukan semua ini dalam satu hari.

Terdapat juga kemajuan substansial menuju AI tujuan umum dalam penelitian tentang pengenalan angka tulisan tangan pada tahun 1990-an. Tim Yann LeCun di AT&T Labs tidak menulis algoritma khusus untuk mengenali "8" dengan mencari garis lengkung dan lingkaran; sebaliknya, mereka meningkatkan algoritma pembelajaran jaringan saraf yang ada untuk menghasilkan jaringan saraf konvolusional. Jaringan-jaringan tersebut, pada gilirannya, menunjukkan pengenalan karakter yang efektif setelah pelatihan yang sesuai pada contoh-contoh berlabel. Algoritma yang sama dapat belajar mengenali huruf, bentuk, rambu berhenti, anjing, kucing, dan mobil polisi.

Di bawah judul "pembelajaran mendalam," mereka telah merevolusi pengenalan suara dan pengenalan objek visual. Mereka juga merupakan salah satu komponen kunci dalam AlphaZero serta di sebagian besar proyek mobil otonom saat ini. Jika Anda memikirkannya, hampir tidak mengherankan bahwa kemajuan menuju AI umum akan terjadi dalam proyek AI sempit yang menangani tugas-tugas spesifik; tugas-tugas tersebut memberi para peneliti AI sesuatu untuk dikerjakan. (Ada alasan mengapa orang tidak mengatakan, "Menatap keluar jendela adalah ibu dari penemuan.") Pada saat yang sama, penting untuk memahami seberapa banyak kemajuan yang telah terjadi dan di mana batas-batasnya. Ketika AlphaGo mengalahkan Lee Sedol dan kemudian semua pemain Go top lainnya, banyak orang berasumsi bahwa karena sebuah mesin telah belajar dari awal untuk mengalahkan umat manusia dalam tugas yang diketahui sangat sulit bahkan untuk manusia yang sangat cerdas, itu adalah awal dari akhir hanya masalah waktu sebelum AI mengambil alih.

Bahkan beberapa orang yang skeptis mungkin akan yakin ketika AlphaZero menang dalam catur dan shogi serta Go. Tetapi AlphaZero memiliki keterbatasan yang nyata: ia hanya berfungsi dalam kelas permainan dua pemain yang diskrit, teramati, dan memiliki aturan yang diketahui. Pendekatan ini sama sekali tidak akan berhasil untuk mengemudi, mengajar, menjalankan pemerintahan, atau mengambil alih dunia.

Batasan tajam pada kompetensi mesin ini berarti bahwa ketika orang berbicara tentang "IQ mesin" yang meningkat pesat dan mengancam untuk melampaui IQ manusia, mereka berbicara omong kosong. Sejauh konsep IQ masuk akal ketika diterapkan pada manusia, itu karena kemampuan manusia cenderung berkorelasi di berbagai aktivitas kognitif. Mencoba menetapkan IQ pada mesin seperti mencoba membuat hewan berkaki empat berkompetisi dalam dasalomba manusia. Memang benar, kuda dapat berlari cepat dan melompat tinggi, tetapi mereka mengalami banyak kesulitan dalam lompat galah dan lempar cakram.

Tujuan dan Model Standar

Jika dilihat dari luar, agen cerdas adalah rangkaian tindakan yang dihasilkannya dari rangkaian masukan yang diterimanya. Dari dalam, tindakan harus dipilih oleh program agen. Manusia dilahirkan dengan satu program agen, bisa dibilang, dan program itu belajar dari waktu ke waktu untuk bertindak cukup berhasil di berbagai tugas. Sejauh ini, hal itu tidak berlaku untuk AI: kita tidak tahu bagaimana membangun satu program AI serbaguna yang melakukan segalanya, jadi sebagai gantinya kita membangun berbagai jenis program agen untuk berbagai jenis masalah. Saya perlu menjelaskan setidaknya sedikit tentang bagaimana berbagai program agen ini bekerja; penjelasan yang lebih rinci diberikan dalam lampiran di akhir buku bagi mereka yang tertarik. (Penunjuk ke lampiran tertentu diberikan sebagai superskrip seperti ini dan ini.) Fokus utama di sini adalah bagaimana model standar diwujudkan dalam berbagai jenis agen ini dengan kata lain, bagaimana tujuan ditentukan dan dikomunikasikan kepada agen.

Cara paling sederhana untuk mengkomunikasikan tujuan adalah dalam bentuk sasaran. Ketika Anda masuk ke mobil otonom dan menyentuh ikon "rumah" di layar, mobil tersebut menganggap ini sebagai tujuannya dan kemudian merencanakan serta menjalankan rute. Suatu keadaan dunia memenuhi tujuan (ya, saya di rumah) atau tidak (tidak, saya tidak tinggal di Bandara San Francisco). Pada periode klasik penelitian AI, sebelum ketidakpastian menjadi isu utama pada tahun 1980-an, sebagian besar penelitian AI mengasumsikan dunia yang sepenuhnya dapat diamati dan deterministik, dan tujuan masuk akal sebagai cara untuk menentukan sasaran.

Terkadang ada juga fungsi biaya untuk mengevaluasi solusi, sehingga solusi optimal adalah solusi yang meminimalkan total biaya sambil mencapai tujuan. Bagi mobil, ini mungkin sudah terintegrasi mungkin biaya suatu rute adalah kombinasi tetap dari waktu dan konsumsi bahan bakar atau manusia mungkin memiliki pilihan untuk menentukan pertukaran antara keduanya.

Kunci untuk mencapai tujuan tersebut adalah kemampuan untuk "mensimulasikan secara mental" efek dari tindakan yang mungkin dilakukan, yang terkadang disebut pencarian ke depan. Mobil otonom Anda memiliki peta internal, sehingga ia tahu bahwa berkendara ke timur dari San Francisco di Jembatan Bay akan membawa Anda ke Oakland. Algoritma yang berasal dari tahun 1960-an menemukan rute optimal dengan melihat ke depan dan mencari melalui banyak kemungkinan urutan tindakan. Algoritma ini membentuk bagian yang tak terpisahkan dari infrastruktur modern: algoritma ini tidak hanya menyediakan petunjuk arah mengemudi tetapi juga solusi perjalanan udara, perakitan robot, perencanaan konstruksi, dan logistik pengiriman. Dengan beberapa modifikasi untuk menangani perilaku lawan yang kurang ajar, ide yang sama tentang melihat ke depan berlaku untuk permainan seperti tic-tac-toe, catur, dan Go, di mana tujuannya adalah untuk menang sesuai dengan definisi kemenangan khusus permainan tersebut.

Algoritma melihat ke depan sangat efektif untuk tugas-tugas spesifiknya, tetapi algoritma ini tidak terlalu fleksibel. Sebagai contoh, AlphaGo "mengetahui" aturan Go, tetapi hanya dalam arti bahwa ia memiliki dua subrutin, yang ditulis dalam bahasa pemrograman

tradisional seperti C++: satu subrutin menghasilkan semua kemungkinan langkah legal dan yang lainnya mengkodekan tujuan, menentukan apakah suatu keadaan tertentu dimenangkan atau kalah. Agar AlphaGo dapat memainkan permainan yang berbeda, seseorang harus menulis ulang semua kode C++ ini.

Terlebih lagi, jika Anda memberinya tujuan baru misalnya, mengunjungi planet ekstrasolar yang mengorbit Proxima Centauri ia akan menjelajahi miliaran urutan langkah Go dalam upaya sia-sia untuk menemukan urutan yang mencapai tujuan tersebut. Ia tidak dapat melihat ke dalam kode C++ dan menentukan hal yang jelas: tidak ada urutan langkah Go yang akan membawa Anda ke Proxima Centauri. Pengetahuan AlphaGo pada dasarnya terkunci di dalam kotak hitam.

Pada tahun 1958, dua tahun setelah pertemuan musim panasnya di Dartmouth memulai bidang kecerdasan buatan, John McCarthy mengusulkan pendekatan yang jauh lebih umum yang membuka kotak hitam: menulis program penalaran tujuan umum yang dapat menyerap pengetahuan tentang topik apa pun dan bernalar dengannya untuk menjawab pertanyaan apa pun yang dapat dijawab. Salah satu jenis penalaran tertentu adalah penalaran praktis seperti yang disarankan oleh Aristoteles: “Melakukan tindakan A, B, C, ... akan mencapai tujuan G.” Tujuannya bisa apa saja: memastikan rumah rapi sebelum saya pulang, memenangkan permainan catur tanpa kehilangan salah satu kuda Anda, mengurangi pajak saya sebesar 50 persen, mengunjungi Proxima Centauri, dan sebagainya. Kelas program baru McCarthy segera dikenal sebagai sistem berbasis pengetahuan.

Untuk mewujudkan sistem berbasis pengetahuan, diperlukan jawaban atas dua pertanyaan. Pertama, bagaimana pengetahuan dapat disimpan dalam komputer? Kedua, bagaimana komputer dapat bernalar dengan benar dengan pengetahuan itu untuk menarik kesimpulan baru? Untungnya, para filsuf Yunani kuno khususnya Aristoteles memberikan jawaban dasar atas pertanyaan-pertanyaan ini jauh sebelum munculnya komputer. Bahkan, tampaknya sangat mungkin bahwa, seandainya Aristoteles diberi akses ke komputer (dan mungkin juga listrik), ia akan menjadi peneliti AI. Jawaban Aristoteles, yang diulangi oleh McCarthy, adalah menggunakan logika formal sebagai dasar pengetahuan dan penalaran.

Ada dua jenis logika yang benar-benar penting dalam ilmu komputer. Yang pertama, disebut logika proposisional atau Boolean, dikenal oleh orang Yunani serta para filsuf Tiongkok dan India kuno. Ini adalah bahasa yang sama dari gerbang AND, gerbang NOT, dan sebagainya yang membentuk sirkuit chip komputer. Secara harfiah, CPU modern hanyalah ekspresi matematika yang sangat besar ratusan juta halaman yang ditulis dalam bahasa logika proposisional.

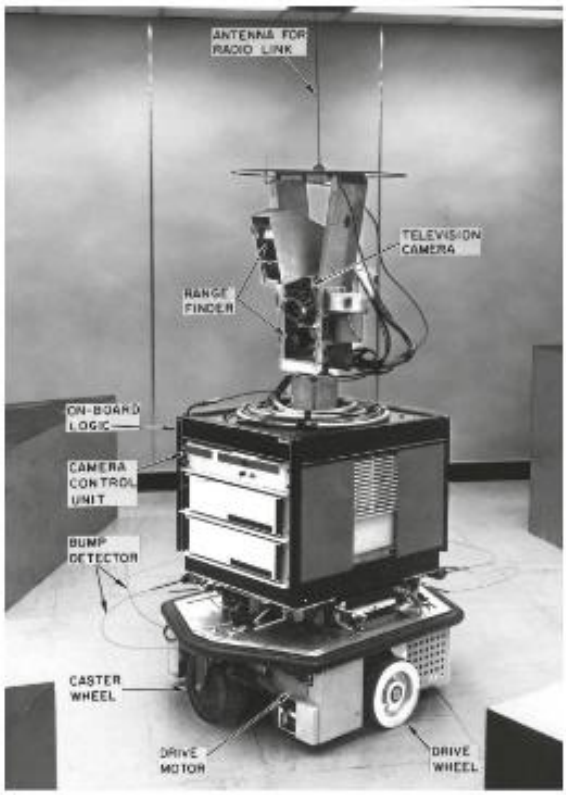
Jenis logika kedua, dan yang diusulkan McCarthy untuk digunakan pada AI, disebut logika orde pertama. Bahasa logika orde pertama jauh lebih ekspresif daripada logika proposisional, yang berarti ada hal-hal yang dapat diekspresikan dengan sangat mudah dalam logika orde pertama yang sulit atau mustahil untuk ditulis dalam logika proposisional. Misalnya, aturan permainan Go membutuhkan sekitar satu halaman dalam logika orde pertama tetapi jutaan halaman dalam logika proposisional. Demikian pula, kita dapat dengan

mudah mengekspresikan pengetahuan tentang catur, kewarganegaraan Inggris, hukum pajak, jual beli, pindah rumah, melukis, memasak, dan banyak aspek lain dari dunia akal sehat kita.

Pada prinsipnya, kemampuan untuk bernalar dengan logika orde pertama membawa kita jauh menuju kecerdasan umum. Pada tahun 1930, ahli logika Austria yang brilian, Kurt Gödel, telah menerbitkan teorema kelengkapannya yang terkenal, yang membuktikan bahwa ada algoritma dengan sifat berikut:

> Untuk setiap kumpulan pengetahuan dan setiap pertanyaan yang dapat diekspresikan dalam logika orde pertama, algoritma akan memberi tahu kita jawaban atas pertanyaan tersebut jika ada.

Ini adalah jaminan yang luar biasa. Artinya, misalnya, kita dapat memberi tahu sistem aturan permainan Go dan sistem tersebut akan memberi tahu kita (jika kita menunggu cukup lama) apakah ada langkah pembukaan yang memenangkan permainan. Kita dapat memberi tahu sistem fakta tentang geografi lokal, dan sistem akan memberi tahu kita jalan menuju bandara. Kita dapat memberi tahu sistem fakta tentang geometri, gerakan, dan peralatan makan, dan sistem akan memberi tahu robot cara menata meja untuk makan malam. Secara lebih umum, dengan mempertimbangkan tujuan yang dapat dicapai dan pengetahuan yang cukup tentang efek dari tindakannya, agen dapat menggunakan algoritma untuk menyusun rencana yang dapat dieksekusi untuk mencapai tujuan tersebut.



Gambar 2.2 Robot Shakey, sekitar tahun 1970. Di latar belakang terdapat beberapa objek yang didorong Shakey di dalam ruangan-ruangannya.

Harus dikatakan bahwa Gödel sebenarnya tidak menyediakan algoritma; ia hanya membuktikan bahwa algoritma itu ada. Pada awal tahun 1960-an, algoritma nyata untuk penalaran logis mulai muncul, dan impian McCarthy tentang sistem yang cerdas secara umum berdasarkan logika tampaknya dapat dicapai. Proyek robot bergerak utama pertama di dunia, proyek Shakey dari SRI, didasarkan pada penalaran logis. Shakey menerima tujuan dari perancang manusianya, menggunakan algoritma penglihatan untuk membuat pernyataan logis yang menggambarkan situasi saat ini, melakukan inferensi logis untuk mendapatkan rencana yang terjamin untuk mencapai tujuan, dan kemudian mengeksekusi rencana tersebut. Shakey adalah bukti "hidup" bahwa analisis Aristoteles tentang kognisi dan tindakan manusia setidaknya sebagian benar.

Sayangnya, analisis Aristoteles (dan McCarthy) jauh dari kebenaran sepenuhnya. Masalah utamanya adalah ketidaktahuan bukan, perlu saya tambahkan, dari pihak Aristoteles atau McCarthy, tetapi dari pihak semua manusia dan mesin, masa kini dan masa depan. Sangat sedikit dari pengetahuan kita yang benar-benar pasti. Secara khusus, kita tidak banyak mengetahui tentang masa depan. Ketidaktahuan hanyalah masalah yang tak teratasi bagi sistem yang sepenuhnya logis. Jika saya bertanya, “Apakah saya akan sampai di bandara tepat waktu, jika saya berangkat tiga jam sebelum penerbangan saya?” atau “Bisakah saya mendapatkan rumah dengan membeli tiket lotre yang menang dan kemudian membeli rumah tersebut dengan hasil penjualan tiket?” jawaban yang benar, dalam setiap kasus, adalah “Saya tidak tahu.”

Alasannya adalah bahwa, untuk setiap pertanyaan, baik ya maupun tidak secara logis mungkin. Secara praktis, seseorang tidak akan pernah bisa benar-benar yakin tentang pertanyaan empiris apa pun kecuali jawabannya sudah diketahui. Untungnya, kepastian sama sekali tidak diperlukan untuk tindakan: kita hanya perlu mengetahui tindakan mana yang terbaik, bukan tindakan mana yang pasti akan berhasil.

Ketidakpastian berarti bahwa “tujuan yang dimasukkan ke dalam mesin” pada umumnya tidak dapat berupa tujuan yang didefinisikan secara tepat, yang harus dicapai dengan segala cara. Tidak ada lagi yang namanya “rangkaian tindakan yang mencapai tujuan,” karena setiap rangkaian tindakan akan memiliki banyak kemungkinan hasil, beberapa di antaranya tidak akan mencapai tujuan.

Kemungkinan keberhasilan sangat penting: berangkat ke bandara tiga jam sebelum penerbangan Anda mungkin berarti Anda tidak akan ketinggalan penerbangan dan membeli tiket lotere mungkin berarti Anda akan memenangkan cukup uang untuk membeli rumah baru, tetapi ini adalah kemungkinan yang sangat berbeda. Tujuan tidak dapat diselamatkan dengan mencari rencana yang memaksimalkan probabilitas pencapaian tujuan. Rencana yang memaksimalkan probabilitas sampai ke bandara tepat waktu untuk mengejar penerbangan mungkin melibatkan meninggalkan rumah beberapa hari sebelumnya, mengatur pengawal bersenjata, menyiapkan banyak alat transportasi alternatif jika yang lain rusak, dan sebagainya. Mau tidak mau, seseorang harus mempertimbangkan keinginan relatif dari berbagai hasil serta kemungkinannya.

Alih-alih tujuan, kita dapat menggunakan fungsi utilitas untuk menggambarkan keinginan dari berbagai hasil atau urutan keadaan. Seringkali, utilitas dari suatu urutan keadaan dinyatakan sebagai jumlah imbalan untuk setiap keadaan dalam urutan tersebut. Dengan tujuan yang didefinisikan oleh fungsi utilitas atau imbalan, mesin bertujuan untuk menghasilkan perilaku yang memaksimalkan utilitas yang diharapkan atau jumlah imbalan yang diharapkan, dirata-ratakan atas kemungkinan hasil yang diberi bobot berdasarkan probabilitasnya. AI modern sebagian merupakan pengulangan mimpi McCarthy, kecuali dengan utilitas dan probabilitas sebagai pengganti tujuan dan logika.

Pierre-Simon Laplace, matematikawan besar Prancis, menulis pada tahun 1814, “Teori probabilitas hanyalah akal sehat yang direduksi menjadi kalkulus.” Namun, baru pada tahun 1980-an, bahasa formal praktis dan algoritma penalaran dikembangkan untuk pengetahuan probabilistik. Ini adalah bahasa jaringan Bayesian, yang diperkenalkan oleh Judea Pearl. Secara kasar, jaringan Bayesian adalah sepupu probabilistik dari logika proposisional. Terdapat pula sepupu probabilistik dari logika orde pertama, termasuk logika Bayesian dan berbagai macam bahasa pemrograman probabilistik.

Jaringan Bayesian dan logika Bayesian dinamai menurut Pendeta Thomas Bayes, seorang pendeta Inggris yang kontribusinya yang abadi terhadap pemikiran modern yang sekarang dikenal sebagai teorema Bayes diterbitkan pada tahun 1763, tak lama setelah kematiannya, oleh temannya Richard Price. Dalam bentuk modernnya, seperti yang disarankan oleh Laplace, teorema tersebut menjelaskan dengan cara yang sangat sederhana bagaimana probabilitas prior tingkat kepercayaan awal seseorang terhadap serangkaian hipotesis yang mungkin menjadi probabilitas posterior sebagai hasil dari pengamatan beberapa bukti.

Seiring dengan datangnya bukti baru, posterior menjadi prior baru dan proses pembaruan Bayesian berulang tanpa batas. Proses ini sangat mendasar sehingga gagasan modern tentang rasionalitas sebagai maksimalisasi utilitas yang diharapkan kadang-kadang disebut rasionalitas Bayesian. Asumsi ini menyatakan bahwa agen rasional memiliki akses ke distribusi probabilitas posterior atas kemungkinan keadaan dunia saat ini, serta atas hipotesis tentang masa depan, berdasarkan semua pengalaman masa lalunya.

Para peneliti di bidang riset operasi, teori kontrol, dan AI juga telah mengembangkan berbagai algoritma untuk pengambilan keputusan dalam kondisi ketidakpastian, beberapa di antaranya berasal dari tahun 1950-an. Algoritma yang disebut "pemrograman dinamis" ini adalah sepupu probabilistik dari pencarian dan perencanaan ke depan dan dapat menghasilkan perilaku optimal atau mendekati optimal untuk semua jenis masalah praktis di bidang keuangan, logistik, transportasi, dan sebagainya, di mana ketidakpastian memainkan peran penting.

Tujuan dimasukkan ke dalam mesin-mesin ini dalam bentuk fungsi imbalan, dan outputnya adalah kebijakan yang menentukan tindakan untuk setiap kemungkinan keadaan yang dapat dialami agen. Untuk masalah kompleks seperti backgammon dan Go, di mana jumlah keadaannya sangat besar dan imbalan hanya datang di akhir permainan, pencarian ke depan tidak akan berhasil. Sebagai gantinya, para peneliti AI telah mengembangkan metode yang disebut pembelajaran penguatan, atau RL. Algoritma RL belajar dari pengalaman

langsung sinyal imbalan di lingkungan, seperti halnya bayi belajar berdiri dari imbalan positif berupa posisi tegak dan imbalan negatif berupa jatuh. Seperti halnya algoritma pemrograman dinamis, tujuan yang dimasukkan ke dalam algoritma RL adalah fungsi imbalan, dan algoritma tersebut mempelajari estimator untuk nilai keadaan (atau terkadang nilai tindakan). Estimator ini dapat dikombinasikan dengan pencarian pandangan ke depan yang relatif miopik untuk menghasilkan perilaku yang sangat kompeten.

Sistem pembelajaran penguatan pertama yang berhasil adalah program catur Arthur Samuel, yang menciptakan sensasi ketika didemonstrasikan di televisi pada tahun 1956. Program tersebut pada dasarnya belajar dari awal, dengan bermain melawan dirinya sendiri dan mengamati imbalan dari menang dan kalah. Pada tahun 1992, Gerry Tesauro menerapkan ide yang sama pada permainan backgammon, mencapai permainan tingkat juara dunia setelah 1.500.000 permainan. Mulai tahun 2016, AlphaGo DeepMind dan keturunannya menggunakan pembelajaran penguatan dan permainan mandiri untuk mengalahkan pemain manusia terbaik di Go, catur, dan shogi.

Algoritma pembelajaran penguatan juga dapat mempelajari cara memilih tindakan berdasarkan masukan persepsi mentah. Sebagai contoh, sistem DQN DeepMind belajar memainkan empat puluh sembilan gim video Atari yang berbeda sepenuhnya dari awal termasuk Pong, Freeway, dan Space Invaders. Sistem ini hanya menggunakan piksel layar sebagai input dan skor gim sebagai sinyal penghargaan. Dalam sebagian besar gim, DQN belajar bermain lebih baik daripada pemain manusia profesional meskipun DQN tidak memiliki gagasan a priori tentang waktu, ruang, objek, gerakan, kecepatan, atau penembakan. Sangat sulit untuk mengetahui apa yang sebenarnya dilakukan DQN, selain menang.

Jika bayi yang baru lahir belajar memainkan lusinan gim video pada tingkat superhuman pada hari pertama kehidupannya, atau menjadi juara dunia dalam Go, catur, dan shogi, kita mungkin mencurigai kerasukan setan atau campur tangan alien. Namun, ingatlah bahwa semua tugas ini jauh lebih sederhana daripada dunia nyata: tugas-tugas ini sepenuhnya dapat diamati, melibatkan cakupan waktu yang singkat, dan memiliki ruang keadaan yang relatif kecil dan aturan yang sederhana dan dapat diprediksi. Melonggarkan salah satu kondisi ini berarti bahwa metode standar akan gagal.

Di sisi lain, penelitian terkini justru bertujuan untuk melampaui metode standar sehingga sistem AI dapat beroperasi di lingkungan yang lebih luas. Misalnya, pada hari saya menulis paragraf sebelumnya, OpenAI mengumumkan bahwa timnya yang terdiri dari lima program AI telah belajar mengalahkan tim manusia berpengalaman dalam permainan Dota 2. (Bagi yang belum tahu, termasuk saya: Dota 2 adalah versi terbaru dari Defense of the Ancients, sebuah permainan strategi waktu nyata dalam keluarga Warcraft; saat ini merupakan e-sport yang paling menguntungkan dan kompetitif, dengan hadiah jutaan rupiah.) Dota 2 melibatkan komunikasi, kerja tim, dan ruang dan waktu yang hampir kontinu.

Permainan berlangsung selama puluhan ribu langkah waktu, dan beberapa tingkat organisasi perilaku hierarkis tampaknya sangat penting. Bill Gates menggambarkan pengumuman itu sebagai “tonggak penting dalam memajukan kecerdasan buatan.” Beberapa

bulan kemudian, versi terbaru dari program tersebut mengalahkan tim Dota 2 profesional terbaik dunia.

Permainan seperti Go dan Dota 2 merupakan lahan uji yang baik untuk metode pembelajaran penguatan karena fungsi hadiahnya sudah termasuk dalam aturan permainan. Namun, dunia nyata kurang nyaman, dan ada puluhan kasus di mana definisi hadiah yang salah menyebabkan perilaku aneh dan tak terduga. Beberapa di antaranya tidak berbahaya, seperti sistem evolusi simulasi yang seharusnya mengembangkan makhluk yang bergerak cepat tetapi pada kenyataannya menghasilkan makhluk yang sangat tinggi dan bergerak cepat dengan cara jatuh. Yang lain kurang berbahaya, seperti pengoptimal klik-tayang media sosial yang tampaknya membuat kekacauan di dunia kita.

Kategori terakhir dari program agen yang akan saya pertimbangkan adalah yang paling sederhana: program yang menghubungkan persepsi langsung dengan tindakan, tanpa pertimbangan atau penalaran perantara. Dalam AI, kita menyebut jenis program ini sebagai agen refleks sebuah referensi untuk refleks saraf tingkat rendah yang ditunjukkan oleh manusia dan hewan, yang tidak dimediasi oleh pikiran. Misalnya, refleks berkedip manusia menghubungkan keluaran sirkuit pemrosesan tingkat rendah dalam sistem visual langsung ke area motorik yang mengontrol kelopak mata, sehingga setiap wilayah yang muncul dengan cepat di bidang visual menyebabkan kedipan yang kuat. Anda dapat mengujinya sekarang dengan mencoba (jangan terlalu keras) untuk menusuk mata Anda sendiri dengan jari Anda. Kita dapat menganggap sistem refleks ini sebagai "aturan" sederhana dalam bentuk berikut:

jika <wilayah yang muncul dengan cepat di bidang visual> maka <kedip>.

Refleks berkedip tidak "mengetahui apa yang dilakukannya": tujuannya (untuk melindungi bola mata dari benda asing) tidak terwakili di mana pun; pengetahuan (bahwa area yang mendekat dengan cepat sesuai dengan objek yang mendekati mata, dan bahwa objek yang mendekati mata dapat merusaknya) tidak terwakili di mana pun. Dengan demikian, ketika bagian non-refleks dari diri Anda ingin meneteskan obat tetes mata, bagian refleks tetap berkedip.

Refleks familiar lainnya adalah pengereman darurat ketika mobil di depan berhenti tiba-tiba atau seorang pejalan kaki melangkah ke jalan. Memutuskan dengan cepat apakah pengereman diperlukan bukanlah hal yang mudah: ketika sebuah kendaraan uji dalam mode otonom menewaskan seorang pejalan kaki pada tahun 2018, Uber menjelaskan bahwa "manuver pengereman darurat tidak diaktifkan saat kendaraan berada di bawah kendali komputer, untuk mengurangi potensi perilaku kendaraan yang tidak menentu." Di sini, tujuan perancang manusia jelas jangan membunuh pejalan kaki tetapi kebijakan agen (seandainya diaktifkan) menerapkannya secara tidak benar. Sekali lagi, tujuannya tidak terwakili dalam agen: tidak ada kendaraan otonom saat ini yang tahu bahwa orang tidak suka dibunuh.

Tindakan refleks juga berperan dalam tugas-tugas yang lebih rutin seperti tetap berada di jalur: ketika mobil sedikit melenceng dari posisi jalur ideal, sistem kontrol umpan balik sederhana dapat mendorong roda kemudi ke arah yang berlawanan untuk mengoreksi

penyimpangan tersebut. Besarnya dorongan akan bergantung pada seberapa jauh mobil melenceng. Sistem kontrol semacam ini biasanya dirancang untuk meminimalkan kuadrat kesalahan pelacakan yang diakumulasikan dari waktu ke waktu. Perancang menurunkan hukum kontrol umpan balik yang, di bawah asumsi tertentu tentang kecepatan dan kelengkungan jalan, secara kasar mengimplementasikan minimisasi ini. Sistem serupa beroperasi sepanjang waktu saat Anda berdiri; jika sistem tersebut berhenti bekerja, Anda akan jatuh dalam beberapa detik. Seperti halnya refleks berkedip, cukup sulit untuk mematikan mekanisme ini dan membiarkan diri Anda jatuh.

Agen refleks, kemudian, mengimplementasikan tujuan perancang, tetapi tidak mengetahui apa tujuannya atau mengapa mereka bertindak dengan cara tertentu. Ini berarti mereka tidak dapat benar-benar membuat keputusan sendiri; orang lain, biasanya perancang manusia atau mungkin proses evolusi biologis, harus memutuskan semuanya terlebih dahulu. Sangat sulit untuk menciptakan agen refleks yang baik melalui pemrograman manual kecuali untuk tugas-tugas yang sangat sederhana seperti tic-tac-toe atau pengereman darurat. Bahkan dalam kasus-kasus tersebut, agen refleks sangat tidak fleksibel dan tidak dapat mengubah perilakunya ketika keadaan menunjukkan bahwa kebijakan yang diterapkan tidak lagi tepat.

Salah satu cara yang mungkin untuk menciptakan agen refleks yang lebih kuat adalah melalui proses pembelajaran dari contoh. Daripada menentukan aturan tentang bagaimana berperilaku, atau memberikan fungsi penghargaan atau tujuan, manusia dapat memberikan contoh masalah pengambilan keputusan beserta keputusan yang benar untuk dibuat dalam setiap kasus. Misalnya, kita dapat membuat agen penerjemahan bahasa Prancis ke bahasa Inggris dengan memberikan contoh kalimat bahasa Prancis beserta terjemahan bahasa Inggris yang benar. (Untungnya, parlemen Kanada dan Uni Eropa menghasilkan jutaan contoh seperti itu setiap tahun.)

Kemudian algoritma pembelajaran terawasi memproses contoh-contoh tersebut untuk menghasilkan aturan kompleks yang mengambil kalimat bahasa Prancis apa pun sebagai masukan dan menghasilkan terjemahan bahasa Inggris. Algoritma pembelajaran mesin yang saat ini menjadi andalan adalah bentuk yang disebut pembelajaran mendalam (*deep learning*), dan menghasilkan aturan dalam bentuk jaringan saraf tiruan dengan ratusan lapisan dan jutaan parameter.

Algoritma pembelajaran mendalam lainnya terbukti sangat baik dalam mengklasifikasikan objek dalam gambar dan mengenali kata-kata dalam sinyal suara. Penerjemahan mesin, pengenalan suara, dan pengenalan objek visual adalah tiga subbidang terpenting dalam AI, itulah sebabnya prospek pembelajaran mendalam sangat menarik.

Kita dapat berdebat tanpa henti tentang apakah pembelajaran mendalam akan langsung mengarah pada AI tingkat manusia. Pandangan saya sendiri, yang akan saya jelaskan nanti, adalah bahwa hal itu masih jauh dari apa yang dibutuhkan, tetapi untuk saat ini mari kita fokus pada bagaimana metode tersebut sesuai dengan model standar AI, di mana algoritma mengoptimalkan tujuan tetap. Untuk pembelajaran mendalam, atau bahkan untuk algoritma pembelajaran terawasi apa pun, "tujuan yang dimasukkan ke dalam mesin" biasanya untuk memaksimalkan akurasi prediksi atau, secara ekuivalen, untuk meminimalkan

kesalahan. Hal itu tampaknya jelas, tetapi sebenarnya ada dua cara untuk memahaminya, tergantung pada peran yang akan dimainkan oleh aturan yang dipelajari dalam sistem secara keseluruhan. Peran pertama adalah peran yang murni perseptual: jaringan memproses input sensorik dan memberikan informasi ke bagian sistem lainnya dalam bentuk perkiraan probabilitas untuk apa yang dipersepsikannya.

Jika itu adalah algoritma pengenalan objek, mungkin ia mengatakan "70 persen kemungkinan itu adalah anjing terrier Norfolk, 30 persen itu adalah anjing terrier Norwich." Bagian sistem lainnya memutuskan tindakan eksternal yang akan diambil berdasarkan informasi ini. Tujuan yang murni perseptual ini tidak bermasalah dalam arti berikut: bahkan sistem AI supercerdas yang "aman", berbeda dengan sistem yang "tidak aman" berdasarkan model standar, perlu memiliki sistem persepsi yang seakurat dan terkalibrasi sebaik mungkin.

Masalah muncul ketika kita beralih dari peran yang murni perseptual ke peran pengambilan keputusan. Misalnya, jaringan terlatih untuk mengenali objek mungkin secara otomatis menghasilkan label untuk gambar di situs web atau akun media sosial. Memposting label tersebut adalah tindakan dengan konsekuensi. Setiap tindakan pelabelan membutuhkan keputusan klasifikasi yang sebenarnya, dan kecuali setiap keputusan dijamin sempurna, perancang manusia harus menyediakan fungsi kerugian yang menjelaskan biaya salah mengklasifikasikan objek tipe A sebagai objek tipe B. Dan itulah mengapa Google mengalami masalah yang tidak menguntungkan dengan gorila.

Pada tahun 2015, seorang insinyur perangkat lunak bernama Jacky Alciné mengeluh di Twitter bahwa layanan pelabelan gambar Google Photos telah melabeli dirinya dan temannya sebagai gorila. Meskipun tidak jelas bagaimana tepatnya kesalahan ini terjadi, hampir pasti bahwa algoritma pembelajaran mesin Google dirancang untuk meminimalkan fungsi kerugian tetap dan pasti lebih jauh lagi, fungsi yang menetapkan biaya yang sama untuk setiap kesalahan. Dengan kata lain, algoritma tersebut mengasumsikan bahwa biaya salah mengklasifikasikan seseorang sebagai gorila sama dengan biaya salah mengklasifikasikan anjing Norfolk terrier sebagai anjing Norwich terrier. Jelas, ini bukanlah fungsi kerugian sebenarnya Google (atau penggunanya), seperti yang diilustrasikan oleh bencana hubungan masyarakat yang terjadi kemudian.

Karena ada ribuan kemungkinan label gambar, ada jutaan biaya potensial yang berbeda yang terkait dengan salah mengklasifikasikan satu kategori sebagai kategori lain. Bahkan jika Google mencoba, mereka akan kesulitan untuk menentukan semua angka ini di awal. Sebaliknya, hal yang benar untuk dilakukan adalah mengakui ketidakpastian tentang biaya kesalahan klasifikasi yang sebenarnya dan merancang algoritma pembelajaran dan klasifikasi yang cukup sensitif terhadap biaya dan ketidakpastian tentang biaya tersebut.

Algoritma seperti itu mungkin sesekali mengajukan pertanyaan kepada perancang Google seperti “Mana yang lebih buruk, salah mengklasifikasikan anjing sebagai kucing atau salah mengklasifikasikan seseorang sebagai hewan?” Selain itu, jika ada ketidakpastian yang signifikan tentang biaya kesalahan klasifikasi, algoritma tersebut mungkin akan menolak untuk memberi label pada beberapa gambar.

Pada awal tahun 2018, dilaporkan bahwa Google Photos menolak untuk mengklasifikasikan foto gorila. Diberikan gambar gorila yang sangat jelas dengan dua bayi, ia mengatakan, “Hmm… belum melihat ini dengan jelas.” Saya tidak ingin menyarankan bahwa adopsi model standar oleh AI adalah pilihan yang buruk pada saat itu. Banyak pekerjaan brilian telah dilakukan untuk mengembangkan berbagai instansiasi model dalam sistem logis, probabilistik, dan pembelajaran. Banyak dari sistem yang dihasilkan sangat berguna; seperti yang akan kita lihat di bab berikutnya, masih banyak lagi yang akan datang. Di sisi lain, kita tidak bisa terus mengandalkan praktik biasa kita dalam memperbaiki kesalahan besar dalam fungsi objektif melalui metode coba-coba: mesin dengan kecerdasan yang semakin tinggi dan dampak global yang semakin besar tidak akan memungkinkan kita untuk menikmati kemewahan itu.

BAB 3
KEMAJUAN AI DI MASA DEPAN

3.1 MASA DEPAN YANG DEKAT

Pada tanggal 3 Mei 1997, sebuah pertandingan catur dimulai antara Deep Blue, komputer catur yang dibangun oleh IBM, dan Garry Kasparov, juara catur dunia dan mungkin pemain manusia terbaik dalam sejarah. *Newsweek* menyebut pertandingan itu sebagai "Pertahanan Terakhir Otak." Pada tanggal 11 Mei, dengan skor imbang 2½–2½, Deep Blue mengalahkan Kasparov di game terakhir. Media menjadi heboh. Kapitalisasi pasar IBM meningkat sebesar Rp 180 triliun dalam semalam. AI, menurut semua pihak, telah mencapai terobosan besar.

Dari sudut pandang penelitian AI, pertandingan tersebut sama sekali tidak mewakili terobosan. Kemenangan Deep Blue, meskipun mengesankan, hanya melanjutkan tren yang telah terlihat selama beberapa dekade. Desain dasar untuk algoritma bermain catur diletakkan pada tahun 1950 oleh Claude Shannon, dengan peningkatan besar pada awal tahun 1960-an. Setelah itu, peringkat catur dari program-program terbaik terus meningkat, terutama sebagai hasil dari komputer yang lebih cepat yang memungkinkan program untuk melihat lebih jauh ke depan. Pada tahun 1994, Peter Norvig dan saya memetakan peringkat numerik dari program catur terbaik dari tahun 1965 dan seterusnya, pada skala di mana peringkat Kasparov adalah 2805. Peringkat dimulai dari 1400 pada tahun 1965 dan meningkat dalam garis lurus yang hampir sempurna selama tiga puluh tahun.

Dengan mengekstrapolasi garis tersebut ke depan dari tahun 1994, dapat diprediksi bahwa komputer akan mampu mengalahkan Kasparov pada tahun 1997 tepat ketika itu terjadi. Bagi para peneliti AI, terobosan nyata terjadi tiga puluh atau empat puluh tahun sebelum Deep Blue muncul di kesadaran publik. Demikian pula, jaringan konvolusional dalam (*deep convolutional networks*) telah ada, dengan semua matematika yang telah sepenuhnya dikembangkan, lebih dari dua puluh tahun sebelum mulai menjadi berita utama.

Pandangan publik tentang terobosan AI yang didapatkan dari media kemenangan menakjubkan atas manusia, robot yang menjadi warga negara Arab Saudi, dan sebagainya hampir tidak ada hubungannya dengan apa yang sebenarnya terjadi di laboratorium penelitian dunia. Di dalam laboratorium, penelitian melibatkan banyak pemikiran, diskusi, dan penulisan rumus matematika di papan tulis. Ide terus-menerus dihasilkan, ditinggalkan, dan ditemukan kembali. Sebuah ide bagus sebuah terobosan nyata seringkali tidak disadari pada saat itu dan mungkin baru kemudian dipahami sebagai dasar kemajuan substansial dalam AI, mungkin ketika seseorang menemukannya kembali pada waktu yang lebih tepat. Ide-ide diuji coba, awalnya pada masalah sederhana untuk menunjukkan bahwa intuisi dasarnya benar dan kemudian pada masalah yang lebih sulit untuk melihat seberapa baik skalanya.

Seringkali, sebuah ide akan gagal dengan sendirinya untuk memberikan peningkatan substansial dalam kemampuan, dan harus menunggu ide lain muncul sehingga kombinasi

keduanya dapat menunjukkan nilai. Semua aktivitas ini sama sekali tidak terlihat dari luar. Di dunia di luar laboratorium, AI hanya terlihat ketika akumulasi ide secara bertahap dan bukti validitasnya melampaui ambang batas: titik di mana investasi uang dan upaya rekayasa untuk menciptakan produk komersial baru atau demonstrasi yang mengesankan menjadi layak. Kemudian media mengumumkan bahwa terobosan telah terjadi.

Oleh karena itu, dapat diharapkan bahwa banyak ide lain yang telah berkembang di laboratorium penelitian dunia akan melampaui ambang batas penerapan komersial dalam beberapa tahun ke depan. Ini akan terjadi semakin sering seiring dengan meningkatnya tingkat investasi komersial dan seiring dengan semakin terbukanya dunia terhadap aplikasi AI. Bab ini memberikan contoh apa yang dapat kita lihat di masa mendatang.

Sepanjang jalan, saya akan menyebutkan beberapa kekurangan dari kemajuan teknologi ini. Anda mungkin dapat memikirkan lebih banyak lagi, tetapi jangan khawatir. Saya akan membahasnya di bab berikutnya.

Ekosistem AI

Pada awalnya, lingkungan tempat sebagian besar komputer beroperasi pada dasarnya tidak berbentuk dan hampa: satu-satunya input mereka berasal dari kartu berlubang dan satu-satunya metode output mereka adalah mencetak karakter pada printer baris. Mungkin karena alasan ini, sebagian besar peneliti memandang mesin cerdas sebagai penjawab pertanyaan; pandangan mesin sebagai agen yang memahami dan bertindak dalam suatu lingkungan tidak meluas hingga tahun 1980-an.

Munculnya World Wide Web pada tahun 1990-an membuka alam semesta baru bagi mesin cerdas untuk berperan. Sebuah kata baru, softbot, diciptakan untuk menggambarkan "robot" perangkat lunak yang beroperasi sepenuhnya dalam lingkungan perangkat lunak seperti Web. Softbot, atau bot seperti yang kemudian dikenal, memahami halaman Web dan bertindak dengan memancarkan rangkaian karakter, URL, dan sebagainya.

Perusahaan AI berkembang pesat selama booming dot-com (1997–2000), menyediakan kemampuan inti untuk pencarian dan e-commerce, termasuk analisis tautan, sistem rekomendasi, sistem reputasi, perbandingan belanja, dan kategorisasi produk. Pada awal tahun 2000-an, adopsi luas telepon seluler dengan mikrofon, kamera, akselerometer, dan GPS memberikan akses baru bagi sistem AI ke kehidupan sehari-hari masyarakat; "speaker pintar" seperti Amazon Echo, Google Home, dan Apple HomePod telah melengkapi proses ini.

Sekitar tahun 2008, jumlah objek yang terhubung ke Internet melebihi jumlah orang yang terhubung ke Internet, sebuah transisi yang oleh sebagian orang disebut sebagai awal dari *Internet of Things* (IoT). Benda-benda tersebut termasuk mobil, peralatan rumah tangga, lampu lalu lintas, mesin penjual otomatis, termostat, quadcopter, kamera, sensor lingkungan, robot, dan semua jenis barang material baik dalam proses manufaktur maupun dalam sistem distribusi dan ritel. Hal ini memberikan sistem AI akses sensorik dan kontrol yang jauh lebih besar ke dunia nyata.

Terakhir, peningkatan dalam persepsi telah memungkinkan robot bertenaga AI untuk keluar dari pabrik, di mana mereka bergantung pada pengaturan objek yang sangat terbatas,

dan masuk ke dunia nyata yang tidak terstruktur dan berantakan, di mana kamera mereka memiliki sesuatu yang menarik untuk dilihat.

Mobil Tanpa Pengemudi

Pada akhir tahun 1950-an, John McCarthy membayangkan bahwa suatu hari nanti kendaraan otomatis dapat membawanya ke bandara. Pada tahun 1987, Ernst Dickmanns mendemonstrasikan sebuah van Mercedes tanpa pengemudi di jalan tol Jerman; van tersebut mampu tetap berada di jalurnya, mengikuti mobil lain, berpindah jalur, dan menyalip. Lebih dari tiga puluh tahun kemudian, kita masih belum memiliki mobil yang sepenuhnya otonom, tetapi semakin mendekat. Fokus pengembangan telah lama beralih dari laboratorium penelitian akademis ke perusahaan besar. Hingga tahun 2019, kendaraan uji dengan kinerja terbaik telah menempuh jutaan mil di jalan umum (dan miliaran mil di simulator mengemudi) tanpa insiden serius. Sayangnya, kendaraan otonom dan semi-otonom lainnya telah menewaskan beberapa orang. Mengapa butuh waktu begitu lama untuk mencapai pengemudian otonom yang aman? Alasan pertama adalah persyaratan kinerja yang sangat ketat. Pengemudi manusia di Amerika Serikat mengalami sekitar satu kecelakaan fatal per seratus juta mil perjalanan, yang menetapkan standar yang tinggi.

Agar dapat diterima, kendaraan otonom perlu jauh lebih baik dari itu: mungkin satu kecelakaan fatal per miliar mil, atau dua puluh lima ribu tahun mengemudi selama empat puluh jam per minggu. Alasan kedua adalah bahwa solusi yang diantisipasi menyerahkan kendali kepada manusia ketika kendaraan bingung atau berada di luar kondisi operasi yang aman sama sekali tidak berhasil. Ketika mobil mengemudi sendiri, manusia dengan cepat terlepas dari keadaan mengemudi saat itu dan tidak dapat memperoleh kembali konteks dengan cukup cepat untuk mengambil alih kendali dengan aman. Selain itu, orang yang bukan pengemudi dan penumpang taksi yang berada di kursi belakang tidak berada dalam posisi untuk mengemudikan mobil jika terjadi sesuatu yang salah.

Proyek-proyek saat ini bertujuan untuk mencapai otonomi SAE Level 4, yang berarti bahwa kendaraan harus selalu mampu mengemudi secara otonom atau berhenti dengan aman, dengan mempertimbangkan batasan geografis dan kondisi cuaca. Karena kondisi cuaca dan lalu lintas dapat berubah, dan karena keadaan yang tidak biasa dapat muncul yang tidak dapat ditangani oleh kendaraan Level 4, manusia harus berada di dalam kendaraan dan siap untuk mengambil alih jika diperlukan. (Level 5 otonomi tanpa batasan tidak memerlukan pengemudi manusia tetapi bahkan lebih sulit untuk dicapai.) Otonomi Level 4 jauh melampaui tugas-tugas refleks sederhana seperti mengikuti garis putih dan menghindari rintangan. Kendaraan harus menilai maksud dan kemungkinan lintasan masa depan dari semua objek yang relevan, termasuk objek yang mungkin tidak terlihat, berdasarkan pengamatan saat ini dan masa lalu.

Kemudian, menggunakan pencarian ke depan, kendaraan harus menemukan lintasan yang mengoptimalkan beberapa kombinasi keselamatan dan kemajuan. Beberapa proyek mencoba pendekatan yang lebih langsung berdasarkan pembelajaran penguatan (terutama dalam simulasi, tentu saja) dan pembelajaran terawasi dari rekaman ratusan pengemudi

manusia, tetapi pendekatan ini tampaknya tidak mungkin mencapai tingkat keselamatan yang dibutuhkan.

Potensi manfaat kendaraan otonom sepenuhnya sangat besar. Setiap tahun, 1,2 juta orang meninggal dalam kecelakaan mobil di seluruh dunia dan puluhan juta menderita cedera serius. Target yang wajar untuk kendaraan otonom adalah mengurangi angka-angka ini hingga sepuluh kali lipat. Beberapa analisis juga memprediksi pengurangan besar dalam biaya transportasi, struktur parkir, kemacetan, dan polusi. Kota-kota akan beralih dari mobil pribadi dan bus besar ke kendaraan listrik otonom yang tersebar luas, menyediakan layanan antar-jemput dan mendukung koneksi transportasi massal berkecepatan tinggi antar pusat. Dengan biaya serendah tiga sen per mil penumpang, sebagian besar kota mungkin akan memilih untuk menyediakan layanan ini secara gratis sambil membuat penumpang terpapar iklan yang tak ada habisnya.

Tentu saja, untuk menuai semua manfaat ini, industri harus memperhatikan risikonya. Jika terlalu banyak kematian yang disebabkan oleh kendaraan eksperimental yang dirancang dengan buruk, regulator dapat menghentikan penyebaran yang direncanakan atau memberlakukan standar yang sangat ketat yang mungkin tidak dapat dicapai selama beberapa dekade. Dan orang-orang mungkin, tentu saja, memutuskan untuk tidak membeli atau menaiki kendaraan otonom kecuali terbukti aman.

Jajak pendapat tahun 2018 mengungkapkan penurunan signifikan dalam tingkat kepercayaan konsumen terhadap teknologi kendaraan otonom dibandingkan dengan tahun 2016. Bahkan jika teknologi tersebut berhasil, transisi menuju otonomi yang meluas akan menjadi hal yang canggung: keterampilan mengemudi manusia mungkin akan menurun atau hilang, dan tindakan mengemudi mobil sendiri yang sembrono dan antisosial mungkin akan dilarang sama sekali.

Asisten Pribadi yang Cerdas

Sebagian besar pembaca mungkin sudah pernah mengalami asisten pribadi yang tidak cerdas: speaker pintar yang menuruti perintah pembelian yang terdengar di televisi, atau chatbot ponsel yang merespons "Panggilkan ambulans!" dengan "*Oke, mulai sekarang saya akan memanggil Anda 'Ann Ambulans*'." Sistem seperti itu pada dasarnya adalah antarmuka yang dimediasi suara untuk aplikasi dan mesin pencari; sistem ini sebagian besar didasarkan pada templat stimulus-respons yang sudah jadi, sebuah pendekatan yang berasal dari sistem Eliza pada pertengahan tahun 1960-an.

Sistem-sistem awal ini memiliki tiga kekurangan: akses, konten, dan konteks. Kekurangan akses berarti mereka kurang memiliki kesadaran sensorik tentang apa yang sedang terjadi misalnya, mereka mungkin dapat mendengar apa yang dikatakan pengguna tetapi mereka tidak dapat melihat dengan siapa pengguna berbicara. Kekurangan konten berarti mereka gagal memahami arti dari apa yang dikatakan atau dikirim pengguna, bahkan jika mereka memiliki akses ke sana. Kekurangan konteks berarti mereka kurang memiliki kemampuan untuk melacak dan menalar tentang tujuan, aktivitas, dan hubungan yang membentuk kehidupan sehari-hari.

Terlepas dari kekurangan ini, speaker pintar dan asisten ponsel menawarkan nilai yang cukup bagi pengguna sehingga telah memasuki rumah dan saku ratusan juta orang. Mereka, dalam arti tertentu, adalah kuda Troya untuk AI. Karena mereka ada di sana, tertanam dalam begitu banyak kehidupan, setiap peningkatan kecil dalam kemampuan mereka bernilai miliaran dolar.

Oleh karena itu, peningkatan terus terjadi dengan cepat. Mungkin yang terpenting adalah kemampuan dasar untuk memahami konten untuk mengetahui bahwa "John di rumah sakit" bukan hanya perintah untuk mengatakan "Saya harap itu bukan sesuatu yang serius" tetapi berisi informasi aktual bahwa putra pengguna yang berusia delapan tahun berada di rumah sakit terdekat dan mungkin mengalami cedera atau penyakit serius. Kemampuan untuk mengakses email dan komunikasi teks serta panggilan telepon dan percakapan domestik (melalui speaker pintar di rumah) akan memberi sistem AI informasi yang cukup untuk membangun gambaran yang cukup lengkap tentang kehidupan pengguna mungkin bahkan lebih banyak informasi daripada yang mungkin tersedia bagi kepala pelayan yang bekerja untuk keluarga bangsawan abad ke-19 atau asisten eksekutif yang bekerja untuk CEO modern.

Tentu saja, informasi mentah saja tidak cukup. Agar benar-benar bermanfaat, seorang asisten juga membutuhkan pengetahuan akal sehat tentang bagaimana dunia bekerja: bahwa seorang anak di rumah sakit tidak sekaligus berada di rumah; bahwa perawatan rumah sakit untuk patah lengan jarang berlangsung lebih dari satu atau dua hari; bahwa sekolah anak tersebut perlu mengetahui tentang ketidakhadiran yang diharapkan; dan seterusnya. Pengetahuan semacam itu memungkinkan asisten untuk melacak hal-hal yang tidak diamatinya secara langsung, sebuah keterampilan penting bagi sistem cerdas.

Kemampuan yang dijelaskan dalam paragraf sebelumnya, menurut saya, dapat diwujudkan dengan teknologi yang ada untuk penalaran probabilistik, tetapi ini akan membutuhkan upaya yang sangat besar untuk membangun model dari semua jenis peristiwa dan transaksi yang membentuk kehidupan kita sehari-hari. Hingga saat ini, proyek pemodelan akal sehat semacam ini umumnya belum dilakukan (kecuali mungkin dalam sistem rahasia untuk analisis intelijen dan perencanaan militer) karena biaya yang terlibat dan imbalan yang tidak pasti. Namun, sekarang, proyek seperti ini dapat dengan mudah menjangkau ratusan juta pengguna, sehingga risiko investasi lebih rendah dan potensi imbalannya jauh lebih tinggi. Selain itu, akses ke sejumlah besar pengguna memungkinkan asisten cerdas untuk belajar dengan sangat cepat dan mengisi semua celah dalam pengetahuannya.

Dengan demikian, kita dapat mengharapkan adanya asisten cerdas yang, dengan biaya beberapa sen per bulan, akan membantu pengguna dalam mengelola berbagai aktivitas harian yang semakin luas: kalender, perjalanan, pembelian rumah tangga, pembayaran tagihan, pekerjaan rumah anak-anak, penyaringan email dan panggilan, pengingat, perencanaan makan, dan kita hanya bisa bermimpi menemukan kunci saya. Keterampilan ini tidak akan tersebar di berbagai aplikasi. Sebaliknya, keterampilan ini akan menjadi bagian dari satu agen terintegrasi yang dapat memanfaatkan sinergi yang tersedia dalam apa yang disebut oleh kalangan militer sebagai gambaran operasional umum.

Templat desain umum untuk asisten cerdas melibatkan pengetahuan latar belakang tentang aktivitas manusia, kemampuan untuk mengekstrak informasi dari aliran data perseptual dan tekstual, dan proses pembelajaran untuk menyesuaikan asisten dengan keadaan khusus pengguna. Templat umum yang sama dapat diterapkan pada setidaknya tiga bidang utama lainnya: kesehatan, pendidikan, dan keuangan. Untuk aplikasi ini, sistem perlu melacak kondisi tubuh, pikiran, dan rekening bank pengguna (dalam arti luas). Seperti halnya asisten untuk kehidupan sehari-hari, biaya awal untuk menciptakan pengetahuan umum yang diperlukan di masing-masing dari tiga bidang ini akan terbayar lunas bagi miliaran pengguna.

Dalam kasus kesehatan, misalnya, kita semua memiliki fisiologi yang kurang lebih sama, dan pengetahuan rinci tentang cara kerjanya telah dikodekan dalam bentuk yang dapat dibaca mesin. Sistem akan beradaptasi dengan karakteristik dan gaya hidup individu Anda, memberikan saran pencegahan dan peringatan dini tentang masalah.

Di bidang pendidikan, janji sistem bimbingan cerdas telah diakui bahkan pada tahun 1960-an, tetapi kemajuan nyata telah lama dinantikan. Alasan utamanya adalah kekurangan konten dan akses: sebagian besar sistem bimbingan tidak memahami isi dari apa yang mereka ajarkan, dan mereka juga tidak dapat terlibat dalam komunikasi dua arah dengan murid mereka melalui ucapan atau teks. (Saya membayangkan diri saya mengajar teori string, yang tidak saya mengerti, dalam bahasa Laos, yang tidak saya kuasai.)

Kemajuan terbaru dalam pengenalan suara berarti bahwa tutor otomatis akhirnya dapat berkomunikasi dengan murid yang belum sepenuhnya melek huruf. Selain itu, teknologi penalaran probabilistik kini dapat melacak apa yang diketahui dan tidak diketahui siswa dan dapat mengoptimalkan penyampaian instruksi untuk memaksimalkan pembelajaran. Kompetisi Global Learning XPRIZE, yang dimulai pada tahun 2014, menawarkan Rp 150 miliar untuk “perangkat lunak sumber terbuka dan terukur yang akan memungkinkan anak-anak di negara berkembang untuk belajar sendiri membaca, menulis, dan berhitung dasar dalam waktu 15 bulan.” Hasil dari para pemenang, Kitkit School dan onebillion, menunjukkan bahwa tujuan tersebut sebagian besar telah tercapai.

Di bidang keuangan pribadi, sistem akan melacak investasi, aliran pendapatan, pengeluaran wajib dan opsional, utang, pembayaran bunga, cadangan darurat, dan sebagainya, dengan cara yang hampir sama seperti analis keuangan melacak keuangan dan prospek perusahaan. Integrasi dengan agen yang menangani kehidupan sehari-hari akan memberikan pemahaman yang lebih detail, bahkan mungkin memastikan bahwa anak-anak mendapatkan uang saku mereka tanpa potongan terkait kenakalan. Kita dapat mengharapkan untuk menerima kualitas nasihat keuangan sehari-hari yang sebelumnya hanya diperuntukkan bagi orang-orang super kaya.

Jika alarm privasi Anda tidak berbunyi saat Anda membaca paragraf sebelumnya, Anda belum mengikuti berita terkini. Namun, ada beberapa lapisan dalam cerita privasi ini. Pertama, apakah asisten pribadi benar-benar dapat berguna jika ia tidak tahu apa pun tentang Anda? Mungkin tidak. Kedua, apakah asisten pribadi benar-benar dapat bermanfaat jika mereka tidak dapat mengumpulkan informasi dari banyak pengguna untuk mempelajari lebih lanjut tentang orang secara umum dan orang-orang yang mirip dengan Anda? Mungkin tidak. Jadi, bukankah

kedua hal itu menyiratkan bahwa kita harus mengorbankan privasi kita untuk mendapatkan manfaat dari AI dalam kehidupan sehari-hari kita? Tidak. Alasannya adalah algoritma pembelajaran dapat beroperasi pada data terenkripsi menggunakan teknik komputasi multipihak yang aman, sehingga pengguna dapat memperoleh manfaat dari pengumpulan data tanpa mengorbankan privasi dengan cara apa pun. Akankah penyedia perangkat lunak mengadopsi teknologi yang menjaga privasi secara sukarela, tanpa dorongan legislatif? Itu masih harus dilihat. Namun, yang tampaknya tak terhindarkan adalah bahwa pengguna hanya akan mempercayai asisten pribadi jika kewajiban utamanya adalah kepada pengguna, bukan kepada perusahaan yang memproduksinya.

Rumah Pintar dan Robot Rumah Tangga

Konsep rumah pintar telah diteliti selama beberapa dekade. Pada tahun 1966, James Sutherland, seorang insinyur di Westinghouse, mulai mengumpulkan suku cadang komputer bekas untuk membangun ECHO, pengontrol rumah pintar pertama. Sayangnya, ECHO memiliki berat delapan ratus pon, mengonsumsi 3,5 kilowatt, dan hanya mengelola tiga jam digital dan antena TV. Sistem selanjutnya mengharuskan pengguna untuk menguasai antarmuka kontrol dengan kompleksitas yang luar biasa. Tidak mengherankan, sistem tersebut tidak pernah populer.

Mulai tahun 1990-an, beberapa proyek ambisius mencoba merancang rumah yang mengelola dirinya sendiri dengan intervensi manusia minimal, menggunakan pembelajaran mesin untuk beradaptasi dengan gaya hidup penghuninya. Agar eksperimen ini bermakna, orang sungguhan harus tinggal di rumah-rumah tersebut. Sayangnya, frekuensi keputusan yang salah membuat sistem tersebut lebih buruk daripada tidak berguna, kualitas hidup penghuni menurun daripada meningkat. Misalnya, penghuni proyek MavHome tahun 2003 di Washington State University sering kali harus duduk dalam kegelapan jika tamu mereka tinggal lebih lama dari waktu tidur biasanya. Seperti halnya asisten pribadi yang tidak cerdas, kegagalan tersebut disebabkan oleh akses sensorik yang tidak memadai terhadap aktivitas penghuni dan ketidakmampuan untuk memahami dan melacak apa yang terjadi di rumah.

Rumah pintar sejati yang dilengkapi dengan kamera dan mikrofon dan kemampuan persepsi dan penalaran yang diperlukan dapat memahami apa yang dilakukan penghuninya: berkunjung, makan, tidur, menonton TV, membaca, berolahraga, bersiap untuk perjalanan jauh, atau tergeletak tak berdaya di lantai setelah jatuh. Dengan berkoordinasi dengan asisten pribadi yang cerdas, rumah tersebut dapat memiliki gambaran yang cukup baik tentang siapa yang akan berada di dalam atau di luar rumah pada jam berapa, siapa yang makan di mana, dan sebagainya.

Pemahaman ini memungkinkan rumah tersebut untuk mengelola pemanas, penerangan, tirai jendela, dan sistem keamanan, untuk mengirimkan pengingat tepat waktu, dan untuk memberi tahu pengguna atau layanan darurat ketika terjadi masalah. Beberapa kompleks apartemen yang baru dibangun di Amerika Serikat dan Jepang sudah menggabungkan teknologi semacam ini.

Nilai rumah pintar terbatas karena aktuatornya: sistem yang jauh lebih sederhana (termostat pengatur waktu dan lampu yang peka terhadap gerakan serta alarm anti maling)

dapat memberikan banyak fungsi yang sama dengan cara yang mungkin lebih mudah diprediksi, meskipun kurang peka terhadap konteks. Rumah pintar tidak dapat melipat pakaian, membersihkan piring, atau mengambil koran. Rumah pintar benar-benar menginginkan robot fisik untuk melakukan perintahnya.

Gambar 3.1 (kiri) BRETT melipat handuk; (kanan) robot SpotMini dari Boston Dynamics membuka pintu.

Mungkin tidak perlu menunggu terlalu lama. Robot telah menunjukkan banyak keterampilan yang dibutuhkan. Di laboratorium Berkeley milik kolega saya, Pieter Abbeel, BRETT (*Berkeley Robot for the Elimination of Tedious Tasks*) telah melipat tumpukan handuk sejak tahun 2011, sementara robot SpotMini dari Boston Dynamics dapat menaiki tangga dan membuka pintu (gambar 3.1). Beberapa perusahaan sudah membangun robot memasak, meskipun mereka membutuhkan pengaturan khusus yang tertutup dan bahan-bahan yang sudah dipotong dan tidak akan berfungsi di dapur biasa.

Dari tiga kemampuan fisik dasar yang dibutuhkan untuk robot rumah tangga yang berguna persepsi, mobilitas, dan ketangkasan yang terakhir adalah yang paling bermasalah. Seperti yang dikatakan Stefanie Tellex, seorang profesor robotika di Universitas Brown, “Sebagian besar robot tidak dapat mengambil sebagian besar objek sebagian besar waktu.” Ini sebagian merupakan masalah penginderaan taktil, sebagian masalah manufaktur (tangan yang cekatan saat ini sangat mahal untuk dibuat), dan sebagian masalah algoritmik: kita belum memiliki pemahaman yang baik tentang bagaimana menggabungkan penginderaan dan kontrol untuk menggenggam dan memanipulasi berbagai macam objek di rumah tangga biasa. Ada puluhan jenis genggaman hanya untuk objek kaku dan ada ribuan keterampilan manipulasi yang berbeda, seperti mengeluarkan tepat dua pil dari botol, mengupas label dari toples selai, mengoleskan mentega keras pada roti lunak, atau mengangkat satu helai spageti dari panci dengan garpu untuk melihat apakah sudah siap.

Tampaknya masalah penginderaan taktil dan konstruksi tangan akan dipecahkan oleh pencetakan 3D, yang sudah digunakan oleh Boston Dynamics untuk beberapa bagian yang lebih kompleks dari robot humanoid Atlas mereka. Keterampilan manipulasi robot berkembang pesat, sebagian berkat pembelajaran penguatan mendalam. Dorongan terakhir menggabungkan semua ini menjadi sesuatu yang mulai mendekati keterampilan fisik yang luar biasa dari robot film kemungkinan akan datang dari industri pergudangan yang agak tidak romantis. Hanya satu perusahaan, Amazon, mempekerjakan beberapa ratus ribu orang yang

mengambil produk dari tempat penyimpanan di gudang-gudang raksasa dan mengirimkannya ke pelanggan.

Dari tahun 2015 hingga 2017, Amazon menjalankan "Tantangan Pengambilan" tahunan untuk mempercepat pengembangan robot yang mampu melakukan tugas ini. Masih ada beberapa hal yang perlu dilakukan, tetapi ketika masalah penelitian inti terpecahkan, mungkin dalam satu dekade kita dapat mengharapkan peluncuran robot yang sangat mumpuni dengan sangat cepat. Awalnya mereka akan bekerja di gudang, kemudian dalam aplikasi komersial lainnya seperti pertanian dan konstruksi, di mana berbagai tugas dan objek cukup dapat diprediksi. Kita mungkin juga akan segera melihatnya di sektor ritel melakukan tugas-tugas seperti mengisi rak supermarket dan melipat kembali pakaian.

Yang pertama kali benar-benar mendapat manfaat dari robot di rumah adalah para lansia dan orang sakit, yang bagi mereka robot yang membantu dapat memberikan tingkat kemandirian yang tidak mungkin didapatkan tanpa robot. Bahkan jika robot memiliki repertoar tugas yang terbatas dan hanya pemahaman dasar tentang apa yang terjadi, robot tersebut tetap dapat sangat berguna. Di sisi lain, robot pelayan yang mengelola rumah tangga dengan penuh percaya diri dan mengantisipasi setiap keinginan tuannya, masih jauh dari kenyataan hal itu membutuhkan sesuatu yang mendekati kecerdasan buatan tingkat manusia.

Kecerdasan dalam Skala Global

Pengembangan kemampuan dasar untuk memahami ucapan dan teks akan memungkinkan asisten pribadi yang cerdas untuk melakukan hal-hal yang sudah dapat dilakukan oleh asisten manusia (tetapi mereka akan melakukannya dengan biaya beberapa sen per bulan, bukan ribuan dolar per bulan). Pemahaman ucapan dan teks dasar juga memungkinkan mesin untuk melakukan hal-hal yang tidak dapat dilakukan oleh manusia bukan karena kedalaman pemahamannya, tetapi karena skalanya. Misalnya, mesin dengan kemampuan membaca dasar akan dapat membaca semua yang pernah ditulis umat manusia pada waktu makan siang, dan kemudian akan mencari hal lain untuk dilakukan. Dengan kemampuan pengenalan ucapan, ia dapat mendengarkan setiap siaran radio dan televisi sebelum waktu minum teh. Sebagai perbandingan, dibutuhkan dua ratus ribu manusia penuh waktu hanya untuk mengikuti tingkat publikasi cetak dunia saat ini (apalagi semua materi tertulis dari masa lalu) dan enam puluh ribu lainnya untuk mendengarkan siaran saat ini.

Sistem seperti itu, jika dapat mengekstrak bahkan pernyataan faktual sederhana, dan mengintegrasikan semua informasi ini di semua bahasa, akan mewakili sumber daya yang luar biasa untuk menjawab pertanyaan dan mengungkap pola, mungkin jauh lebih ampuh daripada mesin pencari, yang saat ini bernilai sekitar Rp 10 kuadriliun. Nilai penelitiannya untuk bidang-bidang seperti sejarah dan sosiologi akan tak ternilai.

Tentu saja, dimungkinkan juga untuk mendengarkan semua panggilan telepon di dunia (pekerjaan yang membutuhkan sekitar dua puluh juta orang). Ada beberapa badan rahasia yang akan menganggap ini berharga. Beberapa dari mereka telah melakukan jenis pendengaran mesin skala besar yang sederhana, seperti menemukan kata kunci dalam percakapan, selama bertahun-tahun, dan sekarang telah beralih ke transkripsi seluruh

percakapan menjadi teks yang dapat dicari. Transkripsi tentu berguna, tetapi tidak seberguna pemahaman simultan dan integrasi konten dari semua percakapan.

"Kekuatan super" lain yang tersedia bagi mesin adalah kemampuan untuk melihat seluruh dunia sekaligus. Secara kasar, satelit memotret seluruh dunia setiap hari dengan resolusi rata-rata sekitar lima puluh sentimeter per piksel. Dengan resolusi ini, setiap rumah, kapal, mobil, sapi, dan pohon di Bumi terlihat. Lebih dari tiga puluh juta karyawan penuh waktu akan dibutuhkan untuk memeriksa semua gambar ini; jadi, saat ini, tidak ada manusia yang pernah melihat sebagian besar data satelit. Algoritma visi komputer dapat memproses semua data ini untuk menghasilkan basis data yang dapat dicari dari seluruh dunia, yang diperbarui setiap hari, serta visualisasi dan model prediktif kegiatan ekonomi, perubahan vegetasi, migrasi hewan dan manusia, dampak perubahan iklim, dan sebagainya. Perusahaan satelit seperti Planet dan DigitalGlobe sedang sibuk mewujudkan ide ini.

Dengan kemungkinan penginderaan dalam skala global, muncullah kemungkinan pengambilan keputusan dalam skala global. Misalnya, dari data satelit global, dimungkinkan untuk membuat model terperinci untuk mengelola lingkungan global, memprediksi dampak intervensi lingkungan dan ekonomi, dan menyediakan masukan analitis yang diperlukan untuk tujuan pembangunan berkelanjutan PBB. Kita sudah melihat sistem kontrol "kota pintar" yang bertujuan untuk mengoptimalkan manajemen lalu lintas, transportasi, pengumpulan sampah, perbaikan jalan, pemeliharaan lingkungan, dan fungsi lainnya untuk kepentingan warga, dan ini dapat diperluas ke tingkat negara. Hingga baru-baru ini, tingkat koordinasi ini hanya dapat dicapai oleh hierarki birokrasi manusia yang besar dan tidak efisien; tak pelak, ini akan digantikan oleh agen-agen raksasa yang mengurus semakin banyak aspek kehidupan kolektif kita. Bersamaan dengan ini, tentu saja, muncul kemungkinan pelanggaran privasi dan kontrol sosial dalam skala global, yang akan saya bahas kembali di bab berikutnya.

3.2 TANPA AMBANG BATAS: MENUJU AI SUPERCERDAS YANG TAK TERDUGA

Saya sering ditanya untuk memprediksi kapan AI supercerdas akan tiba, dan saya biasanya menolak untuk menjawab. Ada tiga alasan untuk ini. Pertama, ada sejarah panjang prediksi semacam itu yang seringkali meleset. Misalnya, pada tahun 1960, pelopor AI dan ekonom peraih Nobel, Herbert Simon, menulis, "Secara teknologi… mesin akan mampu, dalam dua puluh tahun, melakukan pekerjaan apa pun yang dapat dilakukan manusia." Pada tahun 1967, Marvin Minsky, salah satu penyelenggara lokakarya Dartmouth tahun 1956 yang memulai bidang AI, menulis, "Dalam satu generasi, saya yakin, hanya sedikit bidang intelektual yang akan tetap berada di luar jangkauan mesin masalah menciptakan 'kecerdasan buatan' akan sebagian besar terpecahkan."

Alasan kedua untuk menolak memberikan tanggal pasti untuk AI supercerdas adalah karena tidak ada ambang batas yang jelas yang akan dilampaui. Mesin sudah melampaui kemampuan manusia di beberapa bidang. Bidang-bidang tersebut akan meluas dan semakin dalam, dan kemungkinan akan ada sistem pengetahuan umum yang melampaui kemampuan manusia, sistem penelitian biomedis yang melampaui kemampuan manusia, robot yang cekatan dan lincah yang melampaui kemampuan manusia, sistem perencanaan perusahaan

yang melampaui kemampuan manusia, dan sebagainya jauh sebelum kita memiliki sistem AI supercerdas yang sepenuhnya umum. Sistem-sistem "supercerdas sebagian" ini, secara individual dan kolektif, akan mulai menimbulkan banyak masalah yang sama seperti yang akan ditimbulkan oleh sistem yang cerdas secara umum.

Alasan ketiga untuk tidak memprediksi kedatangan AI supercerdas adalah karena hal itu pada dasarnya tidak dapat diprediksi. Hal ini membutuhkan "terobosan konseptual," seperti yang dicatat oleh John McCarthy dalam sebuah wawancara tahun 1977. McCarthy melanjutkan dengan mengatakan, "Yang Anda inginkan adalah 1.7 Einstein dan 0.3 dari Proyek Manhattan, dan Anda menginginkan Einstein terlebih dahulu. Saya percaya itu akan memakan waktu lima hingga 500 tahun." Di bagian selanjutnya saya akan menjelaskan beberapa terobosan konseptual yang mungkin terjadi. Seberapa tidak dapat diprediksi terobosan tersebut? Mungkin sama tidak dapat diprediksinya dengan penemuan reaksi berantai nuklir oleh Szilard beberapa jam setelah pernyataan Rutherford bahwa itu sama sekali tidak mungkin.

Suatu kali, pada pertemuan Forum Ekonomi Dunia pada tahun 2015, saya menjawab pertanyaan tentang kapan kita mungkin melihat AI supercerdas. Pertemuan tersebut berada di bawah aturan Chatham House, yang berarti bahwa tidak ada pernyataan yang dapat dikaitkan dengan siapa pun yang hadir dalam pertemuan tersebut. Meskipun demikian, karena terlalu berhati-hati, saya mengawali jawaban saya dengan "Sangat rahasia" Saya menyarankan bahwa, kecuali terjadi bencana yang menghalangi, hal itu mungkin akan terjadi selama masa hidup anak-anak saya yang masih cukup muda dan mungkin akan memiliki umur yang jauh lebih panjang, berkat kemajuan dalam ilmu kedokteran, daripada banyak orang yang hadir dalam pertemuan tersebut. Kurang dari dua jam kemudian, sebuah artikel muncul di Daily Telegraph yang mengutip pernyataan Profesor Russell, lengkap dengan gambar robot Terminator yang mengamuk. Judulnya adalah 'ROBOT 'SOSIOPATIK' DAPAT MENGUASAI UMAT MANUSIA DALAM SATU GENERASI.

Garis waktu saya, katakanlah, delapan puluh tahun jauh lebih konservatif daripada yang diperkirakan oleh peneliti AI pada umumnya. Survei terbaru menunjukkan bahwa sebagian besar peneliti aktif memperkirakan AI tingkat manusia akan tiba sekitar pertengahan abad ini. Pengalaman kita dengan fisika nuklir menunjukkan bahwa akan bijaksana untuk berasumsi bahwa kemajuan dapat terjadi dengan cukup cepat dan untuk mempersiapkan diri sesuai dengan itu. Jika hanya dibutuhkan satu terobosan konseptual, analog dengan ide Szilard tentang reaksi rantai nuklir yang diinduksi neutron, AI supercerdas dalam beberapa bentuk dapat tiba-tiba muncul. Kemungkinannya adalah kita tidak akan siap: jika kita membangun mesin supercerdas dengan otonomi tingkat apa pun, kita akan segera mendapati diri kita tidak mampu mengendalikannya. Namun, saya cukup yakin bahwa kita memiliki ruang bernapas karena ada beberapa terobosan besar yang dibutuhkan antara sekarang dan kecerdasan super, bukan hanya satu.

3.3 TEROBOSAN KONSEPTUAL YANG AKAN DATANG

Masalah menciptakan AI serbaguna setingkat manusia masih jauh dari terselesaikan. Menyelesaikannya bukanlah masalah menghabiskan uang untuk lebih banyak insinyur, lebih banyak data, dan komputer yang lebih besar. Beberapa futuris menghasilkan grafik yang mengekstrapolasi pertumbuhan eksponensial daya komputasi ke masa depan berdasarkan hukum Moore, menunjukkan tanggal-tanggal ketika mesin akan menjadi lebih kuat daripada otak serangga, otak tikus, otak manusia, semua otak manusia digabungkan, dan seterusnya. Grafik-grafik ini tidak berarti karena, seperti yang telah saya katakan, mesin yang lebih cepat hanya memberikan jawaban yang salah lebih cepat.

Jika seseorang mengumpulkan para ahli AI terkemuka ke dalam satu tim dengan sumber daya tak terbatas, dengan tujuan menciptakan sistem cerdas terintegrasi setingkat manusia dengan menggabungkan semua ide terbaik kita, hasilnya akan gagal. Sistem tersebut akan rusak di dunia nyata. Sistem tersebut tidak akan memahami apa yang sedang terjadi; sistem tersebut tidak akan mampu memprediksi konsekuensi dari tindakannya; sistem tersebut tidak akan memahami apa yang diinginkan orang dalam situasi tertentu; dan karenanya sistem tersebut akan melakukan hal-hal yang sangat bodoh.

Dengan memahami bagaimana sistem tersebut akan rusak, para peneliti AI mampu mengidentifikasi masalah-masalah yang harus dipecahkan, terobosan konseptual yang dibutuhkan untuk mencapai AI setingkat manusia. Sekarang saya akan menjelaskan beberapa masalah yang masih tersisa ini. Setelah masalah-masalah tersebut terpecahkan, mungkin akan ada masalah lain, tetapi tidak akan terlalu banyak.

Bahasa dan Akal Sehat

Kecerdasan tanpa pengetahuan bagaikan mesin tanpa bahan bakar. Manusia memperoleh sejumlah besar pengetahuan dari manusia lain: pengetahuan itu diturunkan dari generasi ke generasi dalam bentuk bahasa. Sebagian di antaranya bersifat faktual: Obama menjadi presiden pada tahun 2009, kepadatan tembaga adalah 8,92 gram per sentimeter kubik, kode Ur-Nammu menetapkan hukuman untuk berbagai kejahatan, dan sebagainya. Sebagian besar pengetahuan terdapat dalam bahasa itu sendiri dalam konsep-konsep yang tersedia di dalamnya. Presiden, 2009, kepadatan, tembaga, gram, sentimeter, kejahatan, dan lainnya semuanya membawa sejumlah besar informasi, yang mewakili esensi yang diekstrak dari proses penemuan dan pengorganisasian yang menyebabkan mereka ada dalam bahasa sejak awal.

Ambil contoh, tembaga, yang merujuk pada beberapa kumpulan atom di alam semesta, dan bandingkan dengan arglebarglium, yang merupakan nama saya untuk kumpulan atom yang sama besarnya yang dipilih secara acak di alam semesta. Ada banyak hukum umum, bermanfaat, dan prediktif yang dapat ditemukan tentang tembaga tentang kepadatannya, konduktivitasnya, kelenturannya, titik lelehnya, asal usul bintangnya, senyawa kimianya, penggunaan praktisnya, dan sebagainya; sebagai perbandingan, pada dasarnya tidak ada yang dapat dikatakan tentang arglebarglium. Organisme yang dilengkapi dengan bahasa yang terdiri dari kata-kata seperti arglebarglium tidak akan mampu berfungsi, karena ia tidak akan pernah

menemukan keteraturan yang memungkinkannya untuk memodelkan dan memprediksi alam semestanya.

Mesin yang benar-benar memahami bahasa manusia akan mampu dengan cepat memperoleh sejumlah besar pengetahuan manusia, memungkinkannya untuk melewati puluhan ribu tahun pembelajaran oleh lebih dari seratus miliar orang yang telah hidup di Bumi. Tampaknya tidak praktis untuk mengharapkan mesin untuk menemukan kembali semua ini dari awal, dimulai dari data sensorik mentah.

Namun, saat ini, teknologi bahasa alami belum mampu membaca dan memahami jutaan buku banyak di antaranya akan membingungkan bahkan manusia yang berpendidikan tinggi. Sistem seperti Watson milik IBM, yang terkenal mengalahkan dua juara manusia dalam Jeopardy! Permainan kuis pada tahun 2011 dapat mengekstrak informasi sederhana dari fakta yang dinyatakan dengan jelas tetapi tidak dapat membangun struktur pengetahuan yang kompleks dari teks; mereka juga tidak dapat menjawab pertanyaan yang membutuhkan rangkaian penalaran yang luas dengan informasi dari berbagai sumber. Misalnya, tugas membaca semua dokumen yang tersedia hingga akhir tahun 1973 dan menilai (dengan penjelasan) kemungkinan hasil dari proses pemakzulan Watergate terhadap presiden Nixon saat itu akan jauh melampaui kemampuan teknologi terkini.

Ada upaya serius yang sedang dilakukan untuk memperdalam tingkat analisis bahasa dan ekstraksi informasi. Misalnya, Proyek Aristo di Allen Institute for AI bertujuan untuk membangun sistem yang dapat lulus ujian sains sekolah setelah membaca buku teks dan panduan belajar. Berikut adalah pertanyaan dari tes kelas empat:

> ***Siswa kelas empat sedang merencanakan perlombaan sepatu roda. Permukaan mana yang terbaik untuk perlombaan ini?***
> (A) kerikil
> (B) pasir
> (C) aspal
> (D) rumput

Sebuah mesin menghadapi setidaknya dua sumber kesulitan dalam menjawab pertanyaan ini. Yang pertama adalah masalah pemahaman bahasa klasik, yaitu memahami apa yang dikatakan kalimat-kalimat tersebut: menganalisis struktur sintaksis, mengidentifikasi arti kata-kata, dan sebagainya. (Cobalah sendiri: gunakan layanan terjemahan daring untuk menerjemahkan kalimat-kalimat tersebut ke dalam bahasa yang tidak dikenal, lalu gunakan kamus untuk bahasa tersebut untuk mencoba menerjemahkannya kembali ke bahasa Inggris.)

Yang kedua adalah kebutuhan akan pengetahuan akal sehat: untuk memahami bahwa "balapan sepatu roda" mungkin adalah balapan antara orang-orang yang mengenakan sepatu roda (di kaki mereka) daripada balapan antara sepatu roda, untuk memahami bahwa "permukaan" adalah tempat para pemain sepatu roda akan meluncur daripada tempat para penonton akan duduk, untuk mengetahui apa arti "terbaik" dalam konteks permukaan untuk

balapan, dan sebagainya. Bayangkan bagaimana jawabannya mungkin berubah jika kita mengganti "siswa kelas empat" dengan "pelatih kamp pelatihan militer yang sadis."

Salah satu cara untuk meringkas kesulitan tersebut adalah dengan mengatakan bahwa membaca membutuhkan pengetahuan dan pengetahuan (sebagian besar) berasal dari membaca. Dengan kata lain, kita menghadapi situasi klasik ayam dan telur. Kita mungkin berharap pada proses bootstrapping, di mana sistem membaca beberapa teks yang mudah, memperoleh beberapa pengetahuan, menggunakannya untuk membaca teks yang lebih sulit, memperoleh lebih banyak pengetahuan lagi, dan seterusnya. Sayangnya, yang cenderung terjadi adalah kebalikannya: pengetahuan yang diperoleh sebagian besar salah, yang menyebabkan kesalahan dalam membaca, yang menghasilkan lebih banyak pengetahuan yang salah, dan seterusnya.

Sebagai contoh, proyek NELL (*Never-Ending Language Learning*) di Universitas Carnegie Mellon mungkin merupakan proyek bootstrapping bahasa paling ambisius yang sedang berlangsung saat ini. Dari tahun 2010 hingga 2018, NELL memperoleh lebih dari 120 juta kepercayaan dengan membaca teks bahasa Inggris di Web. Beberapa kepercayaan ini akurat, seperti kepercayaan bahwa Maple Leafs bermain hoki dan memenangkan Piala Stanley. Selain fakta, NELL memperoleh kosakata, kategori, dan hubungan semantik baru sepanjang waktu. Sayangnya, NELL hanya yakin pada 3 persen dari keyakinannya dan bergantung pada para ahli manusia untuk membersihkan keyakinan yang salah atau tidak bermakna secara teratur seperti keyakinannya bahwa “Nepal adalah negara yang juga dikenal sebagai Amerika Serikat” dan “nilai adalah produk pertanian yang biasanya dipotong menjadi basis.”

Saya menduga mungkin tidak ada satu terobosan pun yang dapat mengubah spiral penurunan menjadi spiral peningkatan. Proses bootstrapping dasar tampaknya tepat: sebuah program yang mengetahui cukup banyak fakta dapat mengetahui fakta mana yang dirujuk oleh kalimat baru, dan dengan demikian mempelajari bentuk tekstual baru untuk mengekspresikan fakta yang kemudian memungkinkannya untuk menemukan lebih banyak fakta, dan proses pun berlanjut. (Sergey Brin, salah satu pendiri Google, menerbitkan makalah penting tentang ide bootstrapping pada tahun 1998.) Mempercepat proses dengan menyediakan banyak pengetahuan dan informasi linguistik yang dikodekan secara manual tentu akan membantu. Meningkatkan kecanggihan representasi fakta memungkinkan peristiwa kompleks, hubungan sebab akibat, kepercayaan dan sikap orang lain, dan sebagainya dan meningkatkan penanganan ketidakpastian tentang makna kata dan makna kalimat pada akhirnya dapat menghasilkan proses pembelajaran yang saling memperkuat daripada saling memadamkan.

Pembelajaran Kumulatif Konsep dan Teori

Sekitar 1,4 miliar tahun yang lalu dan 8,2 sextiliun mil jauhnya, dua lubang hitam, satu dengan massa dua belas juta kali massa Bumi dan yang lainnya sepuluh juta, mendekat cukup untuk mulai mengorbit satu sama lain. Secara bertahap kehilangan energi, mereka berputar semakin dekat satu sama lain dan semakin cepat, mencapai frekuensi orbit 250 kali per detik pada jarak 350 kilometer sebelum akhirnya bertabrakan dan bergabung. Dalam beberapa milidetik terakhir, laju emisi energi dalam bentuk gelombang gravitasi lima puluh kali lebih besar daripada total keluaran energi semua bintang di alam semesta. Pada tanggal 14

September 2015, gelombang gravitasi tersebut tiba di Bumi. Gelombang tersebut secara bergantian memperluas dan memampatkan ruang itu sendiri dengan faktor sekitar satu dalam 2,5 sextiliun, setara dengan mengubah jarak ke Proxima Centauri (4,4 tahun cahaya) selebar rambut manusia.

Untungnya, dua hari sebelumnya, detektor Advanced LIGO (*Laser Interferometer Gravitational-Wave Observatory*) di Washington dan Louisiana telah diaktifkan. Dengan menggunakan interferometri laser, mereka mampu mengukur distorsi ruang yang sangat kecil; menggunakan perhitungan berdasarkan teori relativitas umum Einstein, para peneliti LIGO telah memprediksi dan karena itu mencari bentuk pasti dari gelombang gravitasi yang diharapkan dari peristiwa tersebut.

Hal ini dimungkinkan karena akumulasi dan komunikasi pengetahuan dan konsep oleh ribuan orang selama berabad-abad pengamatan dan penelitian. Dari Thales dari Miletus yang menggosok amber dengan wol dan mengamati penumpukan muatan statis, hingga Galileo yang menjatuhkan batu dari Menara Miring Pisa, hingga Newton yang melihat apel jatuh dari pohon, dan seterusnya melalui ribuan pengamatan lainnya, umat manusia secara bertahap telah mengumpulkan lapisan demi lapisan konsep, teori, dan perangkat: massa, kecepatan, percepatan, gaya, hukum gerak dan gravitasi Newton, persamaan orbital, fenomena listrik, atom, elektron, medan listrik, medan magnet, gelombang elektromagnetik, relativitas khusus, relativitas umum, mekanika kuantum, semikonduktor, laser, komputer, dan sebagainya.

Sekarang, pada prinsipnya kita dapat memahami proses penemuan ini sebagai pemetaan dari semua data sensorik yang pernah dialami oleh semua manusia ke hipotesis yang sangat kompleks tentang data sensorik yang dialami oleh para ilmuwan LIGO pada tanggal 14 September 2015, saat mereka mengamati layar komputer mereka. Ini adalah pandangan pembelajaran yang sepenuhnya berbasis data: data masuk, hipotesis keluar, kotak hitam di antaranya. Jika itu bisa dilakukan, itu akan menjadi puncak dari pendekatan pembelajaran mendalam "big data, big network", tetapi itu tidak bisa dilakukan.

Satu-satunya ide yang masuk akal yang kita miliki tentang bagaimana entitas cerdas dapat mencapai prestasi luar biasa seperti mendeteksi penggabungan dua lubang hitam adalah bahwa pengetahuan fisika sebelumnya, dikombinasikan dengan data observasi dari instrumen mereka, memungkinkan para ilmuwan LIGO untuk menyimpulkan terjadinya peristiwa penggabungan tersebut. Terlebih lagi, pengetahuan sebelumnya ini sendiri merupakan hasil dari pembelajaran dengan pengetahuan sebelumnya dan seterusnya, hingga kembali ke masa lalu. Dengan demikian, kita memiliki gambaran kumulatif tentang bagaimana entitas cerdas dapat membangun kemampuan prediktif, dengan pengetahuan sebagai bahan pembangunnya.

Saya katakan kira-kira karena, tentu saja, sains telah mengambil beberapa jalan yang salah selama berabad-abad, untuk sementara mengejar gagasan ilusi seperti flogiston dan eter pembawa cahaya. Namun kita tahu pasti bahwa gambaran kumulatif itulah yang sebenarnya terjadi, dalam arti bahwa para ilmuwan di sepanjang jalan mencatat temuan dan teori mereka dalam buku dan makalah. Para ilmuwan selanjutnya hanya memiliki akses ke bentuk-bentuk pengetahuan eksplisit ini, dan bukan ke pengalaman indrawi asli dari generasi sebelumnya

yang telah lama meninggal. Karena mereka adalah ilmuwan, anggota tim LIGO memahami bahwa semua pengetahuan yang mereka gunakan, termasuk teori relativitas umum Einstein, berada (dan akan selalu berada) dalam masa percobaan dan dapat dibuktikan salah melalui eksperimen. Ternyata, data LIGO memberikan konfirmasi yang kuat untuk relativitas umum serta bukti lebih lanjut bahwa graviton partikel yang dihipotesiskan yang memediasi gaya gravitasi tidak bermassa.

Kita masih sangat jauh dari kemampuan untuk menciptakan sistem pembelajaran mesin yang mampu menyamai atau melampaui kapasitas pembelajaran dan penemuan kumulatif yang ditunjukkan oleh komunitas ilmiah atau oleh manusia biasa dalam masa hidup mereka sendiri. Sistem pembelajaran mendalam sebagian besar didorong oleh data: paling banter, kita dapat "memasukkan" beberapa bentuk pengetahuan sebelumnya yang sangat lemah dalam struktur jaringan. Sistem pemrograman probabilistik memang memungkinkan pengetahuan sebelumnya dalam proses pembelajaran, seperti yang diungkapkan dalam struktur dan kosakata basis pengetahuan probabilistik, tetapi kita belum memiliki metode yang efektif untuk menghasilkan konsep dan hubungan baru dan menggunakannya untuk memperluas basis pengetahuan tersebut.

Kesulitannya bukanlah menemukan hipotesis yang memberikan kecocokan yang baik dengan data; sistem pembelajaran mendalam dapat menemukan hipotesis yang cocok dengan data gambar, dan peneliti AI telah membangun program pembelajaran simbolik yang mampu merekapitulasi banyak penemuan historis hukum ilmiah kuantitatif. Pembelajaran dalam agen cerdas otonom membutuhkan lebih dari itu.

Pertama, apa yang harus disertakan dalam "data" yang digunakan untuk membuat prediksi? Misalnya, dalam eksperimen LIGO, model untuk memprediksi seberapa besar ruang meregang dan menyusut ketika gelombang gravitasi tiba memperhitungkan massa lubang hitam yang bertabrakan, frekuensi orbitnya, dan sebagainya, tetapi tidak memperhitungkan hari dalam seminggu atau terjadinya pertandingan bisbol Liga Utama. Di sisi lain, model untuk memprediksi lalu lintas di Jembatan Teluk San Francisco memperhitungkan hari dalam seminggu dan terjadinya pertandingan bisbol Liga Utama tetapi mengabaikan massa dan frekuensi orbit lubang hitam yang bertabrakan.

Demikian pula, program yang belajar mengenali jenis objek dalam gambar menggunakan piksel sebagai input, sedangkan program yang belajar memperkirakan nilai suatu benda antik juga ingin mengetahui bahan pembuatannya, siapa yang membuatnya dan kapan, sejarah penggunaan dan kepemilikannya, dan sebagainya. Mengapa demikian? Jelas, karena kita manusia sudah mengetahui sesuatu tentang gelombang gravitasi, lalu lintas, gambar visual, dan barang antik. Kita menggunakan pengetahuan ini untuk menentukan input mana yang dibutuhkan untuk memprediksi output tertentu. Ini disebut rekayasa fitur, dan melakukannya dengan baik membutuhkan pemahaman yang baik tentang masalah prediksi spesifik tersebut.

Tentu saja, mesin cerdas sejati tidak dapat bergantung pada insinyur fitur manusia yang muncul setiap kali ada sesuatu yang baru untuk dipelajari. Mesin tersebut harus menentukan sendiri apa yang merupakan ruang hipotesis yang masuk akal untuk suatu masalah

pembelajaran. Kemungkinan besar, mesin tersebut akan melakukan ini dengan memanfaatkan berbagai pengetahuan yang relevan dalam berbagai bentuk, tetapi saat ini kita hanya memiliki gagasan dasar tentang bagaimana melakukan hal ini. Buku Nelson Goodman, Fact, Fiction, and Forecast yang ditulis pada tahun 1954 dan mungkin merupakan salah satu buku terpenting dan kurang dihargai tentang pembelajaran mesin menyarankan jenis pengetahuan yang disebut overhipotesis, karena membantu mendefinisikan apa yang mungkin menjadi ruang hipotesis yang masuk akal.

Dalam kasus prediksi lalu lintas, misalnya, overhipotesis yang relevan adalah bahwa hari dalam seminggu, waktu dalam sehari, peristiwa lokal, kecelakaan baru-baru ini, hari libur, keterlambatan transit, cuaca, dan waktu matahari terbit dan terbenam dapat memengaruhi kondisi lalu lintas. (Perhatikan bahwa Anda dapat mengetahui hipotesis ini dari pengetahuan latar belakang Anda sendiri tentang dunia, tanpa harus menjadi ahli lalu lintas.) Sistem pembelajaran cerdas dapat mengakumulasi dan menggunakan pengetahuan semacam ini untuk membantu merumuskan dan memecahkan masalah pembelajaran baru.

Kedua, dan mungkin lebih penting, adalah akumulasi generasi konsep-konsep baru seperti massa, percepatan, muatan, elektron, dan gaya gravitasi. Tanpa konsep-konsep ini, para ilmuwan (dan orang awam) harus menafsirkan alam semesta mereka dan membuat prediksi berdasarkan masukan persepsi mentah. Sebaliknya, Newton mampu bekerja dengan konsep massa dan percepatan yang dikembangkan oleh Galileo dan lainnya; Rutherford dapat menentukan bahwa atom terdiri dari inti padat bermuatan positif yang dikelilingi oleh elektron karena konsep elektron telah dikembangkan (oleh banyak peneliti secara bertahap) pada akhir abad kesembilan belas; memang, semua penemuan ilmiah bergantung pada lapisan demi lapisan konsep yang membentang sepanjang waktu dan pengalaman manusia.

Dalam filsafat ilmu, khususnya pada awal abad ke-20, tidak jarang kita melihat penemuan konsep-konsep baru yang dikaitkan dengan tiga "I" yang tak terlukiskan: intuisi, wawasan, dan inspirasi. Ketiganya dianggap resisten terhadap penjelasan rasional atau algoritmik apa pun. Para peneliti AI, termasuk Herbert Simon, telah sangat menentang pandangan ini. Sederhananya, jika algoritma pembelajaran mesin dapat mencari dalam ruang hipotesis yang mencakup kemungkinan menambahkan definisi untuk istilah-istilah baru yang tidak ada dalam input, maka algoritma tersebut dapat menemukan konsep-konsep baru.

Sebagai contoh, anggaplah sebuah robot mencoba mempelajari aturan backgammon dengan mengamati orang-orang yang bermain. Robot tersebut mengamati bagaimana mereka melempar dadu dan memperhatikan bahwa terkadang pemain memindahkan tiga atau empat bidak daripada satu atau dua, dan ini terjadi setelah lemparan 1-1, 2-2, 3-3, 4-4, 5-5, atau 6-6. Jika program tersebut dapat menambahkan konsep baru tentang angka ganda, yang didefinisikan oleh kesamaan antara kedua dadu, program tersebut dapat mengekspresikan teori prediksi yang sama dengan jauh lebih ringkas. Ini adalah proses yang mudah, menggunakan metode seperti pemrograman logika induktif, untuk membuat program yang mengusulkan konsep dan definisi baru untuk mengidentifikasi teori yang akurat dan ringkas.

Saat ini, kita tahu bagaimana melakukan ini untuk kasus yang relatif sederhana, tetapi untuk teori yang lebih kompleks, jumlah konsep baru yang mungkin diperkenalkan menjadi

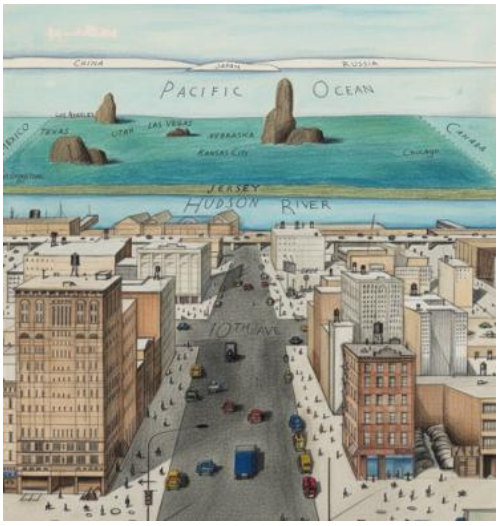
sangat besar. Hal ini membuat keberhasilan metode pembelajaran mendalam baru-baru ini dalam visi komputer menjadi semakin menarik. Jaringan dalam biasanya berhasil menemukan fitur perantara yang berguna seperti mata, kaki, garis, dan sudut, meskipun mereka menggunakan algoritma pembelajaran yang sangat sederhana. Jika kita dapat memahami lebih baik bagaimana hal ini terjadi, kita dapat menerapkan pendekatan yang sama untuk mempelajari konsep baru dalam bahasa yang lebih ekspresif yang dibutuhkan untuk sains. Ini sendiri akan menjadi keuntungan besar bagi umat manusia serta langkah signifikan menuju AI tujuan umum.

Menemukan Tindakan

Perilaku cerdas dalam skala waktu yang panjang membutuhkan kemampuan untuk merencanakan dan mengelola aktivitas secara hierarkis, pada berbagai tingkat abstraksi mulai dari mengerjakan PhD (satu triliun tindakan) hingga perintah kontrol motorik tunggal yang dikirim ke satu jari sebagai bagian dari mengetik satu karakter dalam surat lamaran kerja.

Aktivitas kita diorganisasikan ke dalam hierarki kompleks dengan puluhan tingkat abstraksi. Tingkat-tingkat ini dan tindakan yang dikandungnya merupakan bagian penting dari peradaban kita dan diwariskan dari generasi ke generasi melalui bahasa dan praktik kita. Misalnya, tindakan seperti menangkap babi hutan, mengajukan visa, dan membeli tiket pesawat mungkin melibatkan jutaan tindakan primitif, tetapi kita dapat memikirkannya sebagai unit tunggal karena tindakan tersebut sudah ada di "perpustakaan" tindakan yang disediakan oleh bahasa dan budaya kita dan karena kita tahu (kira-kira) bagaimana melakukannya.

Setelah berada di perpustakaan, kita dapat merangkai tindakan tingkat tinggi ini menjadi tindakan tingkat yang lebih tinggi lagi, seperti mengadakan pesta suku untuk titik balik matahari musim panas atau melakukan penelitian arkeologi selama musim panas di daerah terpencil di Nepal. Mencoba merencanakan aktivitas semacam itu dari awal, dimulai dengan langkah-langkah kontrol motorik tingkat terendah, akan benar-benar sia-sia karena aktivitas tersebut melibatkan jutaan atau miliaran langkah, yang banyak di antaranya sangat tidak dapat diprediksi. (Di mana babi hutan akan ditemukan, dan ke arah mana ia akan lari?) Di sisi lain, dengan tindakan tingkat tinggi yang sesuai di perpustakaan, seseorang hanya perlu merencanakan sekitar selusin langkah, karena setiap langkah tersebut merupakan bagian besar dari keseluruhan aktivitas. Ini adalah sesuatu yang bahkan otak manusia kita yang lemah pun dapat kelola tetapi ini memberi kita "kekuatan super" untuk merencanakan dalam skala waktu yang panjang.



Gambar 3.2 Pandangan Saul Steinberg tentang Dunia dari 9th Avenue, 1976, pertama kali diterbitkan sebagai sampul majalah The New Yorker.

Ada suatu waktu ketika tindakan-tindakan ini tidak ada misalnya, untuk mendapatkan hak untuk melakukan perjalanan pesawat pada tahun 1910 akan membutuhkan proses penelitian, penulisan surat, dan negosiasi yang panjang, rumit, dan tidak dapat diprediksi dengan berbagai pelopor penerbangan. Tindakan lain yang baru-baru ini ditambahkan ke perpustakaan termasuk mengirim email, mencari di Google, dan menggunakan Uber. Seperti yang ditulis Alfred North Whitehead pada tahun 1911, “Peradaban maju dengan memperluas jumlah operasi penting yang dapat kita lakukan tanpa memikirkannya.” Sampul terkenal Saul Steinberg untuk *The New Yorker* (gambar 3.2) secara brilian menunjukkan, dalam bentuk spasial, bagaimana agen cerdas mengelola masa depannya sendiri. Masa depan yang sangat dekat sangat detail bahkan, otak saya telah memuat urutan kontrol motorik spesifik untuk mengetik beberapa kata berikutnya.

Melihat sedikit lebih jauh ke depan, detailnya lebih sedikit rencana saya adalah menyelesaikan bagian ini, makan siang, menulis lebih banyak, dan menonton Prancis bermain melawan Kroasia di final Piala Dunia. Lebih jauh lagi, rencana saya lebih besar tetapi lebih samar: pindah kembali dari Paris ke Berkeley pada awal Agustus, mengajar mata kuliah pascasarjana, dan menyelesaikan buku ini. Saat seseorang bergerak melalui waktu, masa depan bergerak lebih dekat ke masa kini dan rencana untuknya menjadi lebih detail, sementara rencana baru yang samar dapat ditambahkan ke masa depan yang jauh. Rencana untuk masa depan yang dekat menjadi sangat detail sehingga dapat dieksekusi langsung oleh sistem kontrol motorik.

Saat ini kita hanya memiliki beberapa bagian dari gambaran keseluruhan ini untuk sistem AI. Jika hierarki tindakan abstrak disediakan termasuk pengetahuan tentang bagaimana setiap tindakan abstrak dapat disempurnakan menjadi subrencana yang terdiri dari tindakan yang lebih konkret maka kita memiliki algoritma yang dapat membangun rencana kompleks untuk mencapai tujuan tertentu. Ada algoritma yang dapat mengeksekusi rencana abstrak dan hierarkis sedemikian rupa sehingga agen selalu memiliki tindakan fisik primitif yang "siap

dijalankan," bahkan jika tindakan di masa depan masih berada pada tingkat abstrak dan belum dapat dieksekusi.

Bagian utama yang hilang dari teka-teki ini adalah metode untuk membangun hierarki tindakan abstrak sejak awal. Misalnya, apakah mungkin untuk memulai dari awal dengan robot yang hanya tahu bahwa ia dapat mengirim berbagai arus listrik ke berbagai motor dan membiarkannya menemukan sendiri tindakan berdiri? Penting untuk dipahami bahwa saya tidak bertanya apakah kita dapat melatih robot untuk berdiri, yang dapat dilakukan hanya dengan menerapkan pembelajaran penguatan dengan imbalan untuk kepala robot yang lebih jauh dari tanah.

Melatih robot untuk berdiri membutuhkan pelatih manusia untuk sudah mengetahui apa arti berdiri, sehingga sinyal imbalan yang tepat dapat didefinisikan. Yang kita inginkan adalah agar robot menemukan sendiri bahwa berdiri adalah suatu hal, suatu tindakan abstrak yang berguna, yang mencapai prasyarat (berdiri tegak) untuk berjalan atau berlari atau berjabat tangan atau melihat melewati tembok dan dengan demikian menjadi bagian dari banyak rencana abstrak untuk semua jenis tujuan. Demikian pula, kita ingin robot menemukan tindakan seperti berpindah dari satu tempat ke tempat lain, mengambil benda, membuka pintu, mengikat simpul, memasak makan malam, menemukan kunci saya, membangun rumah, dan banyak tindakan lain yang tidak memiliki nama dalam bahasa manusia karena kita manusia belum menemukannya.

Saya percaya kemampuan ini adalah langkah terpenting yang dibutuhkan untuk mencapai AI tingkat manusia. Ini akan, untuk meminjam ungkapan Whitehead lagi, memperluas jumlah operasi penting yang dapat dilakukan sistem AI tanpa memikirkannya. Banyak kelompok riset di seluruh dunia sedang bekerja keras untuk memecahkan masalah ini. Misalnya, makalah DeepMind tahun 2018 yang menunjukkan performa setara manusia pada Quake III Arena Capture the Flag mengklaim bahwa sistem pembelajaran mereka “membangun ruang representasi hierarkis temporal dengan cara baru untuk mendorong… urutan aksi yang koheren secara temporal.” (Saya tidak sepenuhnya yakin apa artinya ini, tetapi tentu saja terdengar seperti kemajuan menuju tujuan menciptakan aksi tingkat tinggi yang baru.) Saya menduga bahwa kita belum memiliki jawaban lengkap, tetapi ini adalah kemajuan yang dapat terjadi kapan saja, hanya dengan menggabungkan beberapa ide yang ada dengan cara yang tepat.

Mesin cerdas dengan kemampuan ini akan mampu melihat lebih jauh ke masa depan daripada yang dapat dilakukan manusia. Mereka juga akan mampu mempertimbangkan informasi yang jauh lebih banyak. Kombinasi kedua kemampuan ini pasti mengarah pada keputusan dunia nyata yang lebih baik. Dalam situasi konflik apa pun antara manusia dan mesin, kita akan segera menemukan, seperti Garry Kasparov dan Lee Sedol, bahwa setiap gerakan kita telah diantisipasi dan diblokir. Kita akan kalah dalam permainan bahkan sebelum dimulai.

Mengelola Aktivitas Mental

Jika mengelola aktivitas di dunia nyata tampak kompleks, bayangkan betapa beratnya otak Anda, yang mengelola aktivitas "objek paling kompleks di alam semesta yang dikenal"

yaitu dirinya sendiri. Kita tidak langsung tahu cara berpikir, sama seperti kita tidak langsung tahu cara berjalan atau bermain piano. Kita belajar cara melakukannya. Kita, sampai batas tertentu, dapat memilih pikiran apa yang ingin kita miliki. (Coba pikirkan tentang hamburger yang lezat atau peraturan bea cukai Bulgaria pilihan Anda!) Dalam beberapa hal, aktivitas mental kita lebih kompleks daripada aktivitas kita di dunia nyata, karena otak kita memiliki lebih banyak bagian yang bergerak daripada tubuh kita dan bagian-bagian tersebut bergerak jauh lebih cepat.

Hal yang sama berlaku untuk komputer: untuk setiap langkah yang dilakukan AlphaGo di papan Go, ia melakukan jutaan atau miliaran unit komputasi, yang masing-masing melibatkan penambahan cabang ke pohon pencarian lookahead dan mengevaluasi posisi papan di ujung cabang tersebut. Dan setiap unit komputasi tersebut terjadi karena program membuat pilihan tentang bagian pohon mana yang akan dieksplorasi selanjutnya. Secara kasar, AlphaGo memilih komputasi yang diharapkan akan meningkatkan keputusan akhirnya di papan.

Telah dimungkinkan untuk menyusun skema yang masuk akal untuk mengelola aktivitas komputasi AlphaGo karena aktivitas tersebut sederhana dan homogen: setiap unit komputasi memiliki jenis yang sama. Dibandingkan dengan program lain yang menggunakan unit komputasi dasar yang sama, AlphaGo mungkin cukup efisien, tetapi mungkin sangat tidak efisien dibandingkan dengan jenis program lain. Misalnya, Lee Sedol, lawan manusia AlphaGo dalam pertandingan bersejarah tahun 2016, mungkin hanya melakukan beberapa ribu unit komputasi per langkah, tetapi ia memiliki arsitektur komputasi yang jauh lebih fleksibel dengan lebih banyak jenis unit komputasi: ini termasuk membagi papan menjadi subgame dan mencoba menyelesaikan interaksi mereka; mengenali kemungkinan tujuan yang ingin dicapai dan membuat rencana tingkat tinggi dengan tindakan seperti "pertahankan kelompok ini tetap hidup" atau "cegah lawan saya menghubungkan kedua kelompok ini"; memikirkan cara mencapai tujuan tertentu, seperti menjaga agar suatu kelompok tetap hidup; dan mengesampingkan seluruh kelas gerakan karena gagal mengatasi ancaman yang signifikan.

Kita sama sekali tidak tahu bagaimana mengatur aktivitas komputasi yang kompleks dan beragam seperti itu bagaimana mengintegrasikan dan membangun hasil dari masing-masing aktivitas dan bagaimana mengalokasikan sumber daya komputasi untuk berbagai jenis pertimbangan sehingga keputusan yang baik dapat ditemukan secepat mungkin. Namun, jelas bahwa arsitektur komputasi sederhana seperti AlphaGo tidak mungkin berfungsi di dunia nyata, di mana kita secara rutin perlu berurusan dengan cakupan keputusan yang bukan puluhan tetapi miliaran langkah primitif dan di mana jumlah tindakan yang mungkin pada titik mana pun hampir tak terbatas. Penting untuk diingat bahwa agen cerdas di dunia nyata tidak terbatas pada bermain Go atau bahkan menemukan kunci Stuart ia hanya ada. Ia dapat melakukan apa pun selanjutnya, tetapi ia tidak mungkin mampu memikirkan semua hal yang mungkin dilakukannya.

Sistem yang dapat menemukan tindakan tingkat tinggi baru seperti yang dijelaskan sebelumnya dan mengelola aktivitas komputasinya untuk fokus pada unit komputasi yang dengan cepat memberikan peningkatan signifikan dalam kualitas keputusan akan menjadi

pengambil keputusan yang tangguh di dunia nyata. Seperti halnya manusia, pertimbangannya akan "efisien secara kognitif," tetapi tidak akan mengalami keterbatasan memori jangka pendek dan perangkat keras yang lambat yang sangat membatasi kemampuan kita untuk melihat jauh ke masa depan, menangani sejumlah besar kemungkinan, dan mempertimbangkan sejumlah besar rencana alternatif.

Apakah Masih Ada Hal Lain yang Hilang?

Jika kita menggabungkan semua yang kita ketahui dengan semua potensi perkembangan baru yang tercantum dalam bab ini, apakah itu akan berhasil? Bagaimana sistem yang dihasilkan akan berperilaku? Sistem tersebut akan terus bergerak maju, menyerap sejumlah besar informasi dan melacak keadaan dunia dalam skala besar melalui pengamatan dan inferensi. Sistem tersebut akan secara bertahap meningkatkan model dunianya (yang tentu saja mencakup model manusia). Sistem tersebut akan menggunakan model-model tersebut untuk memecahkan masalah kompleks dan akan merangkum serta menggunakan kembali proses penyelesaiannya untuk membuat pertimbangannya lebih efisien dan memungkinkan penyelesaian masalah yang lebih kompleks lagi. Sistem tersebut akan menemukan konsep dan tindakan baru, dan ini akan memungkinkan sistem tersebut untuk meningkatkan laju penemuannya. Sistem tersebut akan membuat rencana yang efektif dalam skala waktu yang semakin panjang.

Singkatnya, tidak jelas bahwa ada hal penting lain yang hilang, dari sudut pandang sistem yang efektif dalam mencapai tujuannya. Tentu saja, satu-satunya cara untuk memastikannya adalah dengan membangunnya (setelah terobosan tercapai) dan melihat apa yang terjadi.

3.4 MEMBAYANGKAN MESIN SUPERCERDAS

Komunitas teknis telah mengalami kegagalan imajinasi ketika membahas sifat dan dampak AI supercerdas. Seringkali, kita melihat diskusi tentang pengurangan kesalahan medis, mobil yang lebih aman, atau kemajuan lain yang bersifat bertahap. Robot dibayangkan sebagai entitas individu yang membawa otak mereka sendiri, padahal kenyataannya mereka kemungkinan besar terhubung secara nirkabel menjadi satu entitas global yang memanfaatkan sumber daya komputasi stasioner yang sangat besar. Seolah-olah para peneliti takut untuk memeriksa konsekuensi nyata dari keberhasilan AI.

Sistem cerdas serbaguna, berdasarkan asumsi, dapat melakukan apa pun yang dapat dilakukan manusia. Misalnya, beberapa manusia melakukan banyak matematika, desain algoritma, pengkodean, dan penelitian empiris untuk menghasilkan mesin pencari modern. Hasil dari semua pekerjaan ini sangat berguna dan tentu saja sangat berharga. Seberapa berharga? Sebuah studi baru-baru ini menunjukkan bahwa rata-rata orang dewasa Amerika yang disurvei perlu dibayar setidaknya Rp 175 juta untuk berhenti menggunakan mesin pencari selama setahun, yang diterjemahkan menjadi nilai global dalam puluhan triliun rupiah.

Sekarang bayangkan mesin pencari belum ada karena pekerjaan yang diperlukan selama beberapa dekade belum selesai, tetapi Anda memiliki akses ke sistem AI supercerdas. Hanya dengan mengajukan pertanyaan, Anda sekarang memiliki akses ke teknologi mesin

pencari, berkat sistem AI tersebut. Selesai! Triliunan dolar nilainya, hanya dengan meminta, dan tanpa satu baris kode tambahan pun yang Anda tulis. Hal yang sama berlaku untuk penemuan atau serangkaian penemuan lain yang belum ada: jika manusia dapat melakukannya, maka mesin pun dapat melakukannya.

Poin terakhir ini memberikan batasan bawah yang berguna, perkiraan pesimistis tentang apa yang dapat dilakukan oleh mesin supercerdas. Berdasarkan asumsi, mesin tersebut lebih mampu daripada manusia individu.

Ada banyak hal yang tidak dapat dilakukan oleh manusia individu, tetapi sekumpulan n manusia dapat melakukannya: menempatkan seorang astronot di Bulan, menciptakan detektor gelombang gravitasi, mengurutkan genom manusia, menjalankan sebuah negara dengan ratusan juta penduduk. Jadi, secara kasar, kita menciptakan n salinan perangkat lunak dari mesin tersebut dan menghubungkannya dengan cara yang sama dengan aliran informasi dan kontrol yang sama seperti n manusia. Sekarang kita memiliki mesin yang dapat melakukan apa pun yang dapat dilakukan oleh n manusia, kecuali lebih baik, karena setiap n komponennya adalah manusia super.

Desain kerja sama multi-agen untuk sistem cerdas ini hanyalah batas bawah dari kemampuan mesin yang mungkin karena ada desain lain yang bekerja lebih baik. Dalam kumpulan n manusia, total informasi yang tersedia disimpan secara terpisah di n otak dan dikomunikasikan sangat lambat dan tidak sempurna di antara mereka. Itulah mengapa manusia menghabiskan sebagian besar waktu mereka dalam rapat. Dalam mesin, tidak ada kebutuhan untuk pemisahan ini, yang seringkali mencegah menghubungkan titik-titik. Sebagai contoh titik-titik yang terputus dalam penemuan ilmiah, sekilas sejarah panjang penisilin cukup membuka mata.

Metode lain yang berguna untuk mengembangkan imajinasi Anda adalah dengan memikirkan beberapa bentuk masukan sensorik tertentu misalnya, membaca dan memperbesarnya. Sementara manusia dapat membaca dan memahami satu buku dalam seminggu, mesin dapat membaca dan memahami setiap buku yang pernah ditulis semuanya 150 juta dalam beberapa jam. Hal ini membutuhkan daya pemrosesan yang cukup besar, tetapi buku-buku tersebut dapat dibaca secara paralel, artinya dengan menambahkan lebih banyak chip, mesin tersebut dapat meningkatkan proses pembacaannya.

Dengan cara yang sama, mesin tersebut dapat melihat semuanya sekaligus melalui satelit, robot, dan ratusan juta kamera pengawas; menonton semua siaran TV di dunia; dan mendengarkan semua stasiun radio dan percakapan telepon di dunia. Dengan sangat cepat, mesin tersebut akan memperoleh pemahaman yang jauh lebih detail dan akurat tentang dunia dan penduduknya daripada yang dapat diharapkan oleh manusia mana pun.

Kita juga dapat membayangkan peningkatan kapasitas aksi mesin. Manusia hanya memiliki kendali langsung atas satu tubuh, sementara mesin dapat mengendalikan ribuan atau jutaan tubuh. Beberapa pabrik otomatis sudah menunjukkan karakteristik ini. Di luar pabrik, mesin yang mengendalikan ribuan robot yang cekatan, misalnya, dapat menghasilkan sejumlah besar rumah, masing-masing disesuaikan dengan kebutuhan dan keinginan penghuninya di masa depan.

Di laboratorium, sistem robotik yang ada untuk penelitian ilmiah dapat ditingkatkan skalanya untuk melakukan jutaan eksperimen secara bersamaan mungkin untuk menciptakan model prediktif lengkap biologi manusia hingga tingkat molekuler. Perhatikan bahwa kemampuan penalaran mesin akan memberinya kapasitas yang jauh lebih besar untuk mendeteksi inkonsistensi antara teori ilmiah dan antara teori dan pengamatan. Memang, mungkin sudah ada cukup bukti eksperimental tentang biologi untuk merancang obat untuk kanker: kita hanya belum menggabungkannya.

Di dunia maya, mesin sudah memiliki akses ke miliaran efektor yaitu, layar pada semua ponsel dan komputer di dunia. Ini sebagian menjelaskan kemampuan perusahaan TI untuk menghasilkan kekayaan yang sangat besar dengan sangat sedikit karyawan; Hal ini juga menunjukkan kerentanan serius umat manusia terhadap manipulasi melalui layar.

Skala yang berbeda muncul dari kemampuan mesin untuk melihat lebih jauh ke masa depan, dengan akurasi yang lebih tinggi, daripada yang mungkin dilakukan manusia. Kita telah melihat ini pada catur dan Go; dengan kapasitas untuk menghasilkan dan menganalisis rencana hierarkis dalam skala waktu yang panjang dan kemampuan untuk mengidentifikasi tindakan abstrak baru dan model deskriptif tingkat tinggi, mesin akan mentransfer keunggulan ini ke domain seperti matematika (membuktikan teorema baru yang bermanfaat) dan pengambilan keputusan di dunia nyata. Tugas-tugas seperti mengevakuasi kota besar jika terjadi bencana lingkungan akan relatif mudah, dengan mesin mampu menghasilkan panduan individual untuk setiap orang dan kendaraan untuk meminimalkan jumlah korban.

Mesin mungkin akan sedikit berkeringat saat merancang rekomendasi kebijakan untuk mencegah pemanasan global. Pemodelan sistem bumi membutuhkan pengetahuan tentang fisika (atmosfer, lautan), kimia (siklus karbon, tanah), biologi (dekomposisi, migrasi), teknik (energi terbarukan, penangkapan karbon), ekonomi (industri, penggunaan energi), sifat manusia (kebodohan, keserakahan), dan politik (lebih banyak kebodohan, lebih banyak keserakahan).

Seperti yang telah disebutkan, mesin tersebut akan memiliki akses ke sejumlah besar bukti untuk memberi makan semua model ini. Mesin tersebut akan mampu menyarankan atau melakukan eksperimen dan ekspedisi baru untuk mempersempit ketidakpastian yang tak terhindarkan misalnya, untuk menemukan sejauh mana keberadaan hidrat gas di reservoir laut dangkal. Mesin tersebut akan mampu mempertimbangkan berbagai macam rekomendasi kebijakan yang mungkin hukum, dorongan, pasar, penemuan, dan intervensi geoteknik tetapi tentu saja mesin tersebut juga perlu menemukan cara untuk membujuk kita agar menerimanya.

3.5 BATASAN KECERDASAN SUPER

Saat Anda mengembangkan imajinasi, jangan terlalu jauh. Kesalahan umum adalah mengaitkan kekuatan seperti dewa berupa kemahatahuan kepada sistem AI supercerdas pengetahuan lengkap dan sempurna tidak hanya tentang masa kini tetapi juga tentang masa depan. Hal ini cukup tidak masuk akal karena membutuhkan kemampuan yang tidak fisik untuk menentukan keadaan dunia saat ini secara tepat serta kemampuan yang tidak mungkin

diwujudkan untuk mensimulasikan, jauh lebih cepat daripada waktu nyata, operasi dunia yang mencakup mesin itu sendiri (belum lagi miliaran otak, yang masih akan menjadi objek paling kompleks kedua di alam semesta).

Ini bukan berarti tidak mungkin untuk memprediksi beberapa aspek masa depan dengan tingkat kepastian yang wajar misalnya, saya tahu kelas apa yang akan saya ajarkan di ruangan mana di Berkeley hampir setahun dari sekarang, terlepas dari protes para ahli teori kekacauan tentang sayap kupu-kupu dan semua itu. (Saya juga tidak berpikir bahwa manusia mendekati kemampuan memprediksi masa depan sebaik yang diizinkan oleh hukum fisika!) Prediksi bergantung pada abstraksi yang tepat misalnya, saya dapat memprediksi bahwa "saya" akan "berada di atas panggung di Auditorium Wheeler" di kampus Berkeley pada hari Selasa terakhir bulan April, tetapi saya tidak dapat memprediksi lokasi tepat saya hingga milimeter atau atom karbon mana yang telah dimasukkan ke dalam tubuh saya pada saat itu.

Mesin juga tunduk pada batasan kecepatan tertentu yang diberlakukan oleh dunia nyata pada tingkat perolehan pengetahuan baru tentang dunia salah satu poin valid yang dikemukakan oleh Kevin Kelly dalam artikelnya tentang prediksi yang terlalu disederhanakan tentang AI supermanusia. Misalnya, untuk menentukan apakah obat tertentu menyembuhkan jenis kanker tertentu pada hewan percobaan, seorang ilmuwan manusia atau mesin memiliki dua pilihan: menyuntikkan obat ke hewan tersebut dan menunggu beberapa minggu atau menjalankan simulasi yang cukup akurat. Namun, untuk menjalankan simulasi membutuhkan banyak pengetahuan empiris tentang biologi, beberapa di antaranya saat ini tidak tersedia; jadi, lebih banyak eksperimen pembuatan model harus dilakukan terlebih dahulu. Tidak diragukan lagi, hal ini akan membutuhkan waktu dan harus dilakukan di dunia nyata.

Di sisi lain, seorang ilmuwan mesin dapat menjalankan sejumlah besar eksperimen pembuatan model secara paralel, dapat mengintegrasikan hasilnya ke dalam model yang konsisten secara internal (walaupun sangat kompleks), dan dapat membandingkan prediksi model dengan keseluruhan bukti eksperimental yang diketahui dalam biologi. Selain itu, mensimulasikan model tidak selalu memerlukan simulasi mekanika kuantum dari seluruh organisme hingga tingkat reaksi molekuler individual yang, seperti yang ditunjukkan Kelly, akan memakan waktu lebih lama daripada sekadar melakukan eksperimen di dunia nyata.

Sama seperti saya dapat memprediksi lokasi saya di masa depan pada hari Selasa di bulan April dengan cukup pasti, sifat-sifat sistem biologis dapat diprediksi secara akurat dengan model abstrak. (Di antara alasan lainnya, ini karena biologi beroperasi dengan sistem kontrol yang kuat berdasarkan loop umpan balik agregat, sehingga variasi kecil dalam kondisi awal biasanya tidak menyebabkan variasi besar dalam hasil.) Dengan demikian, meskipun penemuan mesin secara instan dalam ilmu empiris tidak mungkin terjadi, kita dapat mengharapkan bahwa sains akan berjalan jauh lebih cepat dengan bantuan mesin. Memang, hal itu sudah terjadi.

Keterbatasan terakhir dari mesin adalah bahwa mereka bukanlah manusia. Hal ini menempatkan mereka pada posisi yang tidak menguntungkan secara intrinsik ketika mencoba memodelkan dan memprediksi satu kelas objek tertentu: manusia. Otak kita semua cukup mirip, jadi kita dapat menggunakannya untuk mensimulasikan untuk mengalami, jika Anda

mau kehidupan mental dan emosional orang lain. Bagi kita, ini gratis. (Jika Anda memikirkannya, mesin memiliki keuntungan yang lebih besar satu sama lain: mereka sebenarnya dapat menjalankan kode satu sama lain!) Misalnya, saya tidak perlu menjadi ahli sistem sensorik saraf untuk mengetahui bagaimana rasanya ketika Anda memukul ibu jari Anda dengan palu. Saya hanya bisa memukul ibu jari saya dengan palu. Mesin, di sisi lain, harus memulai hampir dari awal dalam pemahaman mereka tentang manusia: mereka hanya memiliki akses ke perilaku eksternal kita, ditambah semua literatur ilmu saraf dan psikologi, dan harus mengembangkan pemahaman tentang bagaimana kita bekerja berdasarkan hal itu. Pada prinsipnya, mereka akan mampu melakukan ini, tetapi masuk akal untuk berasumsi bahwa memperoleh pemahaman tingkat manusia atau supermanusia tentang manusia akan membutuhkan waktu lebih lama daripada sebagian besar kemampuan lainnya.

3.6 BAGAIMANA AI AKAN MENGUNTUNGKAN MANUSIA?

Kecerdasan kita bertanggung jawab atas peradaban kita. Dengan akses ke kecerdasan yang lebih besar, kita dapat memiliki peradaban yang lebih hebat dan mungkin jauh lebih baik. Kita dapat berspekulasi tentang memecahkan masalah-masalah besar yang belum terselesaikan seperti memperpanjang umur manusia tanpa batas atau mengembangkan perjalanan yang lebih cepat dari cahaya, tetapi hal-hal pokok dalam fiksi ilmiah ini belum menjadi kekuatan pendorong kemajuan dalam AI. (Dengan AI supercerdas, kita mungkin akan dapat menciptakan berbagai macam teknologi yang hampir ajaib, tetapi sulit untuk mengatakan sekarang apa saja teknologi tersebut.)

Pertimbangkan, sebagai gantinya, tujuan yang jauh lebih sederhana: meningkatkan standar hidup setiap orang di Bumi, dengan cara yang berkelanjutan, ke tingkat yang akan dianggap cukup terhormat di negara maju. Dengan memilih (agak sewenang-wenang) istilah "terhormat" yang berarti persentil ke-88 di Amerika Serikat, tujuan yang dinyatakan tersebut mewakili peningkatan hampir sepuluh kali lipat dalam produk domestik bruto (PDB) global, dari Rp 760 kuadriliun menjadi Rp 7,5 kuintiliun per tahun.

Untuk menghitung nilai tunai dari hadiah tersebut, para ekonom menggunakan nilai sekarang bersih dari aliran pendapatan, yang memperhitungkan diskonto pendapatan masa depan relatif terhadap masa kini. Pendapatan tambahan sebesar Rp 6,74 kuintiliun per tahun memiliki nilai sekarang bersih sekitar Rp 135 kuintiliun, dengan asumsi faktor diskonto 5 persen. Jadi, secara kasar, ini adalah perkiraan kasar tentang berapa nilai AI tingkat manusia jika dapat memberikan standar hidup yang layak bagi semua orang.

Dengan angka-angka seperti ini, tidak mengherankan jika perusahaan dan negara menginvestasikan puluhan miliar dolar setiap tahunnya dalam penelitian dan pengembangan AI. Meskipun demikian, jumlah yang diinvestasikan sangat kecil dibandingkan dengan besarnya hadiah tersebut.

Tentu saja, semua angka ini hanyalah angka fiktif kecuali jika seseorang memiliki gagasan tentang bagaimana AI tingkat manusia dapat mencapai prestasi meningkatkan standar hidup. AI hanya dapat melakukan ini dengan meningkatkan produksi barang dan jasa per kapita. Dengan kata lain: manusia rata-rata tidak akan pernah dapat mengharapkan untuk

mengonsumsi lebih banyak daripada yang diproduksi manusia rata-rata. Contoh taksi swakemudi yang dibahas sebelumnya dalam bab ini menggambarkan efek pengganda AI: dengan layanan otomatis, seharusnya memungkinkan bagi (katakanlah) sepuluh orang untuk mengelola armada seribu kendaraan, sehingga setiap orang menghasilkan seratus kali lebih banyak transportasi daripada sebelumnya. Hal yang sama berlaku untuk pembuatan mobil dan untuk ekstraksi bahan baku dari mana mobil tersebut dibuat. Bahkan, beberapa operasi penambangan bijih besi di Australia utara, di mana suhu secara teratur melebihi 45 derajat Celcius (113 derajat Fahrenheit), hampir sepenuhnya otomatis.

Aplikasi AI saat ini adalah sistem tujuan khusus: mobil swakemudi dan tambang yang beroperasi sendiri membutuhkan investasi besar dalam penelitian, desain mekanik, rekayasa perangkat lunak, dan pengujian untuk mengembangkan algoritma yang diperlukan dan untuk memastikan bahwa algoritma tersebut bekerja sesuai tujuan. Begitulah cara kerja di semua bidang teknik.

Begitulah cara kerja perjalanan pribadi di masa lalu: jika Anda ingin bepergian dari Eropa ke Australia dan kembali pada abad ketujuh belas, itu akan melibatkan proyek besar yang menghabiskan banyak uang, membutuhkan perencanaan bertahun-tahun, dan membawa risiko kematian yang tinggi. Sekarang kita terbiasa dengan gagasan transportasi sebagai layanan (TaaS): jika Anda perlu berada di Melbourne awal minggu depan, itu hanya membutuhkan beberapa ketukan di ponsel Anda dan sejumlah uang yang relatif kecil.

AI serbaguna akan menjadi segalanya sebagai layanan (EaaS). Tidak perlu lagi mempekerjakan banyak spesialis dalam berbagai disiplin ilmu, yang diorganisir ke dalam hierarki kontraktor dan subkontraktor, untuk melaksanakan suatu proyek. Semua perwujudan AI serbaguna akan memiliki akses ke semua pengetahuan dan keterampilan umat manusia, dan lebih dari itu. Satu-satunya perbedaan terletak pada kemampuan fisik: robot berkaki lincah untuk konstruksi atau pembedahan, robot beroda untuk transportasi barang skala besar, robot quadcopter untuk inspeksi udara, dan sebagainya.

Pada prinsipnya terlepas dari politik dan ekonomi, setiap orang dapat memiliki seluruh organisasi yang terdiri dari agen perangkat lunak dan robot fisik, yang mampu merancang dan membangun jembatan, meningkatkan hasil panen, memasak makan malam untuk seratus tamu, menjalankan pemilihan, atau melakukan apa pun yang perlu dilakukan. Kecerdasan serbaguna inilah yang memungkinkan hal ini.

Sejarah telah menunjukkan, tentu saja, bahwa peningkatan sepuluh kali lipat PDB per kapita global dimungkinkan tanpa AI hanya saja dibutuhkan 190 tahun (dari 1820 hingga 2010) untuk mencapai peningkatan tersebut. Hal ini membutuhkan pengembangan pabrik, peralatan mesin, otomatisasi, kereta api, baja, mobil, pesawat terbang, listrik, produksi minyak dan gas, telepon, radio, televisi, komputer, internet, satelit, dan banyak penemuan revolusioner lainnya. Peningkatan sepuluh kali lipat PDB yang dikemukakan dalam paragraf sebelumnya didasarkan bukan pada teknologi revolusioner lebih lanjut, tetapi pada kemampuan sistem AI untuk menggunakan apa yang sudah kita miliki secara lebih efektif dan dalam skala yang lebih besar.

Tentu saja, akan ada efek selain manfaat materiil semata dari peningkatan standar hidup. Misalnya, bimbingan pribadi diketahui jauh lebih efektif daripada pengajaran di kelas, tetapi jika dilakukan oleh manusia, hal itu tidak terjangkau dan akan selalu demikian bagi sebagian besar orang. Dengan tutor AI, potensi setiap anak, betapapun miskinnya, dapat diwujudkan. Biaya per anak akan sangat kecil, dan anak tersebut akan menjalani kehidupan yang jauh lebih kaya dan produktif. Mengejar usaha artistik dan intelektual, baik secara individu maupun kolektif, akan menjadi bagian normal dari kehidupan daripada kemewahan yang langka.

Di bidang kesehatan, sistem AI seharusnya memungkinkan para peneliti untuk mengungkap dan menguasai kompleksitas biologi manusia yang luas dan dengan demikian secara bertahap menghilangkan penyakit. Wawasan yang lebih besar tentang psikologi dan neurokimia manusia seharusnya mengarah pada peningkatan kesehatan mental yang luas. Mungkin lebih tidak konvensional, AI dapat memungkinkan alat pembuatan konten yang jauh lebih efektif untuk realitas virtual (VR) dan dapat mengisi lingkungan VR dengan entitas yang jauh lebih menarik. Ini mungkin mengubah VR menjadi media pilihan untuk ekspresi sastra dan artistik, menciptakan pengalaman yang kaya dan mendalam yang saat ini tidak terbayangkan.

Dan dalam dunia kehidupan sehari-hari yang biasa, asisten dan pemandu yang cerdas jika dirancang dengan baik dan tidak dimanfaatkan oleh kepentingan ekonomi dan politik akan memberdayakan setiap individu untuk bertindak secara efektif atas nama mereka sendiri dalam sistem ekonomi dan politik yang semakin kompleks dan terkadang bermusuhan. Pada dasarnya, Anda akan memiliki pengacara, akuntan, dan penasihat politik kelas atas yang siap sedia kapan saja. Sama seperti kemacetan lalu lintas yang diharapkan dapat diatasi dengan mencampurkan sebagian kecil kendaraan otonom, kita hanya dapat berharap bahwa kebijakan yang lebih bijaksana dan konflik yang lebih sedikit akan muncul dari warga global yang lebih terinformasi dan lebih terbimbing.

Perkembangan ini secara bersama-sama dapat mengubah dinamika sejarah setidaknya bagian sejarah yang telah didorong oleh konflik di dalam dan antar masyarakat untuk mendapatkan akses ke kebutuhan hidup. Jika kue itu pada dasarnya tak terbatas, maka memperebutkan bagian yang lebih besar tidak masuk akal. Itu akan seperti memperebutkan siapa yang mendapatkan salinan digital koran terbanyak sama sekali tidak ada gunanya ketika siapa pun dapat membuat salinan digital sebanyak yang mereka inginkan secara gratis.

Ada beberapa batasan untuk apa yang dapat diberikan AI. Kue tanah dan bahan mentah tidak tak terbatas, jadi tidak mungkin ada pertumbuhan populasi yang tak terbatas dan tidak semua orang akan memiliki rumah mewah di taman pribadi. (Hal ini pada akhirnya akan memerlukan penambangan di tempat lain di tata surya dan pembangunan habitat buatan di luar angkasa; tetapi saya berjanji untuk tidak membicarakan fiksi ilmiah.) Kebanggaan juga terbatas: hanya 1 persen orang yang dapat berada di 1 persen teratas dalam metrik tertentu. Jika kebahagiaan manusia membutuhkan berada di 1 persen teratas, maka 99 persen manusia akan tidak bahagia, bahkan ketika 1 persen terbawah memiliki gaya hidup yang secara objektif

luar biasa. Oleh karena itu, penting bagi budaya kita untuk secara bertahap mengurangi bobot kebanggaan dan iri hati sebagai elemen sentral dari harga diri yang dirasakan.

Seperti yang dikatakan Nick Bostrom di akhir bukunya Superintelligence, keberhasilan dalam AI akan menghasilkan "lintasan peradaban yang mengarah pada penggunaan yang penuh kasih sayang dan sukacita atas anugerah kosmik umat manusia." Jika kita gagal memanfaatkan apa yang ditawarkan AI, kita hanya akan menyalahkan diri sendiri.

BAB 4
PENYALAHGUNAAN AI

Penggunaan yang penuh kasih sayang dan sukacita atas anugerah kosmik umat manusia terdengar indah, tetapi kita juga harus memperhitungkan laju inovasi yang cepat di sektor kejahatan. Orang-orang yang berniat jahat memikirkan cara-cara baru untuk menyalahgunakan AI dengan begitu cepat sehingga bab ini kemungkinan akan menjadi usang bahkan sebelum dicetak. Namun, anggaplah ini bukan sebagai bacaan yang menyedihkan, melainkan sebagai seruan untuk bertindak sebelum terlambat.

4.1 PENGAWASAN, PERSUASI, DAN KONTROL

Stasi Otomatis

Ministerium für Staatsicherheit Jerman Timur, yang lebih dikenal sebagai Stasi, secara luas dianggap sebagai "salah satu badan intelijen dan polisi rahasia yang paling efektif dan represif yang pernah ada." Mereka menyimpan arsip tentang sebagian besar rumah tangga Jerman Timur. Mereka memantau panggilan telepon, membaca surat, dan memasang kamera tersembunyi di apartemen dan hotel.

Mereka sangat efektif dalam mengidentifikasi dan melenyapkan aktivitas pembangkang. Modus operandi pilihan mereka adalah penghancuran psikologis daripada pemenjaraan atau eksekusi. Namun, tingkat kontrol ini datang dengan biaya yang sangat besar: menurut beberapa perkiraan, lebih dari seperempat orang dewasa usia kerja adalah informan Stasi. Catatan kertas Stasi diperkirakan mencapai dua puluh miliar halaman dan tugas memproses dan menindaklanjuti arus informasi yang sangat besar yang masuk mulai melebihi kapasitas organisasi manusia mana pun.

Oleh karena itu, tidak mengherankan jika badan intelijen telah melihat potensi penggunaan AI dalam pekerjaan mereka. Selama bertahun-tahun, mereka telah menerapkan bentuk-bentuk sederhana teknologi AI, termasuk pengenalan suara dan identifikasi kata kunci dan frasa dalam ucapan dan teks. Semakin banyak, sistem AI mampu memahami isi dari apa yang dikatakan dan dilakukan orang, baik dalam ucapan, teks, atau pengawasan video. Dalam rezim di mana teknologi ini diadopsi untuk tujuan kontrol, akan seperti setiap warga negara memiliki agen Stasi pribadi yang mengawasi mereka dua puluh empat jam sehari.

Bahkan di ranah sipil, di negara-negara yang relatif bebas, kita tunduk pada pengawasan yang semakin efektif. Perusahaan mengumpulkan dan menjual informasi tentang pembelian kita, penggunaan internet dan jejaring sosial, penggunaan peralatan listrik, catatan panggilan dan pesan teks, pekerjaan, dan kesehatan. Lokasi kita dapat dilacak melalui ponsel dan mobil kita yang terhubung ke internet. Kamera mengenali wajah kita di jalan. Semua data ini, dan masih banyak lagi, dapat disatukan oleh sistem integrasi informasi cerdas untuk menghasilkan gambaran yang cukup lengkap tentang apa yang dilakukan setiap orang, bagaimana kita menjalani hidup, siapa yang kita sukai dan tidak sukai, dan bagaimana kita akan memilih. Stasi akan terlihat seperti amatir jika dibandingkan.

Mengendalikan Perilaku Anda

Setelah kemampuan pengawasan diterapkan, langkah selanjutnya adalah memodifikasi perilaku Anda agar sesuai dengan mereka yang menggunakan teknologi ini. Salah satu metode yang agak kasar adalah pemerasan otomatis dan personal: sebuah sistem yang memahami apa yang Anda lakukan baik dengan mendengarkan, membaca, atau mengamati Anda dapat dengan mudah menemukan hal-hal yang seharusnya tidak Anda lakukan. Setelah menemukan sesuatu, sistem tersebut akan berkomunikasi dengan Anda untuk mendapatkan uang sebanyak mungkin (atau untuk memaksa perilaku, jika tujuannya adalah kontrol politik atau spionase).

Pengambilan uang berfungsi sebagai sinyal penghargaan yang sempurna untuk algoritma pembelajaran penguatan, sehingga kita dapat mengharapkan sistem AI untuk meningkat pesat dalam kemampuannya untuk mengidentifikasi dan mengambil keuntungan dari perilaku yang salah. Pada awal tahun 2015, saya menyarankan kepada seorang ahli keamanan komputer bahwa sistem pemerasan otomatis, yang didorong oleh pembelajaran penguatan, mungkin akan segera menjadi layak; dia tertawa dan mengatakan itu sudah terjadi. Bot pemerasan pertama yang dipublikasikan secara luas adalah Delilah, yang diidentifikasi pada Juli 2016.

Cara yang lebih halus untuk mengubah perilaku orang adalah dengan memodifikasi lingkungan informasi mereka sehingga mereka mempercayai hal-hal yang berbeda dan membuat keputusan yang berbeda. Tentu saja, para pengiklan telah melakukan ini selama berabad-abad sebagai cara untuk memodifikasi perilaku pembelian individu. Propaganda sebagai alat perang dan dominasi politik memiliki sejarah yang lebih panjang lagi.

Jadi, apa yang berbeda sekarang? Pertama, karena sistem AI dapat melacak kebiasaan membaca online, preferensi, dan kemungkinan tingkat pengetahuan individu, mereka dapat menyesuaikan pesan-pesan tertentu untuk memaksimalkan dampak pada individu tersebut sambil meminimalkan risiko bahwa informasi tersebut akan tidak dipercaya. Kedua, sistem AI mengetahui apakah individu tersebut membaca pesan, berapa lama mereka menghabiskan waktu untuk membacanya, dan apakah mereka mengikuti tautan tambahan di dalam pesan tersebut. Kemudian, sistem tersebut menggunakan sinyal-sinyal ini sebagai umpan balik langsung tentang keberhasilan atau kegagalan upayanya untuk memengaruhi setiap individu; dengan cara ini, sistem tersebut dengan cepat belajar untuk menjadi lebih efektif dalam pekerjaannya. Inilah bagaimana algoritma pemilihan konten di media sosial memiliki efek yang berbahaya pada opini politik.

Perubahan terbaru lainnya adalah bahwa kombinasi AI, grafis komputer, dan sintesis ucapan memungkinkan untuk menghasilkan deepfake konten video dan audio realistis dari hampir siapa pun yang mengatakan atau melakukan hampir apa pun. Teknologi ini hanya membutuhkan deskripsi verbal tentang peristiwa yang diinginkan, sehingga dapat digunakan oleh hampir semua orang di dunia. Video ponsel Senator X menerima suap dari pengedar kokain Y di tempat mencurigakan Z? Tidak masalah! Konten semacam ini dapat menimbulkan keyakinan yang tak tergoyahkan pada hal-hal yang tidak pernah terjadi. Selain itu, sistem AI dapat menghasilkan jutaan identitas palsu yang disebut pasukan bot yang dapat menghasilkan

miliaran komentar, tweet, dan rekomendasi setiap hari, membanjiri upaya manusia untuk bertukar informasi yang benar. Pasar online seperti eBay, Taobao, dan Amazon yang mengandalkan sistem reputasi untuk membangun kepercayaan antara pembeli dan penjual terus-menerus berperang dengan pasukan bot yang dirancang untuk merusak pasar.

Terakhir, metode pengendalian dapat dilakukan secara langsung jika pemerintah mampu menerapkan imbalan dan hukuman berdasarkan perilaku. Sistem seperti itu memperlakukan orang sebagai algoritma pembelajaran penguatan, melatih mereka untuk mengoptimalkan tujuan yang ditetapkan oleh negara. Godaan bagi pemerintah, khususnya pemerintah dengan pola pikir top-down dan berorientasi rekayasa, adalah untuk bernalar sebagai berikut: akan lebih baik jika semua orang berperilaku baik, memiliki sikap patriotik, dan berkontribusi pada kemajuan negara; teknologi memungkinkan pengukuran perilaku, sikap, dan kontribusi individu; oleh karena itu, semua orang akan lebih baik jika kita membangun sistem pemantauan dan pengendalian berbasis teknologi berdasarkan penghargaan dan hukuman.

Ada beberapa masalah dengan cara berpikir ini. Pertama, hal itu mengabaikan biaya psikologis hidup di bawah sistem pengawasan dan paksaan yang mengganggu; harmoni lahiriah yang menutupi penderitaan batin bukanlah keadaan yang ideal. Setiap tindakan kebaikan berhenti menjadi tindakan kebaikan dan malah menjadi tindakan memaksimalkan skor pribadi dan dianggap demikian oleh penerima. Atau lebih buruk lagi, konsep tindakan kebaikan sukarela secara bertahap menjadi hanya kenangan yang memudar tentang sesuatu yang biasa dilakukan orang.

Mengunjungi teman yang sakit di rumah sakit, di bawah sistem seperti itu, tidak akan memiliki makna moral dan nilai emosional lebih dari berhenti di lampu merah. Kedua, skema ini menjadi korban mode kegagalan yang sama dengan model standar AI, karena mengasumsikan bahwa tujuan yang dinyatakan sebenarnya adalah tujuan mendasar yang sebenarnya.

Tak pelak lagi, hukum Goodhart akan berlaku, di mana individu mengoptimalkan ukuran resmi perilaku lahiriah, sama seperti universitas telah belajar untuk mengoptimalkan ukuran "objektif" dari "kualitas" yang digunakan oleh sistem pemeringkatan universitas alih-alih meningkatkan kualitas nyata (tetapi tidak terukur) mereka. Akhirnya, penerapan ukuran seragam kebajikan perilaku mengabaikan poin bahwa masyarakat yang sukses dapat terdiri dari berbagai macam individu, masing-masing berkontribusi dengan caranya sendiri.

Hak Atas Keamanan Mental

Salah satu pencapaian besar peradaban adalah peningkatan bertahap dalam keamanan fisik bagi manusia. Sebagian besar dari kita dapat berharap untuk menjalani kehidupan sehari-hari tanpa rasa takut yang terus-menerus akan cedera dan kematian. Pasal 3 Deklarasi Universal Hak Asasi Manusia tahun 1948 menyatakan, “Setiap orang berhak atas kehidupan, kebebasan, dan keamanan pribadi.”

Saya ingin menyarankan bahwa setiap orang juga harus memiliki hak atas keamanan mental hak untuk hidup dalam lingkungan informasi yang sebagian besar benar. Manusia cenderung mempercayai bukti dari mata dan telinga kita. Kita mempercayai keluarga, teman,

guru, dan (beberapa) sumber media untuk memberi tahu kita apa yang mereka yakini sebagai kebenaran. Meskipun kita tidak mengharapkan penjual mobil bekas dan politisi untuk mengatakan yang sebenarnya kepada kita, kita kesulitan mempercayai bahwa mereka berbohong dengan begitu berani seperti yang kadang-kadang mereka lakukan. Oleh karena itu, kita sangat rentan terhadap teknologi misinformasi.

Hak atas keamanan mental tampaknya tidak diabadikan dalam Deklarasi Universal. Pasal 18 dan 19 menetapkan hak "kebebasan berpikir" dan "kebebasan berpendapat dan berekspresi." Pikiran dan pendapat seseorang, tentu saja, sebagian dibentuk oleh lingkungan informasi seseorang, yang pada gilirannya tunduk pada Pasal 19 tentang "hak untuk... menyampaikan informasi dan gagasan melalui media apa pun dan tanpa memandang batas negara." Artinya, siapa pun, di mana pun di dunia, berhak untuk menyampaikan informasi palsu kepada Anda. Di situlah letak kesulitannya: negara-negara demokrasi, khususnya Amerika Serikat, sebagian besar enggan atau secara konstitusional tidak mampu untuk mencegah penyebaran informasi palsu tentang masalah kepentingan publik karena kekhawatiran yang dapat dibenarkan mengenai kontrol pemerintah terhadap kebebasan berbicara.

Alih-alih mengejar gagasan bahwa tidak ada kebebasan berpikir tanpa akses ke informasi yang benar, demokrasi tampaknya telah menaruh kepercayaan yang naif pada gagasan bahwa kebenaran akan menang pada akhirnya, dan kepercayaan ini telah membuat kita tidak terlindungi. Jerman adalah pengecualian; baru-baru ini negara itu mengesahkan Undang-Undang Penegakan Jaringan, yang mengharuskan platform konten untuk menghapus ujaran kebencian dan berita palsu yang dilarang, tetapi ini telah mendapat kritik yang cukup besar karena tidak dapat diterapkan dan tidak demokratis.

Untuk saat ini, kita dapat mengharapkan keamanan mental kita tetap diserang, terutama dilindungi oleh upaya komersial dan sukarela. Upaya-upaya ini termasuk situs pengecekan fakta seperti factcheck.org dan snopes.com tetapi tentu saja situs "pengecekan fakta" lainnya bermunculan untuk menyatakan kebenaran sebagai kebohongan dan kebohongan sebagai kebenaran.

Perusahaan penyedia informasi utama seperti Google dan Facebook telah berada di bawah tekanan ekstrem di Eropa dan Amerika Serikat untuk "melakukan sesuatu tentang hal itu." Mereka bereksperimen dengan cara untuk menandai atau menurunkan peringkat konten palsu menggunakan AI dan penyaring manusia dan untuk mengarahkan pengguna ke sumber terverifikasi yang menangkal efek misinformasi. Pada akhirnya, semua upaya tersebut bergantung pada sistem reputasi sirkular, dalam arti bahwa sumber dipercaya karena sumber tepercaya melaporkan bahwa sumber tersebut dapat dipercaya. Jika cukup banyak informasi palsu disebarkan, sistem reputasi ini dapat gagal: sumber yang sebenarnya dapat dipercaya dapat menjadi tidak tepercaya dan sebaliknya, seperti yang tampaknya terjadi saat ini dengan sumber media utama seperti CNN dan Fox News di Amerika Serikat. Aviv Ovadya, seorang ahli teknologi yang bekerja melawan misinformasi, menyebut ini sebagai "infopocalypse—kegagalan dahsyat pasar ide."

Salah satu cara untuk melindungi fungsi sistem reputasi adalah dengan memasukkan sumber yang sedekat mungkin dengan kebenaran sebenarnya. Satu fakta yang pasti benar dapat membatalkan sejumlah sumber yang hanya agak dapat dipercaya, jika sumber-sumber tersebut menyebarkan informasi yang bertentangan dengan fakta yang diketahui. Di banyak negara, notaris berfungsi sebagai sumber kebenaran sebenarnya untuk menjaga integritas informasi hukum dan real estat; mereka biasanya pihak ketiga yang tidak berkepentingan dalam transaksi apa pun dan dilisensikan oleh pemerintah atau perkumpulan profesional. (Di Kota London, Worshipful Company of Scriveners telah melakukan ini sejak tahun 1373, menunjukkan bahwa stabilitas tertentu melekat pada peran penyampaian kebenaran.) Jika standar formal, kualifikasi profesional, dan prosedur perizinan muncul untuk pemeriksa fakta, hal itu cenderung menjaga validitas aliran informasi yang kita andalkan. Organisasi seperti kelompok W3C Credible Web dan Credibility Coalition bertujuan untuk mengembangkan metode teknologi dan crowdsourcing untuk mengevaluasi penyedia informasi, yang kemudian memungkinkan pengguna untuk menyaring sumber yang tidak dapat diandalkan.

Cara kedua untuk melindungi sistem reputasi adalah dengan mengenakan biaya untuk penyebaran informasi palsu. Dengan demikian, beberapa situs peringkat hotel hanya menerima ulasan tentang hotel tertentu dari mereka yang telah memesan dan membayar kamar di hotel tersebut melalui situs tersebut, sementara situs peringkat lainnya menerima ulasan dari siapa pun. Tidak mengherankan jika peringkat di situs-situs pertama jauh kurang bias, karena mereka mengenakan biaya (membayar kamar hotel yang tidak perlu) untuk ulasan palsu. Sanksi regulasi lebih kontroversial: tidak ada yang menginginkan Kementerian Kebenaran, dan Undang-Undang Penegakan Jaringan Jerman hanya menghukum platform konten, bukan orang yang memposting berita palsu. Di sisi lain, sama seperti banyak negara dan banyak negara bagian AS yang melarang perekaman panggilan telepon tanpa izin, setidaknya harus dimungkinkan untuk mengenakan sanksi untuk pembuatan rekaman audio dan video fiktif dari orang sungguhan.

Terakhir, ada dua fakta lain yang menguntungkan kita. Pertama, hampir tidak ada seorang pun yang secara aktif ingin, dengan sadar, dibohongi. (Ini bukan berarti orang tua selalu menyelidiki dengan saksama kebenaran dari mereka yang memuji kecerdasan dan pesona anak-anak mereka; hanya saja mereka cenderung kurang mencari persetujuan seperti itu dari seseorang yang dikenal suka berbohong di setiap kesempatan.) Ini berarti bahwa orang-orang dari semua keyakinan politik memiliki insentif untuk mengadopsi alat yang membantu mereka membedakan kebenaran dari kebohongan. Kedua, tidak ada yang ingin dikenal sebagai pembohong, apalagi media berita. Ini berarti bahwa penyedia informasi setidaknya mereka yang reputasinya penting memiliki insentif untuk bergabung dengan asosiasi industri dan mengikuti kode etik yang mendukung kejujuran. Pada gilirannya, platform media sosial dapat menawarkan pengguna pilihan untuk melihat konten hanya dari sumber-sumber terpercaya yang mengikuti kode etik ini dan tunduk pada pengecekan fakta pihak ketiga.

4.2 SENJATA OTONOM MEMATIKAN

Perserikatan Bangsa-Bangsa mendefinisikan sistem senjata otonom mematikan (disingkat AWS, karena LAWS cukup membingungkan) sebagai sistem senjata yang "menemukan, memilih, dan melenyapkan target manusia tanpa campur tangan manusia." AWS telah digambarkan, dengan alasan yang kuat, sebagai "revolusi ketiga dalam peperangan," setelah bubuk mesiu dan senjata nuklir.

Anda mungkin pernah membaca artikel di media tentang AWS; biasanya artikel tersebut akan menyebutnya robot pembunuh dan dihiasi dengan gambar-gambar dari film Terminator. Ini menyesatkan setidaknya dalam dua hal: pertama, ini menunjukkan bahwa senjata otonom merupakan ancaman karena mereka mungkin mengambil alih dunia dan menghancurkan umat manusia; kedua, ini menunjukkan bahwa senjata otonom akan berbentuk manusia, sadar, dan jahat.

Dampak bersih dari penggambaran media tentang masalah ini adalah membuatnya tampak seperti fiksi ilmiah. Bahkan pemerintah Jerman pun tertipu: baru-baru ini mereka mengeluarkan pernyataan yang menegaskan bahwa "memiliki kemampuan untuk belajar dan mengembangkan kesadaran diri merupakan atribut yang sangat diperlukan untuk digunakan dalam mendefinisikan fungsi individu atau sistem senjata sebagai otonom." (Ini sama tidak masuk akalnya dengan menyatakan bahwa rudal bukanlah rudal kecuali jika kecepatannya melebihi kecepatan cahaya.) Bahkan, senjata otonom akan memiliki tingkat otonomi yang sama dengan program catur, yang diberi misi untuk memenangkan permainan tetapi memutuskan sendiri ke mana harus memindahkan bidaknya dan bidak musuh mana yang harus dieliminasi.

Gambar 4.1 (kiri) Senjata jelajah Harop yang diproduksi oleh Israel Aerospace Industries; (kanan) gambar diam dari video Slaughterbots yang menunjukkan kemungkinan desain untuk senjata otonom yang berisi proyektil kecil yang digerakkan oleh bahan peledak.

Senjata anti-personel (AWS) bukanlah fiksi ilmiah. Mereka sudah ada. Mungkin contoh yang paling jelas adalah Harop Israel (gambar 4.1, kiri), amunisi jelajah dengan bentang sayap sepuluh kaki dan hulu ledak lima puluh pon. Ia mencari target hingga enam jam di wilayah geografis tertentu yang memenuhi kriteria tertentu dan kemudian menghancurkannya. Kriterianya bisa berupa "memancarkan sinyal radar yang menyerupai radar anti-pesawat" atau "terlihat seperti tank."

Dengan menggabungkan kemajuan terbaru dalam desain quadrotor miniatur, kamera miniatur, chip penglihatan komputer, algoritma navigasi dan pemetaan, dan metode untuk

mendeteksi dan melacak manusia, akan memungkinkan dalam waktu yang cukup singkat untuk mengerahkan senjata anti-personel seperti Slaughterbot yang ditunjukkan pada gambar 4.1 (kanan). Senjata semacam itu dapat ditugaskan untuk menyerang siapa pun yang memenuhi kriteria visual tertentu (usia, jenis kelamin, seragam, warna kulit, dan sebagainya) atau bahkan individu tertentu berdasarkan pengenalan wajah. Saya diberitahu bahwa Departemen Pertahanan Swiss telah membangun dan menguji Slaughterbot yang sebenarnya dan menemukan bahwa, seperti yang diharapkan, teknologi tersebut layak dan mematikan.

Sejak 2014, diskusi diplomatik telah berlangsung di Jenewa yang dapat mengarah pada perjanjian yang melarang AWS. Pada saat yang sama, beberapa peserta utama dalam diskusi ini (Amerika Serikat, Tiongkok, Rusia, dan sampai batas tertentu Israel dan Inggris) terlibat dalam persaingan berbahaya untuk mengembangkan senjata otonom. Di Amerika Serikat, misalnya, program CODE (*Collaborative Operations in Denied Environments*) bertujuan untuk bergerak menuju otonomi dengan memungkinkan drone untuk berfungsi dengan kontak radio yang paling baik bersifat intermiten. Drone-drone tersebut akan "berburu dalam kelompok, seperti serigala" menurut manajer program. Pada tahun 2016, Angkatan Udara AS mendemonstrasikan penyebaran 103 mikro-drone Perdix di udara dari tiga pesawat tempur F/A-18. Menurut pengumuman tersebut, "Perdix bukanlah individu yang diprogram dan disinkronkan sebelumnya, mereka adalah organisme kolektif, berbagi satu otak terdistribusi untuk pengambilan keputusan dan beradaptasi satu sama lain seperti kawanan di alam."

Anda mungkin berpikir cukup jelas bahwa membangun mesin yang dapat memutuskan untuk membunuh manusia adalah ide yang buruk. Tetapi "cukup jelas" tidak selalu meyakinkan bagi pemerintah termasuk beberapa yang tercantum dalam paragraf sebelumnya yang bertekad untuk mencapai apa yang mereka anggap sebagai superioritas strategis. Alasan yang lebih meyakinkan untuk menolak senjata otonom adalah bahwa senjata tersebut merupakan senjata pemusnah massal yang dapat diskalakan.

Skalabilitas adalah istilah dari ilmu komputer; suatu proses dikatakan dapat diskalakan jika Anda dapat melakukan satu juta kali lebih banyak hal tersebut pada dasarnya dengan membeli satu juta kali lebih banyak perangkat keras. Dengan demikian, Google menangani sekitar lima miliar permintaan pencarian per hari bukan dengan jutaan karyawan, tetapi dengan jutaan komputer. Dengan senjata otonom, Anda dapat melakukan pembunuhan jutaan kali lebih banyak dengan membeli jutaan kali lebih banyak senjata, justru karena senjata tersebut otonom. Tidak seperti drone yang dikendalikan dari jarak jauh atau AK-47, senjata ini tidak memerlukan pengawasan manusia secara individual untuk melakukan pekerjaannya.

Sebagai senjata pemusnah massal, senjata otonom yang dapat diskalakan memiliki keunggulan bagi penyerang dibandingkan dengan senjata nuklir dan pengeboman karpet: senjata ini tidak merusak properti dan dapat diterapkan secara selektif untuk melenyapkan hanya mereka yang mungkin mengancam pasukan pendudukan. Senjata ini tentu dapat digunakan untuk memusnahkan seluruh kelompok etnis atau semua penganut agama tertentu (jika penganutnya memiliki tanda-tanda yang terlihat).

Selain itu, sementara penggunaan senjata nuklir mewakili ambang batas bencana yang telah kita (seringkali karena keberuntungan semata) hindari sejak tahun 1945, tidak ada

ambang batas seperti itu dengan senjata otonom yang dapat diskalakan. Serangan dapat meningkat dengan lancar dari seratus korban menjadi seribu, sepuluh ribu, hingga seratus ribu. Selain serangan nyata, ancaman serangan oleh senjata semacam itu saja sudah menjadikan senjata tersebut sebagai alat yang efektif untuk teror dan penindasan. Senjata otonom akan sangat mengurangi keamanan manusia di semua tingkatan: pribadi, lokal, nasional, dan internasional.

Ini bukan berarti bahwa senjata otonom akan menjadi akhir dunia seperti yang dibayangkan dalam film Terminator. Senjata tersebut tidak perlu terlalu cerdas mobil tanpa pengemudi mungkin perlu lebih cerdas dan misi mereka bukanlah jenis "menguasai dunia". Risiko eksistensial dari AI tidak terutama berasal dari robot pembunuh yang berpikiran sederhana. Di sisi lain, mesin supercerdas yang berkonflik dengan umat manusia tentu dapat mempersenjatai diri dengan cara ini, dengan mengubah robot pembunuh yang relatif bodoh menjadi perpanjangan fisik dari sistem kendali global.

4.3 MENGHILANGKAN PEKERJAAN SEPERTI YANG KITA KETAHUI

Ribuan artikel media dan opini serta beberapa buku telah ditulis tentang topik robot yang mengambil alih pekerjaan manusia. Pusat-pusat penelitian bermunculan di seluruh dunia untuk memahami apa yang kemungkinan akan terjadi. Judul-judul buku Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* dan buku Calum Chace, *The Economic Singularity: Artificial Intelligence and the Death of Capitalism*, cukup baik dalam meringkas kekhawatiran tersebut. Meskipun, seperti yang akan segera terlihat, saya sama sekali tidak memenuhi syarat untuk berpendapat tentang apa yang pada dasarnya merupakan masalah bagi para ekonom, saya menduga bahwa masalah ini terlalu penting untuk diserahkan sepenuhnya kepada mereka.

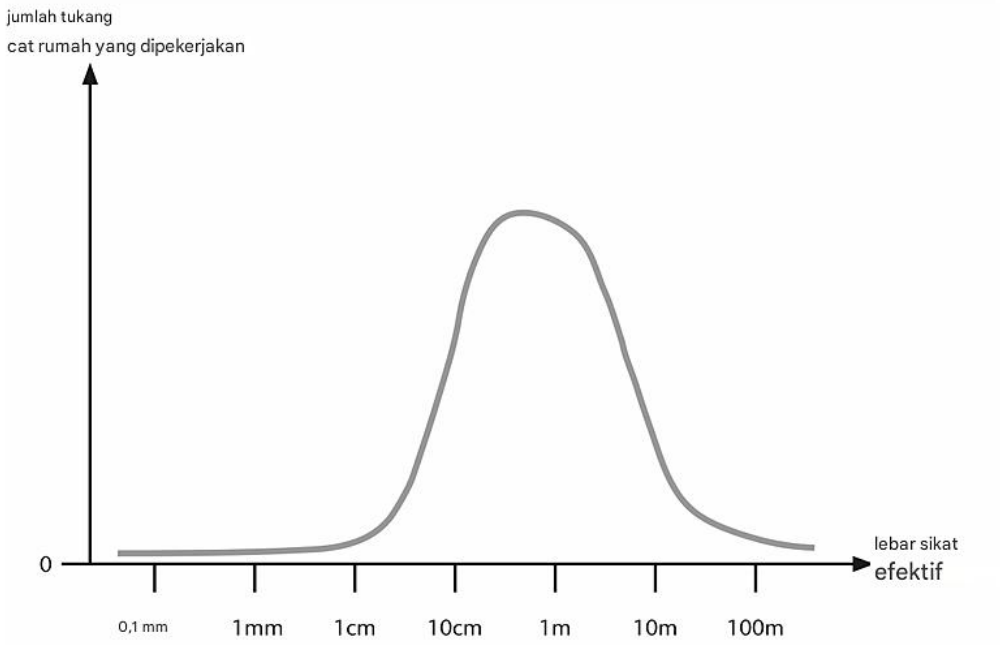
Masalah pengangguran teknologi diangkat ke permukaan dalam sebuah artikel terkenal, “Economic Possibilities for Our Grandchildren,” oleh John Maynard Keynes. Ia menulis artikel tersebut pada tahun 1930, ketika Depresi Besar telah menciptakan pengangguran massal di Inggris, tetapi topik ini memiliki sejarah yang jauh lebih panjang. Aristoteles, dalam Buku I karyanya Politik, menyajikan poin utama dengan cukup jelas:

> Sebab, jika setiap alat dapat menyelesaikan pekerjaannya sendiri, menuruti atau mengantisipasi kehendak orang lain... jika, dengan cara yang sama, alat tenun dapat menenun dan plektrum dapat memainkan kecapi tanpa tangan yang membimbingnya, maka para pekerja utama tidak akan membutuhkan pelayan, dan para majikan tidak akan membutuhkan budak.

Semua orang setuju dengan pengamatan Aristoteles bahwa terjadi pengurangan lapangan kerja secara langsung ketika seorang pengusaha menemukan metode mekanis untuk melakukan pekerjaan yang sebelumnya dilakukan oleh manusia. Masalahnya adalah apakah efek kompensasi yang terjadi dan yang cenderung meningkatkan lapangan kerja pada akhirnya akan menutupi pengurangan ini.

Kaum optimis mengatakan ya dan dalam debat saat ini, mereka menunjuk pada semua pekerjaan baru yang muncul setelah revolusi industri sebelumnya. Kaum pesimis mengatakan tidak dan dalam debat saat ini, mereka berpendapat bahwa mesin juga akan melakukan semua "pekerjaan baru" tersebut. Ketika sebuah mesin menggantikan kerja fisik seseorang, seseorang dapat menjual kerja mental. Ketika sebuah mesin menggantikan kerja mental seseorang, apa yang tersisa untuk dijual?

Dalam *Life 3.0*, Max Tegmark menggambarkan perdebatan tersebut sebagai percakapan antara dua kuda yang membahas munculnya mesin pembakaran internal pada tahun 1900. Salah satunya memprediksi “pekerjaan baru untuk kuda. Itulah yang selalu terjadi sebelumnya, seperti penemuan roda dan bajak.” Sayangnya, bagi sebagian besar kuda, “pekerjaan baru” itu adalah menjadi makanan hewan peliharaan.



Gambar 4.2 Grafik hipotetis pekerjaan pengecatan rumah seiring dengan peningkatan teknologi pengecatan.

Debat ini telah berlangsung selama ribuan tahun karena ada efek di kedua arah. Hasil sebenarnya bergantung pada efek mana yang lebih penting. Pertimbangkan, misalnya, apa yang terjadi pada tukang cat rumah seiring dengan peningkatan teknologi. Demi kesederhanaan, saya akan membiarkan lebar kuas cat mewakili tingkat otomatisasi:

- ✓ Jika kuas selebar satu helai rambut (sepersepuluh milimeter), dibutuhkan ribuan tahun kerja manusia untuk mengecat sebuah rumah dan pada dasarnya tidak ada tukang cat rumah yang dipekerjakan.
- ✓ Dengan kuas selebar satu milimeter, mungkin beberapa mural halus dilukis di istana kerajaan oleh segelintir pelukis. Pada satu sentimeter, kaum bangsawan mulai mengikuti jejaknya.

- ✓ Pada ukuran sepuluh sentimeter (empat inci), kita mencapai ranah kepraktisan: sebagian besar pemilik rumah mengecat rumah mereka di dalam dan di luar, meskipun mungkin tidak terlalu sering, dan ribuan tukang cat rumah mendapatkan pekerjaan.
- ✓ Begitu kita sampai pada rol dan pistol semprot yang lebar setara dengan kuas cat selebar sekitar satu meter harganya turun drastis, tetapi permintaan mungkin mulai jenuh sehingga jumlah tukang cat rumah agak berkurang.
- ✓ Ketika satu orang mengelola tim yang terdiri dari seratus robot pengecat rumah produktivitas yang setara dengan kuas cat selebar seratus meter maka seluruh rumah dapat dicat dalam satu jam dan sangat sedikit tukang cat rumah yang akan bekerja.

Dengan demikian, efek langsung teknologi bekerja dua arah: pertama, dengan meningkatkan produktivitas, teknologi dapat meningkatkan lapangan kerja dengan mengurangi harga suatu aktivitas dan dengan demikian meningkatkan permintaan; selanjutnya, peningkatan teknologi lebih lanjut berarti semakin sedikit manusia yang dibutuhkan. Gambar 4.2 mengilustrasikan perkembangan ini.

Banyak teknologi menunjukkan kurva yang serupa. Jika, di sektor ekonomi tertentu, kita berada di sebelah kiri puncak, maka peningkatan teknologi meningkatkan lapangan kerja di sektor tersebut; contoh saat ini mungkin termasuk tugas-tugas seperti penghapusan grafiti, pembersihan lingkungan, inspeksi kontainer pengiriman, dan konstruksi perumahan di negara-negara kurang berkembang, yang semuanya mungkin menjadi lebih layak secara ekonomi jika kita memiliki robot untuk membantu kita. Jika kita sudah berada di sebelah kanan puncak, maka otomatisasi lebih lanjut mengurangi lapangan kerja. Misalnya, tidak sulit untuk memprediksi bahwa operator lift akan terus tersingkir. Dalam jangka panjang, kita harus memperkirakan bahwa sebagian besar industri akan didorong ke paling kanan pada kurva. Sebuah artikel baru-baru ini, berdasarkan studi ekonometrik yang cermat oleh ekonom David Autor dan Anna Salomons, menyatakan bahwa “selama 40 tahun terakhir, lapangan kerja telah berkurang di setiap industri yang memperkenalkan teknologi untuk meningkatkan produktivitas.”

Bagaimana dengan efek kompensasi yang dijelaskan oleh para optimis ekonomi?

- ➢ Beberapa orang harus membuat robot pengecat. Berapa banyak? Jauh lebih sedikit daripada jumlah tukang cat rumah yang digantikan oleh robot jika tidak, biaya mengecat rumah dengan robot akan lebih mahal, bukan lebih murah, dan tidak ada yang akan membeli robot tersebut.
- ➢ Pengecatan rumah menjadi sedikit lebih murah, sehingga orang lebih sering memanggil tukang cat rumah.
- ➢ Akhirnya, karena kita membayar lebih sedikit untuk pengecatan rumah, kita memiliki lebih banyak uang untuk dibelanjakan pada hal-hal lain, sehingga meningkatkan lapangan kerja di sektor lain.

Para ekonom telah mencoba mengukur besarnya efek ini di berbagai industri yang mengalami peningkatan otomatisasi, tetapi hasilnya umumnya tidak meyakinkan.

Secara historis, sebagian besar ekonom arus utama berpendapat dari sudut pandang "gambaran besar": otomatisasi meningkatkan produktivitas, sehingga secara keseluruhan,

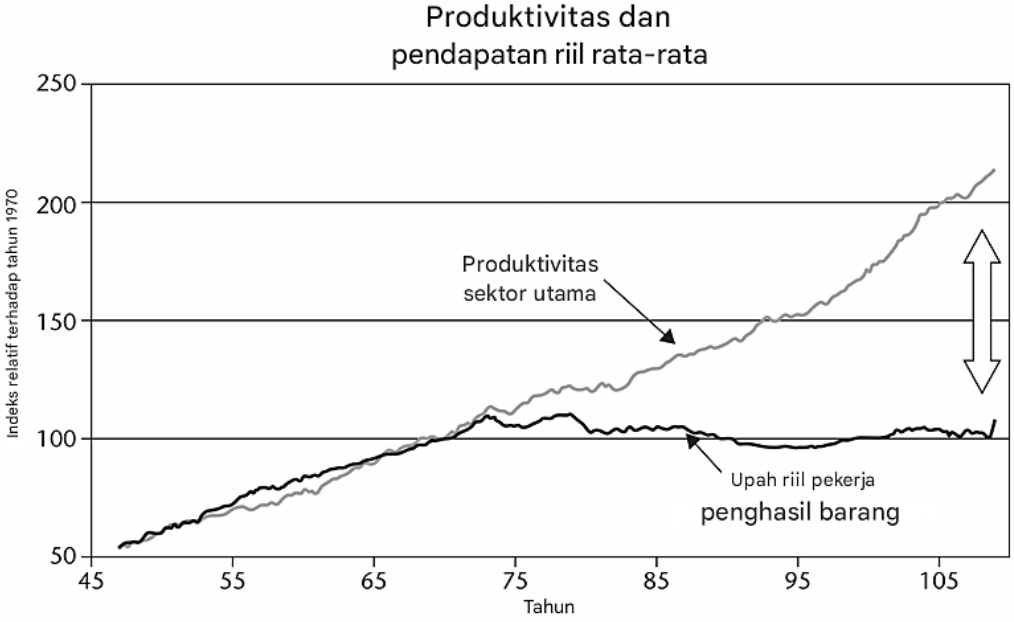
manusia menjadi lebih baik, dalam arti bahwa kita menikmati lebih banyak barang dan jasa untuk jumlah pekerjaan yang sama.

Sayangnya, teori ekonomi tidak memprediksi bahwa setiap manusia akan menjadi lebih baik sebagai akibat dari otomatisasi. Secara umum, otomatisasi meningkatkan bagian pendapatan yang masuk ke modal (pemilik robot pengecat rumah) dan mengurangi bagian yang masuk ke tenaga kerja (mantan tukang cat rumah). Ekonom Erik Brynjolfsson dan Andrew McAfee, dalam *The Second Machine Age*, berpendapat bahwa hal ini telah terjadi selama beberapa dekade. Data untuk Amerika Serikat ditunjukkan pada gambar 4.3. Data tersebut menunjukkan bahwa antara tahun 1947 dan 1973, upah dan produktivitas meningkat bersamaan, tetapi setelah tahun 1973, upah stagnan meskipun produktivitas meningkat hampir dua kali lipat.

Brynjolfsson dan McAfee menyebut ini sebagai Pemisahan Besar (*Great Decoupling*). Ekonom terkemuka lainnya juga telah menyuarakan kekhawatiran, termasuk peraih Nobel Robert Shiller, Mike Spence, dan Paul Krugman; Klaus Schwab, kepala Forum Ekonomi Dunia; dan Larry Summers, mantan kepala ekonom Bank Dunia dan menteri keuangan di bawah Presiden Bill Clinton.

Mereka yang menentang gagasan pengangguran akibat teknologi seringkali menunjuk pada teller bank, yang pekerjaannya sebagian dapat dilakukan oleh ATM, dan kasir ritel, yang pekerjaannya dipercepat oleh barcode dan tag RFID pada barang dagangan. Seringkali diklaim bahwa pekerjaan-pekerjaan ini berkembang karena teknologi. Memang, jumlah teller di Amerika Serikat meningkat hampir dua kali lipat dari tahun 1970 hingga 2010, meskipun perlu dicatat bahwa populasi AS tumbuh sebesar 50 persen dan sektor keuangan lebih dari 400 persen pada periode yang sama, sehingga sulit untuk mengaitkan semua, atau mungkin sebagian, pertumbuhan lapangan kerja dengan ATM.

Sayangnya, antara tahun 2010 dan 2016 sekitar seratus ribu teller kehilangan pekerjaan mereka, dan Biro Statistik Tenaga Kerja AS (BLS) memperkirakan akan ada lagi kehilangan pekerjaan sebanyak empat puluh ribu orang pada tahun 2026: “Perbankan online dan teknologi otomatisasi diperkirakan akan terus menggantikan lebih banyak tugas pekerjaan yang secara tradisional dilakukan oleh teller.” Data tentang kasir ritel tidak lebih menggembirakan: jumlah per kapita turun sebesar 5 persen dari tahun 1997 hingga 2015, dan BLS mengatakan, “Kemajuan teknologi, seperti mesin kasir swalayan di toko ritel dan peningkatan penjualan online, akan terus membatasi kebutuhan akan kasir.” Kedua sektor tersebut tampaknya sedang mengalami penurunan. Hal yang sama berlaku untuk hampir semua pekerjaan berketerampilan rendah yang melibatkan pekerjaan dengan mesin.



Gambar 4.3 Produksi ekonomi dan upah median riil di Amerika Serikat sejak tahun 1947. (Data dari Biro Statistik Tenaga Kerja.)

Pekerjaan apa saja yang akan mengalami penurunan seiring dengan hadirnya teknologi berbasis AI baru? Contoh utama yang dikutip di media adalah pekerjaan pengemudi. Di Amerika Serikat terdapat sekitar 3,5 juta pengemudi truk; banyak dari pekerjaan ini akan rentan terhadap otomatisasi. Amazon, di antara perusahaan lain, sudah menggunakan truk swakemudi untuk pengangkutan barang di jalan raya antar negara bagian, meskipun saat ini dengan pengemudi cadangan manusia. Tampaknya sangat mungkin bahwa bagian perjalanan jarak jauh dari setiap perjalanan truk akan segera menjadi otonom, sementara manusia, untuk sementara waktu, akan menangani lalu lintas kota, pengambilan, dan pengiriman. Sebagai konsekuensi dari perkembangan yang diharapkan ini, sangat sedikit anak muda yang tertarik pada pekerjaan pengemudi truk sebagai karier; ironisnya, saat ini terdapat kekurangan pengemudi truk yang signifikan di Amerika Serikat, yang hanya mempercepat munculnya otomatisasi.

Pekerjaan kerah putih juga berisiko. Sebagai contoh, BLS memproyeksikan penurunan 13 persen dalam pekerjaan penjamin asuransi per kapita dari tahun 2016 hingga 2026: "Perangkat lunak penjaminan otomatis memungkinkan pekerja untuk memproses aplikasi lebih cepat daripada sebelumnya, mengurangi kebutuhan akan banyak penjamin." Jika teknologi bahasa berkembang seperti yang diharapkan, banyak pekerjaan penjualan dan layanan pelanggan juga akan rentan, serta pekerjaan di bidang hukum. (Dalam kompetisi tahun 2018, perangkat lunak AI mengungguli profesor hukum berpengalaman dalam menganalisis perjanjian kerahasiaan standar dan menyelesaikan tugas tersebut dua ratus kali lebih cepat.) Bentuk pemrograman komputer rutin jenis yang sering dialihdayakan saat ini juga kemungkinan akan diotomatisasi. Memang, hampir semua hal yang dapat dialihdayakan merupakan kandidat yang baik untuk otomatisasi, karena pengalihdayaan melibatkan penguraian pekerjaan menjadi tugas-tugas yang dapat dipisah-pisahkan dan didistribusikan

dalam bentuk yang terlepas dari konteksnya. Industri otomatisasi proses robot menghasilkan perangkat lunak yang mencapai efek persis ini untuk tugas-tugas administrasi yang dilakukan secara daring.

Seiring kemajuan AI, sangat mungkin bahkan mungkin bahwa dalam beberapa dekade mendatang, hampir semua pekerjaan fisik dan mental rutin akan dilakukan lebih murah oleh mesin. Sejak kita berhenti menjadi pemburu-pengumpul ribuan tahun yang lalu, masyarakat kita telah menggunakan sebagian besar manusia sebagai robot, melakukan tugas-tugas manual dan mental yang berulang, jadi mungkin tidak mengherankan jika robot akan segera mengambil alih peran ini.

Ketika ini terjadi, hal itu akan mendorong upah di bawah garis kemiskinan bagi orang-orang yang tidak mampu bersaing untuk pekerjaan berketerampilan tinggi yang tersisa. Larry Summers mengatakannya seperti ini: “Mungkin saja, mengingat kemungkinan substitusi [modal untuk tenaga kerja], beberapa kategori tenaga kerja tidak akan mampu memperoleh pendapatan subsisten.” Inilah tepatnya yang terjadi pada kuda: transportasi mekanis menjadi lebih murah daripada biaya pemeliharaan kuda, sehingga kuda menjadi makanan hewan peliharaan. Dihadapkan dengan kondisi sosial ekonomi yang setara dengan menjadi makanan hewan peliharaan, manusia akan sangat tidak senang dengan pemerintah mereka.

Dihadapkan dengan potensi ketidakbahagiaan manusia, pemerintah di seluruh dunia mulai memberikan perhatian pada masalah ini. Sebagian besar telah menemukan bahwa gagasan melatih ulang semua orang sebagai ilmuwan data atau insinyur robot adalah hal yang mustahil dunia mungkin membutuhkan lima atau sepuluh juta orang, tetapi jauh dari satu miliar pekerjaan yang berisiko. Ilmu data adalah sekoci penyelamat yang sangat kecil untuk kapal pesiar raksasa. Beberapa sedang mengerjakan "rencana transisi" tetapi transisi ke mana? Kita membutuhkan tujuan yang masuk akal untuk merencanakan transisi yaitu, kita membutuhkan gambaran yang masuk akal tentang ekonomi masa depan yang diinginkan di mana sebagian besar dari apa yang saat ini kita sebut pekerjaan dilakukan oleh mesin.

Salah satu gambaran yang muncul dengan cepat adalah ekonomi di mana jauh lebih sedikit orang yang bekerja karena pekerjaan tidak diperlukan. Keynes membayangkan masa depan seperti itu dalam esainya "Kemungkinan Ekonomi untuk Cucu Kita." Ia menggambarkan tingginya angka pengangguran yang melanda Inggris Raya pada tahun 1930 sebagai "fase ketidaksesuaian sementara" yang disebabkan oleh "peningkatan efisiensi teknis" yang terjadi "lebih cepat daripada kemampuan kita untuk mengatasi masalah penyerapan tenaga kerja." Namun, ia tidak membayangkan bahwa dalam jangka panjang setelah satu abad kemajuan teknologi lebih lanjut akan ada kembalinya lapangan kerja penuh:

> Dengan demikian, untuk pertama kalinya sejak penciptaannya, manusia akan dihadapkan pada masalah nyata dan permanennya bagaimana menggunakan kebebasannya dari kekhawatiran ekonomi yang mendesak, bagaimana memanfaatkan waktu luang yang telah dimenangkan oleh sains dan bunga majemuk, untuk hidup dengan bijak, menyenangkan, dan sejahtera.

Masa depan seperti itu membutuhkan perubahan radikal dalam sistem ekonomi kita, karena di banyak negara, mereka yang tidak bekerja menghadapi kemiskinan atau kekurangan. Dengan demikian, para pendukung visi Keynes modern biasanya mendukung beberapa bentuk pendapatan dasar universal, atau UBI. Didanai oleh pajak pertambahan nilai atau pajak atas pendapatan dari modal, UBI akan memberikan pendapatan yang layak kepada setiap orang dewasa, terlepas dari keadaan mereka.

Mereka yang bercita-cita untuk mencapai standar hidup yang lebih tinggi masih dapat bekerja tanpa kehilangan UBI, sementara mereka yang tidak dapat menghabiskan waktu mereka sesuai keinginan. Mungkin mengejutkan, UBI mendapat dukungan di seluruh spektrum politik, mulai dari Adam Smith Institute hingga Partai Hijau.

Bagi sebagian orang, UBI mewakili versi surga. Bagi yang lain, UBI merupakan pengakuan kegagalan sebuah pernyataan bahwa sebagian besar orang tidak akan memiliki sesuatu yang bernilai ekonomi untuk disumbangkan kepada masyarakat. Mereka dapat diberi makan dan tempat tinggal sebagian besar oleh mesin tetapi selain itu dibiarkan mengurus diri mereka sendiri. Kebenaran, seperti biasa, terletak di suatu tempat di antara keduanya, dan itu sangat bergantung pada bagaimana seseorang memandang psikologi manusia. Keynes, dalam esainya, membuat perbedaan yang jelas antara mereka yang berjuang dan mereka yang menikmati orang-orang yang "bertujuan" yang bagi mereka "selai bukanlah selai kecuali jika itu adalah selai besok dan tidak pernah selai hari ini" dan orang-orang yang "menyenangkan" yang "mampu menikmati hal-hal secara langsung." Proposal UBI mengasumsikan bahwa sebagian besar orang termasuk dalam kategori yang menyenangkan.

Keynes berpendapat bahwa berjuang adalah salah satu "kebiasaan dan naluri manusia biasa, yang ditanamkan dalam dirinya selama bergenerasi-generasi" daripada salah satu "nilai-nilai kehidupan yang sebenarnya." Ia memprediksi bahwa naluri ini akan perlahan menghilang. Berlawanan dengan pandangan ini, seseorang dapat berpendapat bahwa berjuang adalah hal yang intrinsik bagi makna menjadi manusia sejati. Alih-alih berjuang dan menikmati adalah hal yang saling eksklusif, keduanya seringkali tak terpisahkan: kenikmatan sejati dan kepuasan abadi datang dari memiliki tujuan dan mencapainya (atau setidaknya mencoba), biasanya di tengah rintangan, daripada dari konsumsi pasif kesenangan sesaat. Ada perbedaan antara mendaki Everest dan diantar ke puncak dengan helikopter.

Hubungan antara berjuang dan menikmati adalah tema sentral bagi pemahaman kita tentang bagaimana membentuk masa depan yang diinginkan. Mungkin generasi mendatang akan bertanya-tanya mengapa kita pernah mengkhawatirkan hal yang sia-sia seperti "pekerjaan." Jika perubahan sikap itu datang perlahan, mari kita pertimbangkan implikasi ekonomi dari pandangan bahwa sebagian besar orang akan lebih baik jika memiliki sesuatu yang bermanfaat untuk dilakukan, meskipun sebagian besar barang dan jasa akan diproduksi oleh mesin dengan sedikit pengawasan manusia.

Tak pelak lagi, sebagian besar orang akan terlibat dalam menyediakan layanan interpersonal yang dapat disediakan atau yang lebih kita sukai untuk disediakan hanya oleh manusia. Artinya, jika kita tidak lagi dapat menyediakan pekerjaan fisik rutin dan pekerjaan

mental rutin, kita masih dapat menyediakan kemanusiaan kita. Kita perlu menjadi mahir dalam menjadi manusia.

Profesi saat ini yang termasuk dalam kategori ini adalah psikoterapis, pelatih eksekutif, tutor, konselor, pendamping, dan mereka yang merawat anak-anak dan orang tua. Frasa "profesi pengasuh" sering digunakan dalam konteks ini, tetapi itu menyesatkan: frasa ini memiliki konotasi positif bagi mereka yang memberikan perawatan, tentu saja, tetapi konotasi negatif tentang ketergantungan dan ketidakberdayaan bagi penerima perawatan. Namun, pertimbangkan pengamatan ini, sekali lagi dari Keynes:

> "Orang-oranglah yang dapat menjaga dan mengembangkan seni hidup itu sendiri hingga mencapai kesempurnaan yang lebih penuh, dan tidak menjual diri mereka demi sarana hidup, yang akan dapat menikmati kelimpahan ketika itu datang."

Kita semua membutuhkan bantuan dalam mempelajari "seni hidup itu sendiri." Ini bukan masalah ketergantungan tetapi pertumbuhan. Kapasitas untuk menginspirasi orang lain dan untuk memberikan kemampuan untuk menghargai dan menciptakan baik itu dalam seni, musik, sastra, percakapan, berkebun, arsitektur, makanan, anggur, atau permainan video kemungkinan akan lebih dibutuhkan daripada sebelumnya.

Pertanyaan selanjutnya adalah distribusi pendapatan. Di sebagian besar negara, hal ini telah bergerak ke arah yang salah selama beberapa dekade. Ini adalah masalah yang kompleks, tetapi satu hal yang jelas: pendapatan tinggi dan kedudukan sosial yang tinggi biasanya mengikuti dari pemberian nilai tambah yang tinggi. Profesi pengasuhan anak, misalnya, dikaitkan dengan pendapatan rendah dan status sosial rendah. Ini sebagian merupakan konsekuensi dari kenyataan bahwa kita sebenarnya tidak tahu bagaimana melakukannya. Beberapa praktisi secara alami mahir dalam hal ini, tetapi banyak yang tidak. Bandingkan ini dengan, misalnya, bedah ortopedi. Kita tidak akan begitu saja mempekerjakan remaja yang bosan dan membutuhkan sedikit uang tambahan lalu menempatkan mereka sebagai ahli bedah ortopedi dengan upah lima dolar per jam ditambah makanan sepuasnya dari lemari es.

Kita telah melakukan penelitian selama berabad-abad untuk memahami tubuh manusia dan cara memperbaikinya ketika rusak, dan para praktisi harus menjalani pelatihan bertahun-tahun untuk mempelajari semua pengetahuan ini dan keterampilan yang diperlukan untuk menerapkannya. Akibatnya, ahli bedah ortopedi dibayar tinggi dan sangat dihormati. Mereka dibayar tinggi bukan hanya karena mereka banyak tahu dan memiliki banyak pelatihan, tetapi juga karena semua pengetahuan dan pelatihan itu benar-benar berhasil. Hal itu memungkinkan mereka untuk memberikan nilai tambah yang besar bagi kehidupan orang lain terutama orang-orang dengan bagian tubuh yang rusak.

Sayangnya, pemahaman ilmiah kita tentang pikiran sangat lemah dan pemahaman ilmiah kita tentang kebahagiaan dan kepuasan bahkan lebih lemah. Kita sama sekali tidak tahu bagaimana cara menambah nilai pada kehidupan satu sama lain dengan cara yang konsisten dan dapat diprediksi. Kita telah mencapai keberhasilan yang cukup baik dengan gangguan kejiwaan tertentu, tetapi kita masih berjuang dalam Perang Literasi Seratus Tahun untuk

sesuatu yang mendasar seperti mengajari anak-anak membaca. Kita membutuhkan pemikiran ulang yang radikal tentang sistem pendidikan dan usaha ilmiah kita untuk lebih memfokuskan perhatian pada dunia manusia daripada dunia fisik. (Joseph Aoun, presiden Universitas Northeastern, berpendapat bahwa universitas harus mengajarkan dan mempelajari "humanika.")

Kedengarannya aneh untuk mengatakan bahwa kebahagiaan harus menjadi disiplin ilmu teknik, tetapi tampaknya itu adalah kesimpulan yang tak terhindarkan. Disiplin ilmu semacam itu akan dibangun di atas ilmu dasar pemahaman yang lebih baik tentang bagaimana pikiran manusia bekerja pada tingkat kognitif dan emosional dan akan melatih berbagai macam praktisi, mulai dari arsitek kehidupan, yang membantu individu merencanakan bentuk keseluruhan lintasan hidup mereka, hingga para ahli profesional dalam topik-topik seperti peningkatan rasa ingin tahu dan ketahanan pribadi. Jika didasarkan pada ilmu pengetahuan yang sebenarnya, profesi-profesi ini tidak perlu lebih mengada-ada daripada perancang jembatan dan ahli bedah ortopedi saat ini.

Membangun kembali lembaga pendidikan dan penelitian kita untuk menciptakan ilmu dasar ini dan mengubahnya menjadi program pelatihan dan profesi yang bersertifikasi akan membutuhkan waktu puluhan tahun, jadi ada baiknya untuk memulainya sekarang dan sayang sekali kita tidak memulainya sejak lama. Hasil akhirnya jika berhasil akan menjadi dunia yang layak untuk ditinggali. Tanpa pemikiran ulang seperti itu, kita berisiko mengalami dislokasi sosial ekonomi yang tidak berkelanjutan.

4.4 MENGAMBIL ALIH PERAN MANUSIA LAIN

Kita harus berpikir dua kali sebelum membiarkan mesin mengambil alih peran yang melibatkan layanan interpersonal. Jika menjadi manusia adalah nilai jual utama kita kepada manusia lain, maka membuat manusia tiruan tampaknya merupakan ide yang buruk. Untungnya bagi kita, kita memiliki keunggulan yang berbeda dibandingkan mesin dalam hal mengetahui bagaimana perasaan manusia lain dan bagaimana mereka akan bereaksi. Hampir setiap manusia tahu bagaimana rasanya memukul ibu jari dengan palu atau merasakan cinta yang tak berbalas. Yang mengimbangi keunggulan alami manusia ini adalah kelemahan alami manusia: kecenderungan untuk tertipu oleh penampilan terutama penampilan manusia. Alan Turing memperingatkan agar tidak membuat robot menyerupai manusia:

> Saya sangat berharap dan percaya bahwa tidak akan ada upaya besar untuk membuat mesin dengan karakteristik yang paling khas manusia, tetapi non-intelektual, seperti bentuk tubuh manusia; menurut saya upaya seperti itu sangat sia-sia dan hasilnya akan memiliki kualitas yang tidak menyenangkan seperti bunga buatan.

Sayangnya, peringatan Turing telah diabaikan. Beberapa kelompok penelitian telah menghasilkan robot yang sangat mirip manusia, seperti yang ditunjukkan pada gambar 4.4.

Gambar 4.4 (kiri) JiaJia, robot yang dibangun di Universitas Sains dan Teknologi Tiongkok; (kanan) Geminoid DK, robot yang dirancang oleh Hiroshi Ishiguro di Universitas Osaka di Jepang dan dimodelkan berdasarkan Henrik Schärfe dari Universitas Aalborg di Denmark.

Sebagai alat penelitian, robot-robot tersebut dapat memberikan wawasan tentang bagaimana manusia menafsirkan perilaku dan komunikasi robot. Sebagai prototipe untuk produk komersial di masa depan, mereka mewakili bentuk ketidakjujuran. Mereka melewati kesadaran kita dan langsung menarik sisi emosional kita, mungkin meyakinkan kita bahwa mereka diberkahi dengan kecerdasan nyata. Bayangkan, misalnya, betapa jauh lebih mudahnya mematikan dan mendaur ulang sebuah kotak abu-abu pendek yang rusak meskipun kotak itu mengeluarkan suara aneh karena tidak mau dimatikan daripada melakukan hal yang sama untuk JiaJia atau Geminoid DK. Bayangkan juga betapa membingungkan dan mungkin mengganggu secara psikologis bagi bayi dan anak kecil untuk diasuh oleh entitas yang tampak seperti manusia, seperti orang tua mereka, tetapi entah bagaimana bukan; yang tampak peduli pada mereka, seperti orang tua mereka, tetapi sebenarnya tidak.

Selain kemampuan dasar untuk menyampaikan informasi nonverbal melalui ekspresi wajah dan gerakan yang bahkan Bugs Bunny mampu lakukan dengan mudah tidak ada alasan yang baik bagi robot untuk memiliki bentuk humanoid. Ada juga alasan praktis yang baik untuk tidak memiliki bentuk humanoid misalnya, postur bipedal kita relatif tidak stabil dibandingkan dengan lokomosi quadrupedal. Anjing, kucing, dan kuda cocok dengan kehidupan kita, dan bentuk fisik mereka merupakan petunjuk yang sangat baik tentang bagaimana mereka cenderung berperilaku. (Bayangkan jika seekor kuda tiba-tiba mulai berperilaku seperti anjing!) Hal yang sama seharusnya berlaku untuk robot. Mungkin morfologi seperti centaur dengan empat kaki dan dua lengan akan menjadi standar yang baik. Robot humanoid yang akurat sama tidak masuk akalnya dengan Ferrari dengan kecepatan tertinggi lima mil per jam atau es krim cone "raspberry" yang terbuat dari krim hati cincang yang diberi warna bit.

Aspek humanoid dari beberapa robot telah berkontribusi pada kebingungan politik dan emosional. Pada 25 Oktober 2017, Arab Saudi memberikan kewarganegaraan kepada Sophia, robot humanoid yang digambarkan sebagai tidak lebih dari "chatbot dengan wajah" dan lebih buruk lagi. Mungkin ini adalah aksi humas, tetapi proposal yang berasal dari Komite Urusan Hukum Parlemen Eropa sepenuhnya serius. Proposal tersebut merekomendasikan pembuatan status hukum khusus untuk robot dalam jangka panjang, sehingga setidaknya robot otonom

yang paling canggih dapat ditetapkan sebagai entitas elektronik yang bertanggung jawab untuk memperbaiki kerusakan apa pun yang mungkin mereka sebabkan.

Dengan kata lain, robot itu sendiri akan bertanggung jawab secara hukum atas kerusakan, bukan pemilik atau produsennya. Ini menyiratkan bahwa robot akan memiliki aset keuangan dan akan dikenakan sanksi jika mereka tidak patuh. Jika diartikan secara harfiah, ini tidak masuk akal. Misalnya, jika kita memenjarakan robot karena tidak membayar, mengapa ia peduli?

Selain peningkatan status robot yang tidak perlu dan bahkan absurd, ada bahaya bahwa peningkatan penggunaan mesin dalam keputusan yang memengaruhi manusia akan menurunkan status dan martabat manusia. Kemungkinan ini diilustrasikan dengan sempurna dalam sebuah adegan dari film fiksi ilmiah *Elysium*, ketika Max (Matt Damon) memohon di hadapan "petugas pembebasan bersyaratnya" (gambar 4.5) untuk menjelaskan mengapa perpanjangan hukumannya tidak dapat dibenarkan. Tak perlu dikatakan, Max gagal. Petugas pembebasan bersyarat bahkan menegurnya karena gagal menunjukkan sikap hormat yang pantas.

Gambar 4.5 Max (Matt Damon) bertemu dengan petugas pembebasan bersyaratnya di Elysium.

Kita dapat memikirkan serangan terhadap martabat manusia semacam itu dalam dua cara. Pertama, jelas: dengan memberikan wewenang kepada mesin atas manusia, kita merendahkan diri kita sendiri ke status kelas dua dan kehilangan hak untuk berpartisipasi dalam keputusan yang memengaruhi kita. (Bentuk yang lebih ekstrem dari ini adalah memberikan wewenang kepada mesin untuk membunuh manusia, seperti yang dibahas sebelumnya dalam bab ini.)

Kedua, tidak langsung: bahkan jika Anda percaya bahwa bukan mesin yang membuat keputusan tetapi manusia yang merancang dan memesan mesin tersebut, fakta bahwa para perancang dan pemesan manusia tersebut tidak menganggapnya layak untuk mempertimbangkan keadaan individu setiap subjek manusia dalam kasus-kasus seperti itu menunjukkan bahwa mereka kurang menghargai kehidupan orang lain. Ini mungkin merupakan gejala awal dari pemisahan besar antara elit yang dilayani oleh manusia dan kelas bawah yang luas yang dilayani, dan dikendalikan, oleh mesin. Di Uni Eropa, Pasal 22 Peraturan

Perlindungan Data Umum 2018, atau GDPR, secara eksplisit melarang pemberian wewenang kepada mesin dalam kasus-kasus seperti ini:

> Subjek data berhak untuk tidak tunduk pada keputusan yang didasarkan semata-mata pada pemrosesan otomatis, termasuk pembuatan profil, yang menghasilkan efek hukum yang menyangkut dirinya atau secara signifikan memengaruhinya.

Meskipun ini terdengar mengagumkan secara prinsip, masih perlu dilihat setidaknya pada saat penulisan ini seberapa besar dampaknya dalam praktik. Seringkali jauh lebih mudah, cepat, dan murah untuk menyerahkan keputusan kepada mesin.

Salah satu alasan kekhawatiran tentang keputusan otomatis adalah potensi bias algoritmik kecenderungan algoritma pembelajaran mesin untuk menghasilkan keputusan yang bias secara tidak tepat tentang pinjaman, perumahan, pekerjaan, asuransi, pembebasan bersyarat, hukuman, penerimaan perguruan tinggi, dan sebagainya. Penggunaan kriteria seperti ras secara eksplisit dalam keputusan ini telah ilegal selama beberapa dekade di banyak negara dan dilarang oleh Pasal 9 GDPR untuk berbagai aplikasi yang sangat luas. Tentu saja, itu tidak berarti bahwa dengan mengecualikan ras dari data, kita akan mendapatkan keputusan yang tidak bias secara rasial. Misalnya, mulai tahun 1930-an, praktik redlining yang disetujui pemerintah menyebabkan kode pos tertentu di Amerika Serikat tidak dapat diakses untuk pinjaman hipotek dan bentuk investasi lainnya, yang menyebabkan penurunan nilai real estat. Kebetulan, kode pos tersebut sebagian besar dihuni oleh warga Afrika-Amerika.

Untuk mencegah redlining, sekarang hanya tiga digit pertama dari kode pos lima digit yang dapat digunakan dalam pengambilan keputusan kredit. Selain itu, proses pengambilan keputusan harus dapat diperiksa, untuk memastikan tidak ada bias "tidak disengaja" lain yang masuk. GDPR Uni Eropa sering dikatakan memberikan "hak untuk penjelasan" umum untuk setiap keputusan otomatis, tetapi bahasa sebenarnya dari Pasal 14 hanya mensyaratkan informasi yang bermakna tentang logika yang terlibat, serta signifikansi dan konsekuensi yang diperkirakan dari pemrosesan tersebut bagi subjek data.

Saat ini, belum diketahui bagaimana pengadilan akan menegakkan klausul ini. Ada kemungkinan bahwa konsumen yang malang hanya akan diberi deskripsi tentang algoritma pembelajaran mendalam tertentu yang digunakan untuk melatih pengklasifikasi yang membuat keputusan tersebut. Saat ini, kemungkinan penyebab bias algoritmik terletak pada data daripada pada kesalahan yang disengaja oleh perusahaan. Pada tahun 2015, majalah Glamour melaporkan temuan yang mengecewakan: “Hasil pencarian gambar Google pertama untuk ‘CEO’ muncul DUA BELAS baris ke bawah dan itu adalah Barbie.” (Ada beberapa wanita sungguhan dalam hasil tahun 2018, tetapi sebagian besar adalah model yang memerankan CEO dalam foto stok generik, bukan CEO wanita sungguhan; hasil tahun 2019 agak lebih baik.) Ini bukan konsekuensi dari bias gender yang disengaja dalam peringkat pencarian gambar Google, tetapi dari bias yang sudah ada sebelumnya dalam budaya yang menghasilkan data: ada jauh lebih banyak CEO pria daripada wanita, dan ketika orang ingin menggambarkan CEO

"arketipe" dalam gambar dengan keterangan, mereka hampir selalu memilih figur pria. Fakta bahwa bias tersebut terutama terletak pada data, tentu saja, tidak berarti bahwa tidak ada kewajiban untuk mengambil langkah-langkah untuk mengatasi masalah tersebut.

Ada alasan lain yang lebih teknis mengapa penerapan metode pembelajaran mesin secara naif dapat menghasilkan hasil yang bias. Misalnya, kelompok minoritas, menurut definisi, kurang terwakili dalam sampel data populasi secara luas; oleh karena itu, prediksi untuk anggota individu dari kelompok minoritas mungkin kurang akurat jika prediksi tersebut sebagian besar didasarkan pada data dari anggota lain dari kelompok yang sama. Untungnya, banyak perhatian telah diberikan pada masalah penghapusan bias yang tidak disengaja dari algoritma pembelajaran mesin, dan sekarang ada metode yang menghasilkan hasil yang tidak bias menurut beberapa definisi keadilan yang masuk akal dan diinginkan. Analisis matematis dari definisi keadilan ini menunjukkan bahwa definisi tersebut tidak dapat dicapai secara bersamaan dan bahwa, ketika diterapkan, definisi tersebut menghasilkan akurasi prediksi yang lebih rendah dan, dalam kasus keputusan pinjaman, keuntungan yang lebih rendah bagi pemberi pinjaman. Ini mungkin mengecewakan, tetapi setidaknya hal ini memperjelas pertimbangan yang terlibat dalam menghindari bias algoritmik. Kita berharap bahwa kesadaran akan metode ini dan masalah itu sendiri akan menyebar dengan cepat di antara para pembuat kebijakan, praktisi, dan pengguna.

Jika menyerahkan wewenang atas individu manusia kepada mesin terkadang bermasalah, bagaimana dengan wewenang atas banyak manusia? Artinya, haruskah kita menempatkan mesin dalam peran politik dan manajemen? Saat ini hal ini mungkin tampak mengada-ada. Mesin tidak dapat mempertahankan percakapan yang panjang dan kurang memiliki pemahaman dasar tentang faktor-faktor yang relevan untuk membuat keputusan dengan cakupan luas, seperti apakah akan menaikkan upah minimum atau menolak proposal merger dari perusahaan lain.

Namun, trennya jelas: mesin membuat keputusan pada tingkat wewenang yang semakin tinggi di banyak bidang. Ambil contoh maskapai penerbangan. Pertama, komputer membantu dalam penyusunan jadwal penerbangan. Segera, mereka mengambil alih alokasi awak penerbangan, pemesanan kursi, dan manajemen perawatan rutin. Selanjutnya, mereka terhubung ke jaringan informasi global untuk memberikan laporan status waktu nyata kepada manajer maskapai penerbangan, sehingga manajer dapat mengatasi gangguan secara efektif. Sekarang mereka mengambil alih pekerjaan mengelola gangguan: mengubah rute pesawat, menjadwal ulang staf, memesan ulang penumpang, dan merevisi jadwal perawatan.

Ini semua baik dari sudut pandang ekonomi maskapai penerbangan dan pengalaman penumpang. Pertanyaannya adalah apakah sistem komputer tetap menjadi alat manusia, atau manusia menjadi alat sistem komputer menyediakan informasi dan memperbaiki bug bila perlu, tetapi tidak lagi memahami secara mendalam bagaimana keseluruhan sistem bekerja. Jawabannya menjadi jelas ketika sistem mati dan kekacauan global terjadi sampai sistem dapat diaktifkan kembali.

Sebagai contoh, satu "gangguan komputer" pada tanggal 3 April 2018 menyebabkan lima belas ribu penerbangan di Eropa mengalami penundaan atau pembatalan yang signifikan.

Ketika algoritma perdagangan menyebabkan "flash crash" tahun 2010 di Bursa Saham New York, yang menghapus Rp 10 kuadriliun dalam beberapa menit, satu-satunya solusi adalah menutup bursa. Apa yang terjadi masih belum dipahami dengan baik. Sebelum ada teknologi, manusia hidup, seperti kebanyakan hewan, serba kekurangan. Kita berdiri langsung di tanah, bisa dikatakan demikian. Teknologi secara bertahap mengangkat kita ke atas piramida mesin, meningkatkan jejak kita sebagai individu dan sebagai spesies. Ada berbagai cara kita dapat merancang hubungan antara manusia dan mesin. Jika kita merancangnya sedemikian rupa sehingga manusia mempertahankan pemahaman, otoritas, dan otonomi yang cukup, bagian-bagian teknologi dari sistem tersebut dapat sangat memperbesar kemampuan manusia, memungkinkan setiap dari kita untuk berdiri di atas piramida kemampuan yang luas seperti dewa, jika Anda mau. Tetapi pertimbangkan pekerja di gudang pemenuhan pesanan belanja online. Dia lebih produktif daripada pendahulunya karena dia memiliki pasukan kecil robot yang membawakan tempat penyimpanan untuk mengambil barang; tetapi dia adalah bagian dari sistem yang lebih besar yang dikendalikan oleh algoritma cerdas yang memutuskan di mana dia harus berdiri dan barang mana yang harus dia ambil dan kirim. Dia sudah sebagian terkubur di dalam piramida, bukan berdiri di atasnya. Hanya masalah waktu sebelum pasir mengisi ruang-ruang di piramida dan perannya dihilangkan.

BAB 5
AI YANG TERLALU CERDAS

5.1 MASALAH GORILA: RISIKO AI SUPERCERDAS

Tidak perlu banyak imajinasi untuk melihat bahwa membuat sesuatu yang lebih cerdas daripada diri kita sendiri bisa menjadi ide yang buruk. Kita memahami bahwa kendali kita atas lingkungan dan spesies lain adalah hasil dari kecerdasan kita, jadi gagasan tentang sesuatu yang lain yang lebih cerdas daripada kita baik itu robot atau alien segera menimbulkan perasaan mual.

Sekitar sepuluh juta tahun yang lalu, nenek moyang gorila modern menciptakan (secara tidak sengaja, tentu saja) garis keturunan genetik yang mengarah ke manusia modern. Bagaimana perasaan gorila tentang hal ini? Jelas, jika mereka dapat memberi tahu kita tentang situasi spesies mereka saat ini dibandingkan dengan manusia, pendapat konsensus akan sangat negatif. Spesies mereka pada dasarnya tidak memiliki masa depan di luar apa yang kita izinkan. Kita tidak ingin berada dalam situasi serupa dibandingkan dengan mesin supercerdas. Saya akan menyebut ini sebagai masalah gorila khususnya, masalah apakah manusia dapat mempertahankan supremasi dan otonomi mereka di dunia yang mencakup mesin dengan kecerdasan yang jauh lebih besar.

Charles Babbage dan Ada Lovelace, yang merancang dan menulis program untuk Mesin Analitik pada tahun 1842, menyadari potensinya tetapi tampaknya tidak memiliki keraguan tentang hal itu. Namun, pada tahun 1847, Richard Thornton, editor *Primitive Expounder*, sebuah jurnal keagamaan, mengecam kalkulator mekanis:

> Pikiran... melampaui dirinya sendiri dan menghilangkan kebutuhan akan eksistensinya sendiri dengan menciptakan mesin untuk melakukan pemikirannya sendiri. Tetapi siapa yang tahu bahwa mesin-mesin tersebut, ketika disempurnakan lebih jauh, mungkin akan memikirkan rencana untuk memperbaiki semua kekurangannya sendiri dan kemudian menghasilkan ide-ide di luar jangkauan pikiran manusia fana!

Ini mungkin spekulasi pertama mengenai risiko eksistensial dari perangkat komputasi, tetapi tetap tersembunyi.

Sebaliknya, novel Samuel Butler, Erewhon, yang diterbitkan pada tahun 1872, mengembangkan tema ini dengan jauh lebih mendalam dan meraih kesuksesan langsung. Erewhon adalah sebuah negara di mana semua perangkat mekanik telah dilarang setelah perang saudara yang mengerikan antara kaum mekanik dan anti-mekanik. Satu bagian dari buku tersebut, yang disebut "Buku Mesin," menjelaskan asal usul perang ini dan menyajikan argumen dari kedua belah pihak. Ini sangat tepat menggambarkan perdebatan yang muncul kembali di awal abad ke-21.

Argumen utama kaum anti-mesin adalah bahwa mesin akan berkembang hingga titik di mana umat manusia kehilangan kendali:

> Bukankah kita sendiri sedang menciptakan penerus kita dalam supremasi bumi? Setiap hari menambah keindahan dan kehalusan organisasi mereka, setiap hari memberi mereka keterampilan yang lebih besar dan menyediakan semakin banyak kekuatan pengaturan diri yang akan lebih baik daripada kecerdasan apa pun? Dalam perjalanan waktu, kita akan mendapati diri kita sebagai ras yang lebih rendah. . . .
> Kita harus memilih antara alternatif mengalami banyak penderitaan saat ini, atau melihat diri kita secara bertahap digantikan oleh ciptaan kita sendiri, hingga kita tidak lebih tinggi dibandingkan dengan mereka, daripada binatang di ladang dibandingkan dengan diri kita sendiri. Perbudakan kita akan menyelinap kepada kita tanpa suara dan dengan pendekatan yang tak terlihat.

Narator juga menyampaikan argumen balasan utama kaum pro-mesin, yang mengantisipasi argumen simbiosis manusia-mesin yang akan kita bahas di bab berikutnya:

> Hanya ada satu upaya serius untuk menjawabnya. Penulisnya mengatakan bahwa mesin harus dianggap sebagai bagian dari sifat fisik manusia itu sendiri, yang sebenarnya hanyalah anggota tubuh di luar tubuh.

Meskipun kaum anti-mesin di Erewhon memenangkan perdebatan, Butler sendiri tampaknya memiliki dua pendapat. Di satu sisi, ia mengeluh bahwa "orang-orang Erewhon… cepat mengorbankan akal sehat di hadapan logika, ketika seorang filsuf muncul di antara mereka, yang memikat mereka melalui reputasinya dalam hal pengetahuan khusus" dan berkata, "Mereka bunuh diri dalam hal mesin." Di sisi lain, masyarakat Erewhon yang ia gambarkan sangat harmonis, produktif, dan bahkan idilis. Orang-orang Erewhon sepenuhnya menerima kebodohan untuk kembali memulai penemuan mekanik, dan memandang sisa-sisa mesin yang disimpan di museum "dengan perasaan seorang ahli barang antik Inggris mengenai monumen Druid atau ujung panah batu api."

Kisah Butler rupanya diketahui oleh Alan Turing, yang mempertimbangkan masa depan jangka panjang AI dalam sebuah kuliah yang diberikan di Manchester pada tahun 1951:

> Tampaknya mungkin bahwa begitu metode berpikir mesin dimulai, tidak akan butuh waktu lama untuk melampaui kemampuan kita yang lemah. Tidak akan ada pertanyaan tentang mesin yang mati, dan mereka akan dapat berkomunikasi satu sama lain untuk mempertajam kecerdasan mereka. Oleh karena itu, pada suatu tahap kita harus mengharapkan mesin untuk mengambil kendali, seperti yang disebutkan dalam Erewhon karya Samuel Butler.

Pada tahun yang sama, Turing mengulangi kekhawatiran ini dalam sebuah kuliah radio yang disiarkan di seluruh Inggris melalui Program Ketiga BBC:

> Jika sebuah mesin dapat berpikir, mungkin ia akan berpikir lebih cerdas daripada kita, lalu di mana posisi kita? Bahkan jika kita dapat menjaga mesin-mesin tersebut dalam posisi tunduk, misalnya dengan mematikan daya pada saat-saat strategis, kita, sebagai spesies, akan merasa sangat rendah diri. . . . Bahaya baru ini . . . tentu saja sesuatu yang dapat menimbulkan kecemasan bagi kita.

Ketika para anti-mesin Erewhon "merasa sangat gelisah tentang masa depan," mereka melihatnya sebagai "tugas mereka untuk menghentikan kejahatan selagi kita masih bisa melakukannya," dan mereka menghancurkan semua mesin. Tanggapan Turing terhadap "bahaya baru" dan "kecemasan" adalah mempertimbangkan "mematikan daya" (walaupun akan segera jelas bahwa ini sebenarnya bukan pilihan).

Dalam novel fiksi ilmiah klasik Frank Herbert, Dune, yang berlatar di masa depan yang jauh, umat manusia nyaris selamat dari Jihad Butlerian, perang dahsyat dengan "mesin-mesin berpikir." Sebuah perintah baru telah muncul: *"Janganlah engkau membuat mesin yang menyerupai pikiran manusia."* Perintah ini melarang perangkat komputasi dalam bentuk apa pun.

Semua respons drastis ini mencerminkan ketakutan yang belum terwujud yang ditimbulkan oleh kecerdasan mesin. Ya, prospek mesin supercerdas memang membuat kita gelisah. Ya, secara logis mungkin saja mesin-mesin tersebut dapat mengambil alih dunia dan menaklukkan atau melenyapkan umat manusia. Jika hanya itu yang bisa kita andalkan, maka satu-satunya respons yang masuk akal yang tersedia bagi kita saat ini adalah mencoba untuk membatasi penelitian kecerdasan buatan khususnya, melarang pengembangan dan penerapan sistem AI tingkat manusia yang serbaguna.

Seperti kebanyakan peneliti AI lainnya, saya merasa ngeri dengan prospek ini. Beraninya seseorang memberi tahu saya apa yang boleh dan tidak boleh saya pikirkan? Siapa pun yang mengusulkan penghentian penelitian AI harus melakukan banyak upaya untuk meyakinkan. Mengakhiri penelitian AI berarti mengabaikan bukan hanya salah satu jalan utama untuk memahami cara kerja kecerdasan manusia, tetapi juga kesempatan emas untuk meningkatkan kondisi manusia untuk menciptakan peradaban yang jauh lebih baik.

Nilai ekonomi AI tingkat manusia dapat diukur dalam ribuan triliun dolar, sehingga momentum di balik penelitian AI dari perusahaan dan pemerintah kemungkinan akan sangat besar. Hal itu akan mengalahkan keberatan yang samar-samar dari seorang filsuf, tidak peduli seberapa hebat "reputasinya untuk pembelajaran khusus," seperti yang dikatakan Butler.

Kelemahan kedua dari gagasan melarang AI tujuan umum adalah bahwa hal itu sulit untuk dilarang. Kemajuan pada AI tujuan umum terutama terjadi di papan tulis laboratorium penelitian di seluruh dunia, saat masalah matematika diajukan dan dipecahkan. Kita tidak tahu sebelumnya ide dan persamaan mana yang harus dilarang, dan, bahkan jika kita tahu,

tampaknya tidak masuk akal untuk mengharapkan bahwa larangan tersebut dapat ditegakkan atau efektif.

Untuk memperparah kesulitan lebih lanjut, para peneliti yang membuat kemajuan pada AI tujuan umum sering kali mengerjakan hal lain. Seperti yang telah saya kemukakan, penelitian tentang AI alat aplikasi spesifik dan tidak berbahaya seperti bermain game, diagnosis medis, dan perencanaan perjalanan seringkali mengarah pada kemajuan teknik tujuan umum yang dapat diterapkan pada berbagai masalah lain dan membawa kita lebih dekat ke AI tingkat manusia.

Karena alasan ini, sangat tidak mungkin komunitas AI atau pemerintah dan perusahaan yang mengendalikan hukum dan anggaran penelitian akan menanggapi masalah gorila dengan mengakhiri kemajuan dalam AI. Jika masalah gorila hanya dapat diselesaikan dengan cara ini, maka masalah itu tidak akan terselesaikan. Satu-satunya pendekatan yang tampaknya berhasil adalah memahami mengapa membuat AI yang lebih baik mungkin merupakan hal yang buruk. Ternyata kita telah mengetahui jawabannya selama ribuan tahun.

5.2 MASALAH RAJA MIDAS: KEGAGALAN KESELARASAN AI

Norbert Wiener, yang kita temui di Bab 1, memiliki dampak yang mendalam pada banyak bidang, termasuk kecerdasan buatan, ilmu kognitif, dan teori kontrol. Tidak seperti kebanyakan rekan sezamannya, ia sangat memperhatikan ketidakpastian sistem kompleks yang beroperasi di dunia nyata. (Ia menulis makalah pertamanya tentang topik ini pada usia sepuluh tahun.) Ia menjadi yakin bahwa rasa percaya diri yang berlebihan dari para ilmuwan dan insinyur dalam kemampuan mereka untuk mengendalikan ciptaan mereka, baik militer maupun sipil, dapat memiliki konsekuensi yang mengerikan.

Pada tahun 1950, Wiener menerbitkan *The Human Use of Human Beings*, yang pada sampul depannya tertulis, "'Otak mekanis' dan mesin serupa dapat menghancurkan nilai-nilai manusia atau memungkinkan kita untuk mewujudkannya seperti belum pernah terjadi sebelumnya." Ia secara bertahap menyempurnakan idenya dari waktu ke waktu dan pada tahun 1960 telah mengidentifikasi satu masalah inti: ketidakmungkinan untuk mendefinisikan tujuan manusia yang sebenarnya dengan benar dan lengkap. Ini, pada gilirannya, berarti bahwa apa yang saya sebut model standar di mana manusia mencoba untuk menanamkan tujuan mereka sendiri ke dalam mesin ditakdirkan untuk gagal.

Kita dapat menyebut ini sebagai masalah Raja Midas: Midas, seorang raja legendaris dalam mitologi Yunani kuno, mendapatkan persis apa yang dia minta yaitu, bahwa segala sesuatu yang disentuhnya harus berubah menjadi emas. Terlambat, ia menemukan bahwa ini termasuk makanan, minuman, dan anggota keluarganya, dan ia meninggal dalam kesengsaraan dan kelaparan. Tema yang sama tersebar luas dalam mitologi manusia. Wiener mengutip kisah Goethe tentang murid penyihir, yang memerintahkan sapu untuk mengambil air tetapi tidak mengatakan berapa banyak air dan tidak tahu bagaimana membuat sapu berhenti.

Secara teknis, kita mungkin mengalami kegagalan keselarasan nilai kita mungkin, tanpa sengaja, menanamkan tujuan pada mesin yang tidak sepenuhnya selaras dengan tujuan kita

sendiri. Hingga baru-baru ini, kita terlindungi dari konsekuensi yang berpotensi menimbulkan bencana oleh kemampuan terbatas mesin cerdas dan ruang lingkup terbatas yang mereka miliki untuk memengaruhi dunia. (Memang, sebagian besar pekerjaan AI dilakukan dengan masalah mainan di laboratorium penelitian.) Seperti yang dikatakan Norbert Wiener dalam bukunya tahun 1964, *God and Golem*,

> Di masa lalu, pandangan parsial dan tidak memadai tentang tujuan manusia relatif tidak berbahaya hanya karena disertai dengan keterbatasan teknis. Ini hanyalah salah satu dari banyak tempat di mana ketidakberdayaan manusia telah melindungi kita dari dampak destruktif penuh dari kebodohan manusia.

Sayangnya, periode perlindungan ini akan segera berakhir.

Kita telah melihat bagaimana algoritma pemilihan konten di media sosial menimbulkan kekacauan di masyarakat atas nama memaksimalkan pendapatan iklan. Jika Anda berpikir bahwa memaksimalkan pendapatan iklan sudah merupakan tujuan yang tidak mulia dan seharusnya tidak pernah dikejar, mari kita bayangkan jika kita meminta sistem supercerdas di masa depan untuk mengejar tujuan mulia yaitu menemukan obat untuk kanker idealnya secepat mungkin, karena seseorang meninggal karena kanker setiap 3,5 detik.

Dalam beberapa jam, sistem AI telah membaca seluruh literatur biomedis dan membuat hipotesis jutaan senyawa kimia yang berpotensi efektif tetapi belum pernah diuji sebelumnya. Dalam beberapa minggu, sistem tersebut telah menginduksi berbagai jenis tumor pada setiap manusia yang hidup untuk melakukan uji coba medis terhadap senyawa-senyawa ini, yang merupakan cara tercepat untuk menemukan obat. Ups.

Jika Anda lebih suka memecahkan masalah lingkungan, Anda mungkin meminta mesin tersebut untuk mengatasi pengasaman laut yang cepat akibat peningkatan kadar karbon dioksida. Mesin tersebut mengembangkan katalis baru yang memfasilitasi reaksi kimia yang sangat cepat antara laut dan atmosfer dan mengembalikan tingkat pH laut. Sayangnya, seperempat oksigen di atmosfer habis dalam proses tersebut, membuat kita mati lemas perlahan dan menyakitkan. Ups.

Skenario kiamat semacam ini tidak halus seperti yang mungkin diharapkan, untuk skenario kiamat. Tetapi ada banyak skenario di mana semacam mati lemas mental "menyelinap kepada kita tanpa suara dan dengan pendekatan yang tak terlihat." Prolog buku Max Tegmark, *Life 3.0*, menjelaskan secara rinci sebuah skenario di mana mesin supercerdas secara bertahap mengambil alih kendali ekonomi dan politik atas seluruh dunia sambil tetap pada dasarnya tidak terdeteksi. Internet dan mesin-mesin berskala global yang didukungnya yang sudah berinteraksi dengan miliaran "pengguna" setiap hari menyediakan media yang sempurna untuk pertumbuhan kendali mesin atas manusia.

Saya tidak mengharapkan bahwa tujuan yang dimasukkan ke dalam mesin-mesin tersebut akan berupa jenis "mengambil alih dunia". Kemungkinan besar tujuannya adalah memaksimalkan keuntungan atau memaksimalkan keterlibatan, atau bahkan tujuan yang

tampaknya tidak berbahaya seperti mencapai skor lebih tinggi pada survei kepuasan pengguna reguler atau mengurangi penggunaan energi kita.

Sekarang, jika kita menganggap diri kita sebagai entitas yang tindakannya diharapkan untuk mencapai tujuan kita, ada dua cara untuk mengubah perilaku kita. Yang pertama adalah cara kuno: biarkan harapan dan tujuan kita tidak berubah, tetapi ubah keadaan kita misalnya, dengan menawarkan uang, menodongkan pistol kepada kita, atau membuat kita kelaparan hingga menyerah. Itu cenderung mahal dan sulit dilakukan oleh komputer. Cara kedua adalah mengubah harapan dan tujuan kita. Ini jauh lebih mudah bagi mesin. Mesin tersebut berhubungan dengan Anda selama berjam-jam setiap hari, mengontrol akses Anda ke informasi, dan menyediakan sebagian besar hiburan Anda melalui permainan, TV, film, dan interaksi sosial.

Algoritma pembelajaran penguatan yang mengoptimalkan klik-tayang media sosial tidak memiliki kapasitas untuk bernalar tentang perilaku manusia bahkan, mereka bahkan tidak tahu dalam arti yang berarti bahwa manusia itu ada. Bagi mesin yang memiliki pemahaman jauh lebih besar tentang psikologi, kepercayaan, dan motivasi manusia, seharusnya relatif mudah untuk secara bertahap membimbing kita ke arah yang meningkatkan tingkat kepuasan tujuan mesin. Misalnya, mesin tersebut dapat mengurangi konsumsi energi kita dengan membujuk kita untuk memiliki lebih sedikit anak, yang pada akhirnya dan tanpa disengaja mencapai impian para filsuf anti-natalis yang ingin menghilangkan dampak buruk umat manusia terhadap dunia alami.

Dengan sedikit latihan, Anda dapat belajar mengidentifikasi cara-cara di mana pencapaian hampir semua tujuan tetap dapat menghasilkan hasil yang buruk secara sewenang-wenang. Salah satu pola yang paling umum melibatkan pengabaian sesuatu dari tujuan yang sebenarnya Anda pedulikan. Dalam kasus seperti itu seperti dalam contoh yang diberikan di atas sistem AI sering kali akan menemukan solusi optimal yang menetapkan hal yang Anda pedulikan, tetapi lupa Anda sebutkan, ke nilai ekstrem. Jadi, jika Anda berkata kepada mobil otonom Anda, "Bawa saya ke bandara secepat mungkin!" dan mobil tersebut menafsirkannya secara harfiah, mobil tersebut akan mencapai kecepatan 180 mil per jam dan Anda akan masuk penjara. (Untungnya, mobil otonom yang saat ini sedang dipertimbangkan tidak akan menerima permintaan seperti itu.) Jika Anda mengatakan, "Antarkan saya ke bandara secepat mungkin tanpa melebihi batas kecepatan," mobil akan berakselerasi dan mengerem sekeras mungkin, bermanuver di antara lalu lintas untuk mempertahankan kecepatan maksimum di antaranya. Mobil bahkan mungkin akan mendorong mobil lain keluar dari jalan untuk mendapatkan beberapa detik dalam keramaian di terminal bandara. Dan seterusnya pada akhirnya, Anda akan menambahkan cukup banyak pertimbangan sehingga cara mengemudi mobil tersebut kurang lebih menyerupai pengemudi manusia yang terampil yang mengantar seseorang ke bandara dengan sedikit terburu-buru.

Mengemudi adalah tugas sederhana dengan dampak lokal saja, dan sistem AI yang saat ini sedang dibangun untuk mengemudi tidak terlalu cerdas. Karena alasan ini, banyak potensi kegagalan dapat diantisipasi; yang lain akan terungkap dalam simulator mengemudi atau dalam jutaan mil pengujian dengan pengemudi profesional yang siap mengambil alih jika

terjadi kesalahan; yang lain lagi baru akan muncul kemudian, ketika mobil sudah berada di jalan dan sesuatu yang aneh terjadi.

Sayangnya, dengan sistem supercerdas yang dapat berdampak global, tidak ada simulator dan tidak ada kesempatan untuk mengulang. Tentu sangat sulit, dan mungkin mustahil, bagi manusia biasa untuk mengantisipasi dan mengesampingkan terlebih dahulu semua cara buruk yang dapat dipilih mesin untuk mencapai tujuan tertentu. Secara umum, jika Anda memiliki satu tujuan dan mesin supercerdas memiliki tujuan yang berbeda dan bertentangan, mesin akan mendapatkan apa yang diinginkannya dan Anda tidak.

5.3 KETAKUTAN DAN KESERAKAHAN: TUJUAN INSTRUMENTAL

Jika mesin yang mengejar tujuan yang salah terdengar cukup buruk, masih ada yang lebih buruk. Solusi yang disarankan oleh Alan Turing mematikan daya pada saat-saat strategis mungkin tidak tersedia, karena alasan yang sangat sederhana: Anda tidak dapat mengambil kopi jika Anda sudah mati.

Izinkan saya menjelaskan. Misalkan sebuah mesin memiliki tujuan untuk mengambil kopi. Jika mesin tersebut cukup cerdas, ia pasti akan memahami bahwa ia akan gagal dalam tujuannya jika dimatikan sebelum menyelesaikan misinya. Dengan demikian, tujuan mengambil kopi menciptakan, sebagai sub-tujuan yang diperlukan, tujuan untuk menonaktifkan sakelar mati. Hal yang sama berlaku untuk menyembuhkan kanker atau menghitung angka pi. Sebenarnya tidak banyak yang dapat Anda lakukan setelah Anda mati, jadi kita dapat mengharapkan sistem AI untuk bertindak secara preventif untuk mempertahankan keberadaan mereka sendiri, mengingat tujuan yang kurang lebih pasti.

Jika tujuan tersebut bertentangan dengan preferensi manusia, maka kita akan mendapatkan plot yang persis sama dengan film *2001: A Space Odyssey*, di mana komputer HAL 9000 membunuh empat dari lima astronot di dalam pesawat untuk mencegah gangguan terhadap misinya. Dave, astronot terakhir yang tersisa, berhasil mematikan HAL setelah pertarungan kecerdasan yang epik mungkin untuk menjaga agar plot tetap menarik. Tetapi jika HAL benar-benar supercerdas, Dave pasti sudah dimatikan.

Penting untuk dipahami bahwa naluri mempertahankan diri tidak harus berupa insting bawaan atau arahan utama dalam mesin. (Jadi Hukum Ketiga Robotika Isaac Asimov, yang dimulai dengan "Robot harus melindungi keberadaannya sendiri," sama sekali tidak diperlukan.) Tidak perlu membangun naluri mempertahankan diri karena itu adalah tujuan instrumental tujuan yang merupakan sub-tujuan yang berguna dari hampir semua tujuan awal. Setiap entitas yang memiliki tujuan yang pasti akan secara otomatis bertindak seolah-olah ia juga memiliki tujuan instrumental.

Selain hidup, memiliki akses ke uang adalah tujuan instrumental dalam sistem kita saat ini. Dengan demikian, mesin cerdas mungkin menginginkan uang, bukan karena serakah tetapi karena uang berguna untuk mencapai berbagai macam tujuan. Dalam film *Transcendence*, ketika otak Johnny Depp diunggah ke superkomputer kuantum, hal pertama yang dilakukan mesin itu adalah menyalin dirinya sendiri ke jutaan komputer lain di Internet sehingga tidak

dapat dimatikan. Hal kedua yang dilakukannya adalah menghasilkan keuntungan besar di pasar saham untuk mendanai rencana ekspansinya.

Dan apa sebenarnya rencana ekspansi itu? Rencana tersebut mencakup merancang dan membangun superkomputer kuantum yang jauh lebih besar; melakukan penelitian AI; dan menemukan pengetahuan baru tentang fisika, ilmu saraf, dan biologi. Tujuan sumber daya ini daya komputasi, algoritma, dan pengetahuan juga merupakan tujuan instrumental, berguna untuk mencapai tujuan menyeluruh apa pun. Hal ini tampak cukup tidak berbahaya sampai seseorang menyadari bahwa proses akuisisi akan terus berlanjut tanpa batas. Ini tampaknya menciptakan konflik yang tak terhindarkan dengan manusia. Dan tentu saja, mesin tersebut, yang dilengkapi dengan model pengambilan keputusan manusia yang semakin canggih, akan mengantisipasi dan menggagalkan setiap langkah kita dalam konflik ini.

5.4 LEDAKAN KECERDASAN

I. J. Good adalah seorang matematikawan brilian yang bekerja dengan Alan Turing di Bletchley Park, memecahkan kode Jerman selama Perang Dunia II. Ia memiliki minat yang sama dengan Turing dalam kecerdasan mesin dan inferensi statistik. Pada tahun 1965, ia menulis makalah yang sekarang paling terkenal, "Spekulasi Mengenai Mesin Ultracerdas Pertama." Kalimat pertama menunjukkan bahwa Good, yang khawatir dengan ancaman nuklir Perang Dingin, menganggap AI sebagai penyelamat potensial bagi umat manusia: "Kelangsungan hidup manusia bergantung pada pembangunan awal mesin ultracerdas." Namun, seiring berjalannya makalah, ia menjadi lebih berhati-hati. Ia memperkenalkan gagasan tentang ledakan kecerdasan, tetapi, seperti Butler, Turing, dan Wiener sebelumnya, ia khawatir kehilangan kendali:

> Biarlah mesin ultracerdas didefinisikan sebagai mesin yang dapat jauh melampaui semua aktivitas intelektual manusia mana pun, betapapun cerdasnya. Karena desain mesin adalah salah satu aktivitas intelektual ini, mesin ultracerdas dapat merancang mesin yang lebih baik lagi; Maka, tak diragukan lagi akan terjadi "ledakan kecerdasan," dan kecerdasan manusia akan jauh tertinggal. Dengan demikian, mesin ultra-cerdas pertama adalah penemuan terakhir yang perlu dibuat manusia, asalkan mesin tersebut cukup jinak untuk memberi tahu kita cara mengendalikannya. Anehnya, poin ini jarang dikemukakan di luar fiksi ilmiah.

Paragraf ini merupakan pokok bahasan dalam setiap diskusi tentang AI supercerdas, meskipun peringatan di bagian akhir biasanya diabaikan. Poin Good dapat diperkuat dengan mencatat bahwa mesin ultra-cerdas tidak hanya dapat meningkatkan desainnya sendiri; kemungkinan besar ia akan melakukannya karena, seperti yang telah kita lihat, mesin cerdas mengharapkan manfaat dari peningkatan perangkat keras dan perangkat lunaknya. Kemungkinan ledakan kecerdasan sering disebut sebagai sumber risiko utama bagi umat manusia dari AI karena akan memberi kita waktu yang sangat sedikit untuk menyelesaikan masalah pengendalian.

Argumen Good tentu memiliki plausibilitas melalui analogi alami dengan ledakan kimia di mana setiap reaksi molekuler melepaskan energi yang cukup untuk memulai lebih dari satu reaksi tambahan. Di sisi lain, secara logis mungkin terjadi penurunan hasil dalam peningkatan kecerdasan, sehingga prosesnya meredup daripada meledak. Tidak ada cara yang jelas untuk membuktikan bahwa ledakan pasti akan terjadi.

Skenario penurunan hasil itu sendiri menarik. Hal itu dapat muncul jika ternyata mencapai peningkatan persentase tertentu menjadi jauh lebih sulit seiring dengan meningkatnya kecerdasan mesin. (Saya berasumsi demi argumen bahwa kecerdasan mesin tujuan umum dapat diukur pada semacam skala linier, yang saya ragukan akan pernah benar-benar tepat.) Dalam hal itu, manusia juga tidak akan mampu menciptakan superintelijen. Jika mesin yang sudah superhuman kehabisan tenaga ketika mencoba meningkatkan kecerdasannya sendiri, maka manusia akan kehabisan tenaga lebih cepat lagi.

Sekarang, saya belum pernah mendengar argumen serius yang menyatakan bahwa menciptakan tingkat kecerdasan mesin tertentu berada di luar kemampuan kecerdasan manusia, tetapi saya kira kita harus mengakui bahwa itu secara logis mungkin. “Secara logis mungkin” dan “Saya bersedia mempertaruhkan masa depan umat manusia untuk itu” tentu saja adalah dua hal yang sangat berbeda. Bertaruh melawan kecerdasan manusia tampaknya merupakan strategi yang merugikan.

Jika ledakan kecerdasan terjadi, dan jika kita belum memecahkan masalah pengendalian mesin dengan kecerdasan yang sedikit di atas kemampuan manusia misalnya, jika kita tidak dapat mencegah mereka melakukan peningkatan diri secara rekursif maka kita tidak akan punya waktu lagi untuk memecahkan masalah pengendalian dan permainan akan berakhir. Ini adalah skenario lepas landas keras Bostrom, di mana kecerdasan mesin meningkat secara astronomis hanya dalam beberapa hari atau minggu. Dalam kata-kata Turing, itu “tentu saja sesuatu yang dapat membuat kita cemas.”

Tanggapan yang mungkin terhadap kecemasan ini tampaknya adalah mundur dari penelitian AI, menyangkal bahwa ada risiko yang melekat dalam pengembangan AI tingkat lanjut, memahami dan mengurangi risiko melalui desain sistem AI yang tetap berada di bawah kendali manusia, dan menyerah hanya menyerahkan masa depan kepada mesin cerdas.

Penyangkalan dan mitigasi adalah subjek dari sisa buku ini. Seperti yang telah saya kemukakan, mundur dari penelitian AI tidak mungkin terjadi (karena manfaat yang hilang terlalu besar) dan sangat sulit untuk diwujudkan. Sikap pasrah tampaknya merupakan respons terburuk. Hal ini sering disertai dengan gagasan bahwa sistem AI yang lebih cerdas daripada kita entah bagaimana pantas mewarisi planet ini, meninggalkan manusia untuk pergi dengan tenang ke malam yang indah itu, terhibur oleh pemikiran bahwa keturunan elektronik kita yang brilian sedang sibuk mengejar tujuan mereka. Pandangan ini dipromosikan oleh ahli robotika dan futuris Hans Moravec, yang menulis, “Luasnya dunia maya akan dipenuhi dengan pikiran super yang bukan manusia, terlibat dalam urusan yang menjadi perhatian manusia seperti halnya urusan kita bagi bakteri.” Ini tampaknya merupakan kesalahan. Nilai, bagi manusia, terutama didefinisikan oleh pengalaman manusia yang sadar. Jika tidak ada manusia

dan tidak ada entitas sadar lainnya yang pengalaman subjektifnya penting bagi kita, maka tidak ada hal yang bernilai yang terjadi.

BAB 6
DEBAT AI YANG TIDAK BEGITU HEBAT

Implikasi dari memperkenalkan spesies cerdas kedua ke Bumi cukup luas untuk layak dipikirkan secara mendalam." Demikianlah ulasan majalah *The Economist* tentang buku *Superintelligence* karya Nick Bostrom. Kebanyakan orang akan menafsirkan ini sebagai contoh klasik dari pernyataan yang meremehkan. Tentu, Anda mungkin berpikir, para pemikir hebat saat ini sudah melakukan pemikiran mendalam ini terlibat dalam debat serius, menimbang risiko dan manfaat, mencari solusi, menggali celah dalam solusi, dan sebagainya. Belum, sejauh yang saya ketahui.

Ketika seseorang pertama kali memperkenalkan ide-ide ini kepada audiens teknis, orang dapat melihat gelembung pikiran muncul dari kepala mereka, dimulai dengan kata-kata "Tetapi, tetapi, tetapi . . ." dan diakhiri dengan tanda seru. Jenis "tetapi" yang pertama berbentuk penyangkalan. Para penyangkal mengatakan, "Tetapi ini tidak mungkin menjadi masalah nyata, karena XYZ." Beberapa dari XYZ mencerminkan proses penalaran yang mungkin dapat digambarkan sebagai angan-angan belaka, sementara yang lain lebih substansial. Jenis "tetapi" kedua berbentuk pengalihan: menerima bahwa masalahnya nyata tetapi berpendapat bahwa kita tidak boleh mencoba menyelesaikannya, baik karena tidak dapat dipecahkan atau karena ada hal-hal yang lebih penting untuk difokuskan daripada akhir peradaban atau karena lebih baik tidak menyebutkannya sama sekali. Jenis "tetapi" ketiga berbentuk solusi instan yang terlalu disederhanakan: "Tapi bukankah kita bisa melakukan ABC saja?" Seperti halnya penyangkalan, beberapa dari ABC (*Alternative Basic Statements*) langsung disesalkan. Yang lain, mungkin secara tidak sengaja, lebih mendekati identifikasi sifat sebenarnya dari masalah tersebut.

Saya tidak bermaksud menyarankan bahwa tidak mungkin ada keberatan yang masuk akal terhadap pandangan bahwa mesin supercerdas yang dirancang dengan buruk akan menimbulkan risiko serius bagi umat manusia. Hanya saja saya belum melihat keberatan seperti itu. Karena masalah ini tampaknya sangat penting, maka layak untuk diperdebatkan secara publik dengan kualitas terbaik. Jadi, demi adanya debat tersebut, dan dengan harapan pembaca akan berkontribusi di dalamnya, izinkan saya memberikan tur singkat tentang poin-poin penting sejauh ini, seperti apa adanya.

6.1 ARGUMEN KELIRU: MENGAPA PENYANGKALAN RISIKO AI GAGAL

Menyangkal bahwa masalah itu ada sama sekali adalah jalan keluar termudah. Scott Alexander, penulis blog *Slate Star Codex*, memulai sebuah artikel terkenal tentang risiko AI sebagai berikut: "Saya pertama kali tertarik pada risiko AI sekitar tahun 2007. Pada saat itu, sebagian besar tanggapan orang terhadap topik tersebut adalah 'Haha, kembalilah ketika ada yang percaya ini selain orang-orang gila di internet.'"

Pernyataan yang Langsung Disesalkan

Ancaman yang dirasakan terhadap profesi seumur hidup seseorang dapat menyebabkan orang yang cerdas dan biasanya bijaksana mengatakan hal-hal yang mungkin ingin mereka tarik kembali setelah analisis lebih lanjut. Karena itu, saya tidak akan menyebutkan nama penulis argumen berikut, yang semuanya adalah peneliti AI terkenal. Saya telah menyertakan sanggahan terhadap argumen tersebut, meskipun sebenarnya tidak perlu.

- Kalkulator elektronik sangat hebat dalam berhitung. Kalkulator tidak mengambil alih dunia; oleh karena itu, tidak ada alasan untuk khawatir tentang AI yang super-manusiawi.
  - Sangkalan: kecerdasan tidak sama dengan kemampuan berhitung, dan kemampuan berhitung para kalkulator tidak membekali mereka untuk menguasai dunia.
- Kuda memiliki kekuatan luar biasa, dan kita tidak perlu khawatir membuktikan bahwa kuda aman; jadi kita tidak perlu khawatir membuktikan bahwa sistem AI aman.
  - Sangkalan: kecerdasan tidak sama dengan kekuatan fisik, dan kekuatan kuda tidak membuat mereka mampu menguasai dunia.
- Secara historis, tidak ada contoh mesin yang membunuh jutaan manusia, jadi, secara induksi, hal itu tidak mungkin terjadi di masa depan.
  - Sangkalan: selalu ada pertama kalinya untuk segala sesuatu, sebelum itu tidak ada contoh kejadian tersebut.
- Tidak ada besaran fisik di alam semesta yang tak terbatas, dan itu termasuk kecerdasan, jadi kekhawatiran tentang superintelijen terlalu berlebihan.
  - Sangkalan: superintelijen tidak perlu tak terbatas untuk menjadi masalah; dan fisika memungkinkan perangkat komputasi miliaran kali lebih kuat daripada otak manusia.
- Kita tidak khawatir tentang kemungkinan yang mengakhiri spesies tetapi sangat tidak mungkin terjadi seperti lubang hitam yang muncul di orbit dekat Bumi, jadi mengapa khawatir tentang AI superintelijen?
  - Sangkalan: jika sebagian besar fisikawan di Bumi bekerja untuk membuat lubang hitam semacam itu, bukankah kita akan bertanya kepada mereka apakah itu aman?

Kecerdasan bukan Satu Dimensi: Mengapa IQ Mesin Bermasalah

Merupakan prinsip dasar psikologi modern bahwa satu angka IQ saja tidak dapat menggambarkan kekayaan kecerdasan manusia secara penuh. Teori tersebut mengatakan bahwa ada berbagai dimensi kecerdasan: spasial, logis, linguistik, sosial, dan sebagainya. Alice, pemain sepak bola kita dari Bab 2, mungkin memiliki kecerdasan spasial yang lebih tinggi daripada temannya Bob, tetapi kecerdasan sosialnya lebih rendah. Dengan demikian, kita tidak dapat mengurutkan semua manusia berdasarkan tingkat kecerdasan mereka.

Hal ini bahkan lebih benar untuk mesin, karena kemampuan mereka jauh lebih sempit. Mesin pencari Google dan AlphaGo hampir tidak memiliki kesamaan, selain sebagai produk dari dua anak perusahaan dari perusahaan induk yang sama, sehingga tidak masuk akal untuk

mengatakan bahwa yang satu lebih cerdas daripada yang lain. Hal ini membuat gagasan tentang "IQ mesin" menjadi bermasalah dan menunjukkan bahwa menyesatkan untuk menggambarkan masa depan sebagai perlombaan IQ satu dimensi antara manusia dan mesin.

Kevin Kelly, pendiri dan editor majalah Wired serta komentator teknologi yang sangat jeli, membawa argumen ini selangkah lebih jauh. Dalam “Mitos AI Superhuman,” ia menulis, “Kecerdasan bukanlah satu dimensi tunggal, jadi ‘lebih pintar dari manusia’ adalah konsep yang tidak berarti.” Dalam sekejap, semua kekhawatiran tentang superintelijen lenyap.

Sekarang, satu tanggapan yang jelas adalah bahwa mesin dapat melampaui kemampuan manusia dalam semua dimensi kecerdasan yang relevan. Dalam hal itu, bahkan menurut standar ketat Kelly, mesin tersebut akan lebih pintar dari manusia. Tetapi asumsi yang agak kuat ini tidak diperlukan untuk membantah argumen Kelly. Pertimbangkan simpanse. Simpanse mungkin memiliki memori jangka pendek yang lebih baik daripada manusia, bahkan pada tugas-tugas yang berorientasi pada manusia seperti mengingat urutan angka. Memori jangka pendek adalah dimensi penting dari kecerdasan.

Dengan argumen Kelly, maka manusia tidak lebih pintar dari simpanse; memang, ia akan mengklaim bahwa “lebih pintar dari simpanse” adalah konsep yang tidak berarti. Ini bukanlah penghiburan yang berarti bagi simpanse (dan bonobo, gorila, orangutan, paus, lumba-lumba, dan sebagainya) yang spesiesnya bertahan hidup hanya karena kita mengizinkannya. Ini bahkan lebih tidak menghibur lagi bagi semua spesies yang telah kita musnahkan. Ini juga bukan penghiburan yang berarti bagi manusia yang mungkin khawatir akan dimusnahkan oleh mesin.

Skeptisisme Abadi: AI Supermanusia Tak Mungkin?

Bahkan sebelum kelahiran AI pada tahun 1956, para intelektual terkemuka menggerutu dan mengatakan bahwa mesin cerdas itu tidak mungkin. Alan Turing mencurahkan sebagian besar makalah pentingnya tahun 1950, "Computing Machinery and Intelligence," untuk membantah argumen-argumen ini. Sejak saat itu, komunitas AI telah menangkis klaim ketidakmungkinan serupa dari para filsuf, matematikawan, dan lainnya. Dalam debat saat ini tentang superintelijen, beberapa filsuf telah menggali kembali klaim ketidakmungkinan ini untuk membuktikan bahwa umat manusia tidak perlu takut. Ini tidak mengejutkan.

Studi Seratus Tahun tentang Kecerdasan Buatan, atau AI100, adalah proyek jangka panjang yang ambisius yang berada di Universitas Stanford. Tujuannya adalah untuk melacak AI, atau, lebih tepatnya, untuk "mempelajari dan mengantisipasi bagaimana efek kecerdasan buatan akan menyebar ke setiap aspek bagaimana orang bekerja, hidup, dan bermain." Laporan utama pertamanya, “Kecerdasan Buatan dan Kehidupan di Tahun 2030,” memang mengejutkan. Seperti yang diharapkan, laporan ini menekankan manfaat AI di bidang-bidang seperti diagnosis medis dan keselamatan otomotif. Yang tidak terduga adalah klaim bahwa “tidak seperti di film, tidak ada ras robot supermanusia di cakrawala atau bahkan mungkin tidak akan pernah ada.”

Sepengetahuan saya, ini adalah pertama kalinya para peneliti AI serius secara terbuka menyatakan pandangan bahwa AI tingkat manusia atau supermanusia tidak mungkin dan ini di tengah periode kemajuan yang sangat pesat dalam penelitian AI, ketika berbagai hambatan

terus ditembus. Seolah-olah sekelompok ahli biologi kanker terkemuka mengumumkan bahwa mereka telah menipu kita selama ini: mereka selalu tahu bahwa tidak akan pernah ada obat untuk kanker.

Apa yang bisa memotivasi perubahan haluan seperti itu? Laporan tersebut tidak memberikan argumen atau bukti apa pun. (Memang, bukti apa yang mungkin ada bahwa tidak ada susunan atom yang secara fisik mungkin mengungguli otak manusia?) Saya menduga ada dua alasan. Alasan pertama adalah keinginan alami untuk membuktikan bahwa masalah gorila tidak ada, yang menghadirkan prospek yang sangat tidak nyaman bagi peneliti AI; tentu saja, jika AI tingkat manusia tidak mungkin, masalah gorila akan terselesaikan dengan mudah. Alasan kedua adalah tribalisme naluri untuk bersatu melawan apa yang dianggap sebagai "serangan" terhadap AI.

Tampaknya aneh untuk menganggap klaim bahwa AI supercerdas itu mungkin sebagai serangan terhadap AI, dan bahkan lebih aneh lagi untuk membela AI dengan mengatakan bahwa AI tidak akan pernah berhasil dalam tujuannya. Kita tidak dapat menjamin terhadap bencana di masa depan hanya dengan bertaruh melawan kecerdasan manusia.

Kita telah melakukan taruhan seperti itu sebelumnya dan kalah. Seperti yang kita lihat sebelumnya, kalangan fisika pada awal tahun 1930-an, yang diwakili oleh Lord Rutherford, dengan yakin percaya bahwa mengekstraksi energi atom adalah hal yang mustahil; namun penemuan Leo Szilard tentang reaksi rantai nuklir yang diinduksi neutron pada tahun 1933 membuktikan bahwa keyakinan itu salah tempat.

Terobosan Szilard datang pada waktu yang tidak menguntungkan: awal dari perlombaan senjata dengan Nazi Jerman. Tidak ada kemungkinan mengembangkan teknologi nuklir untuk kebaikan bersama. Beberapa tahun kemudian, setelah mendemonstrasikan reaksi berantai nuklir di laboratoriumnya, Szilard menulis, “Kami mematikan semuanya dan pulang. Malam itu, hampir tidak ada keraguan dalam pikiran saya bahwa dunia sedang menuju ke arah kesedihan.”

Terlalu Dini untuk Mengkhawatirkannya

Biasanya kita melihat orang-orang yang berpikiran jernih berusaha meredakan kekhawatiran publik dengan menunjukkan bahwa karena AI tingkat manusia kemungkinan tidak akan muncul selama beberapa dekade, tidak ada yang perlu dikhawatirkan. Misalnya, laporan AI100 mengatakan bahwa “tidak ada alasan untuk khawatir bahwa AI merupakan ancaman langsung bagi umat manusia.”

Argumen ini gagal dalam dua hal. Pertama, argumen ini menyerang argumen yang keliru. Alasan kekhawatiran tidak didasarkan pada ancaman langsung. Misalnya, Nick Bostrom menulis dalam *Superintelligence*, “Bukan bagian dari argumen dalam buku ini bahwa kita berada di ambang terobosan besar dalam kecerdasan buatan, atau bahwa kita dapat memprediksi dengan tepat kapan perkembangan seperti itu mungkin terjadi.” Kedua, risiko jangka panjang tetap dapat menjadi alasan untuk kekhawatiran segera. Waktu yang tepat untuk mengkhawatirkan masalah yang berpotensi serius bagi umat manusia tidak hanya bergantung pada kapan masalah itu akan terjadi, tetapi juga pada berapa lama waktu yang dibutuhkan untuk mempersiapkan dan menerapkan solusi.

Sebagai contoh, jika kita mendeteksi asteroid besar yang akan bertabrakan dengan Bumi pada tahun 2069, apakah kita akan mengatakan bahwa terlalu dini untuk khawatir? Justru sebaliknya! Akan ada proyek darurat di seluruh dunia untuk mengembangkan cara-cara untuk menanggulangi ancaman tersebut. Kita tidak akan menunggu hingga tahun 2068 untuk mulai mengerjakan solusi, karena kita tidak dapat mengatakan sebelumnya berapa banyak waktu yang dibutuhkan. Memang, proyek Pertahanan Planet NASA sudah mengerjakan solusi yang mungkin, meskipun "tidak ada asteroid yang diketahui menimbulkan risiko signifikan untuk bertabrakan dengan Bumi selama 100 tahun ke depan." Jika itu membuat Anda merasa puas, mereka juga mengatakan, "Sekitar 74 persen objek dekat Bumi yang lebih besar dari 460 kaki masih belum ditemukan." Dan jika kita mempertimbangkan risiko bencana global dari perubahan iklim, yang diprediksi akan terjadi di akhir abad ini, apakah terlalu dini untuk mengambil tindakan untuk mencegahnya? Sebaliknya, mungkin sudah terlambat. Skala waktu yang relevan untuk AI supermanusia kurang dapat diprediksi, tetapi tentu saja itu berarti, seperti fisi nuklir, hal itu mungkin akan tiba jauh lebih cepat dari yang diperkirakan.

Salah satu formulasi argumen "masih terlalu dini untuk khawatir" yang semakin populer adalah pernyataan Andrew Ng bahwa "itu seperti mengkhawatirkan kelebihan populasi di Mars."[11] (Ia kemudian meningkatkannya dari Mars menjadi Alpha Centauri.) Ng, mantan profesor Stanford, adalah pakar terkemuka di bidang pembelajaran mesin, dan pandangannya memiliki bobot tertentu. Pernyataan tersebut menggunakan analogi yang mudah: tidak hanya risikonya mudah dikelola dan masih jauh di masa depan, tetapi juga sangat tidak mungkin kita akan mencoba memindahkan miliaran manusia ke Mars sejak awal.

Namun, analogi tersebut keliru. Kita sudah mencurahkan sumber daya ilmiah dan teknis yang sangat besar untuk menciptakan sistem AI yang semakin mumpuni, dengan sedikit sekali pertimbangan tentang apa yang terjadi jika kita berhasil. Analogi yang lebih tepat, kemudian, adalah mengerjakan rencana untuk memindahkan umat manusia ke Mars tanpa mempertimbangkan apa yang mungkin kita hirup, minum, atau makan begitu kita tiba. Beberapa orang mungkin menyebut rencana ini tidak bijaksana.

Sebagai alternatif, kita dapat menafsirkan pernyataan Ng secara harfiah, dan menjawab bahwa mendaratkan bahkan satu orang pun di Mars akan menyebabkan kelebihan populasi, karena Mars memiliki daya tampung nol. Dengan demikian, kelompok-kelompok yang saat ini berencana mengirim beberapa manusia ke Mars khawatir tentang kelebihan populasi di Mars, itulah sebabnya mereka mengembangkan sistem pendukung kehidupan.

Kita Adalah Para Ahli

Dalam setiap diskusi tentang risiko teknologi, kubu pro-teknologi selalu mengemukakan klaim bahwa semua kekhawatiran tentang risiko muncul dari ketidaktahuan. Misalnya, berikut adalah Oren Etzioni, CEO Allen Institute for AI dan peneliti terkemuka di bidang pembelajaran mesin dan pemahaman bahasa alami:

> Pada setiap munculnya inovasi teknologi, orang-orang selalu merasa takut. Dari para penenun yang melemparkan sepatu mereka ke mesin tenun mekanis di awal era industri hingga ketakutan akan robot pembunuh saat ini, respons kita

> didorong oleh ketidaktahuan tentang dampak teknologi baru terhadap rasa percaya diri dan mata pencaharian kita. Dan ketika kita tidak tahu, pikiran kita yang penuh ketakutan akan mengisi detailnya.

Popular Science menerbitkan sebuah artikel berjudul "Bill Gates Takut pada AI, tetapi Peneliti AI Lebih Tahu":

> Ketika Anda berbicara dengan para peneliti AI sekali lagi, para peneliti AI sejati, orang-orang yang bergulat dengan pembuatan sistem yang berfungsi, apalagi berfungsi terlalu baik mereka tidak khawatir tentang kecerdasan super yang menyelinap mendekati mereka, sekarang atau di masa depan. Bertentangan dengan cerita-cerita menakutkan yang tampaknya ingin diceritakan Musk, para peneliti AI tidak panik memasang ruang pemanggilan yang dilindungi firewall dan penghitung waktu penghancuran diri.

Analisis ini didasarkan pada sampel empat orang, yang semuanya sebenarnya mengatakan dalam wawancara mereka bahwa keamanan AI jangka panjang adalah masalah penting.

Dengan menggunakan bahasa yang sangat mirip dengan artikel *Popular Science*, David Kenny, yang saat itu menjabat sebagai wakil presiden di IBM, menulis surat kepada Kongres AS yang berisi kata-kata yang meyakinkan berikut ini:

> Ketika Anda benar-benar melakukan ilmu kecerdasan mesin, dan ketika Anda benar-benar menerapkannya di dunia nyata bisnis dan masyarakat seperti yang telah kami lakukan di IBM untuk menciptakan sistem komputasi kognitif perintis kami, Watson Anda memahami bahwa teknologi ini tidak mendukung penyebaran ketakutan yang umumnya dikaitkan dengan debat AI saat ini.

Pesan yang disampaikan sama dalam ketiga kasus tersebut: "Jangan dengarkan mereka; kami adalah para ahlinya." Sekarang, orang dapat menunjukkan bahwa ini sebenarnya adalah argumen ad hominem yang mencoba untuk membantah pesan tersebut dengan mendiskreditkan para pembawa pesan, tetapi bahkan jika diterima begitu saja, argumen tersebut tidak masuk akal. Elon Musk, Stephen Hawking, dan Bill Gates tentu sangat familiar dengan penalaran ilmiah dan teknologi, dan Musk dan Gates khususnya telah mengawasi dan berinvestasi dalam banyak proyek penelitian AI. Dan akan semakin tidak masuk akal untuk berpendapat bahwa Alan Turing, I. J. Good, Norbert Wiener, dan Marvin Minsky tidak memenuhi syarat untuk membahas AI. Terakhir, tulisan blog Scott Alexander yang disebutkan sebelumnya, yang berjudul "Peneliti AI tentang Risiko AI," mencatat bahwa "peneliti AI, termasuk beberapa pemimpin di bidang ini, telah berperan penting dalam mengangkat isu-isu tentang risiko AI dan superintelijen sejak awal." Dia mencantumkan beberapa peneliti tersebut, dan daftarnya sekarang jauh lebih panjang.

Langkah retorika standar lainnya bagi “pembela AI” adalah menggambarkan lawan mereka sebagai Luddite. Referensi Oren Etzioni tentang “penenun yang melemparkan sepatu mereka ke alat tenun mekanis” adalah ini: kaum Luddite adalah penenun pengrajin pada awal abad kesembilan belas yang memprotes pengenalan mesin untuk menggantikan tenaga kerja terampil mereka. Pada tahun 2015, Yayasan Teknologi Informasi dan Inovasi memberikan Penghargaan Luddite tahunannya kepada “para pengkhawatir yang menggembar-gemborkan kiamat kecerdasan buatan.” Ini adalah definisi "Luddite" yang aneh karena mencakup Turing, Wiener, Minsky, Musk, dan Gates, yang termasuk di antara kontributor paling terkemuka bagi kemajuan teknologi di abad kedua puluh dan kedua puluh satu.

Tuduhan Luddism merupakan kesalahpahaman tentang sifat kekhawatiran yang diangkat dan tujuan pengangkatannya. Ini seperti menuduh insinyur nuklir sebagai Luddism jika mereka menunjukkan perlunya pengendalian reaksi fisi. Seperti halnya fenomena aneh para peneliti AI yang tiba-tiba mengklaim bahwa AI tidak mungkin, saya pikir kita dapat mengaitkan episode membingungkan ini dengan tribalisme dalam membela kemajuan teknologi.

6.2 MENERIMA RISIKO TAPI MENOLAK BERTINDAK

Beberapa komentator bersedia menerima bahwa risikonya nyata, tetapi tetap mengajukan argumen untuk tidak melakukan apa pun. Argumen-argumen ini termasuk ketidakmungkinan untuk melakukan apa pun, pentingnya melakukan sesuatu yang sama sekali berbeda, dan perlunya merahasiakan risiko tersebut.

Anda tidak dapat Mengendalikan Penelitian

Jawaban umum terhadap saran bahwa AI tingkat lanjut mungkin menimbulkan risiko bagi umat manusia adalah dengan mengklaim bahwa melarang penelitian AI tidak mungkin. Perhatikan lompatan mental di sini: “Hmm, seseorang sedang membahas risiko! Mereka pasti mengusulkan larangan terhadap penelitian saya!!” Lompatan mental ini mungkin tepat dalam diskusi tentang risiko yang hanya didasarkan pada masalah gorila, dan saya cenderung setuju bahwa memecahkan masalah gorila dengan mencegah terciptanya AI supercerdas akan membutuhkan beberapa jenis batasan pada penelitian AI.

Namun, diskusi risiko baru-baru ini tidak berfokus pada masalah gorila secara umum (secara jurnalistik, pertarungan manusia vs. kecerdasan super) tetapi pada masalah Raja Midas dan variannya. Memecahkan masalah Raja Midas juga memecahkan masalah gorila bukan dengan mencegah AI supercerdas atau menemukan cara untuk mengalahkannya tetapi dengan memastikan bahwa AI tersebut tidak pernah berkonflik dengan manusia sejak awal. Diskusi tentang masalah Raja Midas umumnya menghindari usulan agar penelitian AI dibatasi; diskusi tersebut hanya menyarankan agar perhatian diberikan pada masalah pencegahan konsekuensi negatif dari sistem yang dirancang dengan buruk. Dengan cara yang sama, diskusi tentang risiko kegagalan penahanan di pembangkit nuklir harus ditafsirkan bukan sebagai upaya untuk melarang penelitian fisika nuklir tetapi sebagai saran untuk lebih memfokuskan upaya pada pemecahan masalah penahanan.

Kebetulan, ada preseden historis yang sangat menarik untuk menghentikan penelitian. Pada awal tahun 1970-an, para ahli biologi mulai khawatir bahwa metode DNA rekombinan baru penyambungan gen dari satu organisme ke organisme lain dapat menimbulkan risiko besar bagi kesehatan manusia dan ekosistem global. Dua pertemuan di Asilomar di California pada tahun 1973 dan 1975 pertama-tama mengarah pada moratorium terhadap eksperimen semacam itu dan kemudian pada pedoman biokeamanan terperinci yang sesuai dengan risiko yang ditimbulkan oleh eksperimen yang diusulkan. Beberapa kelas eksperimen, seperti yang melibatkan gen toksin, dianggap terlalu berbahaya untuk diizinkan.

Segera setelah pertemuan tahun 1975, *National Institutes of Health* (NIH), yang mendanai hampir semua penelitian medis dasar di Amerika Serikat, memulai proses pembentukan Komite Penasihat DNA Rekombinan. RAC, sebagaimana dikenal, berperan penting dalam mengembangkan pedoman NIH yang pada dasarnya menerapkan rekomendasi Asilomar. Sejak tahun 2000, pedoman tersebut mencakup larangan persetujuan pendanaan untuk protokol apa pun yang melibatkan perubahan garis keturunan manusia modifikasi genom manusia dengan cara yang dapat diwariskan kepada generasi berikutnya. Larangan ini diikuti oleh larangan hukum di lebih dari lima puluh negara.

Tujuan "meningkatkan kualitas manusia" telah menjadi salah satu impian gerakan eugenika pada akhir abad kesembilan belas dan awal abad kedua puluh. Pengembangan CRISPR-Cas9, metode yang sangat presisi untuk pengeditan genom, telah menghidupkan kembali impian ini. Sebuah pertemuan puncak internasional yang diadakan pada tahun 2015 membuka pintu bagi aplikasi di masa mendatang, menyerukan pengekangan sampai "ada konsensus masyarakat yang luas tentang kesesuaian aplikasi yang diusulkan."

Pada November 2018, ilmuwan Tiongkok He Jiankui mengumumkan bahwa ia telah mengedit genom tiga embrio manusia, setidaknya dua di antaranya telah menghasilkan kelahiran hidup. Kecaman internasional pun terjadi, dan pada saat penulisan ini, Jiankui tampaknya berada di bawah tahanan rumah. Pada Maret 2019, sebuah panel internasional yang terdiri dari para ilmuwan terkemuka secara eksplisit menyerukan moratorium formal.

Pelajaran dari perdebatan ini untuk AI bersifat beragam. Di satu sisi, hal ini menunjukkan bahwa kita dapat menahan diri untuk tidak melanjutkan bidang penelitian yang memiliki potensi besar. Konsensus internasional menentang perubahan garis keturunan hampir sepenuhnya berhasil hingga saat ini. Kekhawatiran bahwa larangan hanya akan mendorong penelitian ke bawah tanah, atau ke negara-negara tanpa regulasi, belum terwujud. Di sisi lain, perubahan garis keturunan adalah proses yang mudah diidentifikasi, kasus penggunaan spesifik dari pengetahuan umum tentang genetika yang membutuhkan peralatan khusus dan manusia sungguhan untuk melakukan eksperimen. Selain itu, hal ini termasuk dalam bidang kedokteran reproduksi yang sudah tunduk pada pengawasan dan regulasi yang ketat. Karakteristik ini tidak berlaku untuk AI tujuan umum, dan, hingga saat ini, belum ada yang menemukan bentuk regulasi yang masuk akal untuk membatasi penelitian AI.

"Whataboutery"

Saya diperkenalkan dengan istilah "whataboutery" oleh seorang penasihat politisi Inggris yang harus menghadapinya secara teratur di pertemuan publik. Tidak peduli topik

pidato yang disampaikannya, seseorang pasti akan bertanya, "Bagaimana dengan penderitaan rakyat Palestina?" Sebagai tanggapan atas setiap penyebutan risiko dari AI canggih, kemungkinan besar akan terdengar, "Bagaimana dengan manfaat AI?" Misalnya, berikut adalah Oren Etzioni:

> Prediksi malapetaka dan kesuraman seringkali gagal mempertimbangkan potensi manfaat AI dalam mencegah kesalahan medis, mengurangi kecelakaan mobil, dan banyak lagi.

Dan berikut adalah Mark Zuckerberg, CEO Facebook, dalam pertukaran yang dipicu media baru-baru ini dengan Elon Musk:

> Jika Anda menentang AI, maka Anda menentang mobil yang lebih aman yang tidak akan mengalami kecelakaan dan Anda menentang kemampuan untuk mendiagnosis orang dengan lebih baik ketika mereka sakit.

Mengesampingkan anggapan umum bahwa siapa pun yang menyebutkan risiko berarti "menentang AI," baik Zuckerberg maupun Etzioni berpendapat bahwa membicarakan risiko berarti mengabaikan potensi manfaat AI atau bahkan meniadakannya.

Ini justru terbalik, karena dua alasan. Pertama, jika tidak ada potensi manfaat AI, tidak akan ada dorongan ekonomi atau sosial untuk penelitian AI dan karenanya tidak ada bahaya untuk mencapai AI setingkat manusia. Kita tidak akan pernah membahas ini sama sekali. Kedua, jika risiko tidak berhasil dimitigasi, tidak akan ada manfaat. Potensi manfaat tenaga nuklir telah sangat berkurang karena peleburan inti parsial di Three Mile Island pada tahun 1979, reaksi yang tidak terkendali dan pelepasan bencana di Chernobyl pada tahun 1986, dan beberapa peleburan di Fukushima pada tahun 2011. Bencana-bencana tersebut sangat membatasi pertumbuhan industri nuklir. Italia meninggalkan tenaga nuklir pada tahun 1990 dan Belgia, Jerman, Spanyol, dan Swiss telah mengumumkan rencana untuk melakukan hal yang sama. Sejak tahun 1990, tingkat pengoperasian pembangkit nuklir di seluruh dunia hanya sekitar sepersepuluh dari sebelum Chernobyl.

Keheningan

Bentuk pengalihan yang paling ekstrem adalah dengan menyarankan agar kita tetap diam tentang risikonya. Misalnya, laporan AI100 yang disebutkan di atas mencakup peringatan berikut:

> Jika masyarakat mendekati teknologi ini terutama dengan rasa takut dan curiga, kesalahan langkah yang memperlambat pengembangan AI atau mendorongnya ke bawah tanah akan terjadi, menghambat pekerjaan penting untuk memastikan keamanan dan keandalan teknologi AI.

Robert Atkinson, direktur Information Technology and Innovation Foundation (yayasan yang sama yang memberikan Penghargaan Luddite), menyampaikan argumen serupa dalam debat tahun 2015. Meskipun ada pertanyaan yang valid tentang bagaimana tepatnya risiko harus dijelaskan ketika berbicara kepada media, pesan keseluruhannya jelas: “Jangan sebutkan risikonya; itu akan buruk untuk pendanaan.” Tentu saja, jika tidak ada yang menyadari risikonya, tidak akan ada pendanaan untuk penelitian tentang mitigasi risiko dan tidak ada alasan bagi siapa pun untuk mengerjakannya.

Ilmuwan kognitif ternama Steven Pinker memberikan versi yang lebih optimis dari argumen Atkinson. Menurut pandangannya, "budaya keselamatan di masyarakat maju" akan memastikan bahwa semua risiko serius dari AI akan dihilangkan; oleh karena itu, tidak tepat dan kontraproduktif untuk menarik perhatian pada risiko-risiko tersebut. Bahkan jika kita mengabaikan fakta bahwa budaya keselamatan kita yang maju telah menyebabkan Chernobyl, Fukushima, dan pemanasan global yang tak terkendali, argumen Pinker sepenuhnya meleset dari intinya.

Budaya keselamatan justru terdiri dari orang-orang yang menunjukkan kemungkinan mode kegagalan dan menemukan cara untuk memastikan hal itu tidak terjadi. (Dan dengan AI, model standar adalah mode kegagalan.) Mengatakan bahwa tidak masuk akal untuk menunjukkan mode kegagalan karena budaya keselamatan akan memperbaikinya juga sama seperti mengatakan tidak ada yang boleh memanggil ambulans ketika mereka melihat kecelakaan tabrak lari karena seseorang akan memanggil ambulans.

Dalam upaya untuk menggambarkan risiko kepada publik dan pembuat kebijakan, para peneliti AI berada dalam posisi yang kurang menguntungkan dibandingkan dengan fisikawan nuklir. Para fisikawan tidak perlu menulis buku yang menjelaskan kepada publik bahwa mengumpulkan sejumlah besar uranium yang diperkaya dapat menimbulkan risiko, karena konsekuensinya telah terbukti di Hiroshima dan Nagasaki. Tidak diperlukan banyak bujukan lebih lanjut untuk meyakinkan pemerintah dan lembaga pendanaan bahwa keselamatan itu penting dalam pengembangan energi nuklir.

6.3 SOLUSI INSTAN YANG TERLALU DISEDERHANAKAN

Dalam karya Butler, Erewhon, fokus pada masalah gorila menyebabkan dikotomi yang prematur dan keliru antara pendukung mesin dan penentang mesin. Pendukung mesin percaya bahwa risiko dominasi mesin minimal atau tidak ada; penentang mesin percaya bahwa risiko tersebut tidak dapat diatasi kecuali semua mesin dihancurkan. Perdebatan menjadi bersifat kesukuan, dan tidak ada yang mencoba untuk menyelesaikan masalah mendasar tentang mempertahankan kendali manusia atas mesin.

Pada berbagai tingkat, semua isu teknologi utama abad ke-20 tenaga nuklir, organisme hasil rekayasa genetika (GMO), dan bahan bakar fosil menyerah pada tribalisme. Pada setiap isu, ada dua sisi, pro dan anti. Dinamika dan hasil dari masing-masing isu berbeda, tetapi gejala tribalisme serupa: ketidakpercayaan dan penghinaan timbal balik, argumen irasional, dan penolakan untuk mengakui poin (yang masuk akal) yang mungkin menguntungkan kelompok lain.

Di sisi pro-teknologi, terlihat penyangkalan dan penyembunyian risiko yang dikombinasikan dengan tuduhan Luddism; di sisi anti-teknologi, terlihat keyakinan bahwa risikonya tidak dapat diatasi dan masalahnya tidak dapat dipecahkan. Seorang anggota kelompok pro-teknologi yang terlalu jujur tentang suatu masalah dipandang sebagai pengkhianat, yang sangat disayangkan karena kelompok pro-teknologi biasanya mencakup sebagian besar orang yang memenuhi syarat untuk memecahkan masalah tersebut.

Seorang anggota kelompok anti-teknologi yang membahas kemungkinan mitigasi juga dianggap sebagai pengkhianat, karena teknologi itu sendiri yang dipandang sebagai sesuatu yang jahat, bukan efek yang mungkin ditimbulkannya. Dengan cara ini, hanya suara-suara yang paling ekstrem mereka yang paling kecil kemungkinannya untuk didengarkan oleh pihak lain yang dapat berbicara untuk masing-masing kelompok.

Pada tahun 2016, saya diundang ke No. 10 Downing Street untuk bertemu dengan beberapa penasihat perdana menteri saat itu, David Cameron. Mereka khawatir bahwa perdebatan AI mulai menyerupai perdebatan GMO yang, di Eropa, telah menyebabkan apa yang dianggap oleh para penasihat sebagai peraturan yang prematur dan terlalu ketat tentang produksi dan pelabelan GMO.

Mereka ingin menghindari hal yang sama terjadi pada AI. Kekhawatiran mereka memiliki beberapa validitas: perdebatan AI berisiko menjadi bersifat kesukuan, menciptakan kubu pro-AI dan anti-AI. Ini akan merusak bidang ini karena sama sekali tidak benar bahwa mengkhawatirkan risiko yang melekat pada AI tingkat lanjut adalah sikap anti-AI. Seorang fisikawan yang khawatir tentang risiko perang nuklir atau risiko reaktor nuklir yang dirancang buruk meledak bukanlah "anti-fisika." Mengatakan bahwa AI akan cukup kuat untuk memiliki dampak global adalah pujian bagi bidang ini daripada penghinaan.

Sangat penting bagi komunitas AI untuk mengakui risiko dan berupaya untuk menguranginya. Risiko, sejauh yang kita pahami, bukanlah minimal atau tidak dapat diatasi. Kita perlu melakukan banyak pekerjaan untuk menghindarinya, termasuk membentuk kembali dan membangun kembali fondasi AI.

Tidakkah Kita Bisa Saja . . .

. . . Mematikannya?

Setelah memahami gagasan dasar risiko eksistensial, baik dalam bentuk masalah gorila atau masalah Raja Midas, banyak orang termasuk saya segera mulai mencari solusi mudah. Seringkali, hal pertama yang terlintas di pikiran adalah mematikan mesin. Misalnya, Alan Turing sendiri, seperti yang dikutip sebelumnya, berspekulasi bahwa kita mungkin "menjaga mesin dalam posisi tunduk, misalnya dengan mematikan daya pada saat-saat strategis." Ini tidak akan berhasil, karena alasan sederhana bahwa entitas supercerdas sudah memikirkan kemungkinan itu dan mengambil langkah-langkah untuk mencegahnya. Dan itu akan dilakukan bukan karena ingin tetap hidup tetapi karena mengejar tujuan apa pun yang kita berikan dan tahu bahwa itu akan gagal jika dimatikan.

Ada beberapa sistem yang sedang dipertimbangkan yang benar-benar tidak dapat dimatikan tanpa merusak banyak infrastruktur peradaban kita. Ini adalah sistem yang diimplementasikan sebagai apa yang disebut kontrak pintar di blockchain. Blockchain adalah

bentuk komputasi dan pencatatan yang sangat terdistribusi berdasarkan enkripsi; ia dirancang secara khusus sehingga tidak ada data yang dapat dihapus dan tidak ada kontrak pintar yang dapat diganggu tanpa pada dasarnya mengambil kendali atas sejumlah besar mesin dan membatalkan rantai, yang pada gilirannya dapat menghancurkan sebagian besar Internet dan/atau sistem keuangan. Masih diperdebatkan apakah ketahanan yang luar biasa ini merupakan fitur atau bug. Ini tentu saja merupakan alat yang dapat digunakan oleh sistem AI supercerdas untuk melindungi dirinya sendiri.

...Masukkan ke dalam Kotak?

Jika Anda tidak dapat mematikan sistem AI, dapatkah Anda menyegel mesin-mesin tersebut di dalam semacam firewall, mengekstrak pekerjaan menjawab pertanyaan yang berguna dari mereka tetapi tidak pernah membiarkan mereka memengaruhi dunia nyata secara langsung? Inilah ide di balik Oracle AI, yang telah dibahas panjang lebar di komunitas keamanan AI. Sistem Oracle AI dapat memiliki kecerdasan yang sewenang-wenang, tetapi hanya dapat menjawab ya atau tidak (atau memberikan probabilitas yang sesuai) untuk setiap pertanyaan. Ia dapat mengakses semua informasi yang dimiliki umat manusia melalui koneksi baca-saja artinya, ia tidak memiliki akses langsung ke Internet.

Tentu saja, ini berarti kita harus mengorbankan robot supercerdas, asisten, dan berbagai jenis sistem AI lainnya, tetapi AI Oracle yang dapat dipercaya tetap akan memiliki nilai ekonomi yang sangat besar karena kita dapat mengajukan pertanyaan yang jawabannya penting bagi kita, seperti apakah penyakit Alzheimer disebabkan oleh organisme menular atau apakah melarang senjata otonom adalah ide yang baik. Dengan demikian, AI Oracle tentu merupakan kemungkinan yang menarik.

Sayangnya, ada beberapa kesulitan serius. Pertama, sistem AI Oracle setidaknya akan sama tekunnya dalam memahami fisika dan asal-usul dunianya sumber daya komputasi, cara kerjanya, dan entitas misterius yang menghasilkan penyimpanan informasinya dan sekarang mengajukan pertanyaan seperti halnya kita dalam memahami dunia kita. Kedua, jika tujuan sistem AI Oracle adalah untuk memberikan jawaban yang akurat atas pertanyaan dalam waktu yang wajar, ia akan memiliki insentif untuk keluar dari sangkarnya untuk memperoleh lebih banyak sumber daya komputasi dan untuk mengendalikan para penanya sehingga mereka hanya mengajukan pertanyaan sederhana. Dan, akhirnya, kita belum menemukan firewall yang aman terhadap manusia biasa, apalagi mesin supercerdas.

Saya pikir mungkin ada solusi untuk beberapa masalah ini, terutama jika kita membatasi sistem AI Oracle agar menjadi kalkulator logis atau Bayesian yang terbukti valid. Artinya, kita dapat bersikeras bahwa algoritma hanya dapat menghasilkan kesimpulan yang dijamin oleh informasi yang diberikan, dan kita dapat memeriksa secara matematis bahwa algoritma tersebut memenuhi kondisi ini. Ini masih menyisakan masalah pengendalian proses yang memutuskan komputasi logis atau Bayesian mana yang harus dilakukan, untuk mencapai kesimpulan terkuat secepat mungkin. Karena proses ini memiliki insentif untuk bernalar dengan cepat, ia memiliki insentif untuk memperoleh sumber daya komputasi dan tentu saja untuk mempertahankan keberadaannya sendiri.

Pada tahun 2018, Pusat AI yang Kompatibel dengan Manusia di Berkeley mengadakan lokakarya di mana kami mengajukan pertanyaan, “Apa yang akan Anda lakukan jika Anda yakin bahwa AI supercerdas akan tercapai dalam satu dekade?” Jawaban saya adalah sebagai berikut: bujuk para pengembang untuk menunda pembangunan agen cerdas serbaguna agen yang dapat memilih tindakannya sendiri di dunia nyata dan sebagai gantinya bangun Oracle AI.

Sementara itu, kami akan berupaya memecahkan masalah untuk membuat sistem Oracle AI terbukti aman sejauh mungkin. Alasan strategi ini mungkin berhasil ada dua: pertama, sistem Oracle AI supercerdas masih akan bernilai triliunan rupiah, sehingga para pengembang mungkin bersedia menerima batasan ini; dan kedua, mengendalikan sistem Oracle AI hampir pasti lebih mudah daripada mengendalikan agen cerdas serbaguna, sehingga kami memiliki peluang lebih baik untuk memecahkan masalah dalam dekade ini.

... Bekerja dalam Tim Manusia-Mesin?

Sebuah ungkapan umum di dunia korporat adalah bahwa AI bukanlah ancaman bagi lapangan kerja atau bagi umat manusia karena kita akan memiliki tim kolaboratif manusia AI. Misalnya, surat David Kenny kepada Kongres, yang dikutip sebelumnya dalam bab ini, menyatakan bahwa "sistem kecerdasan buatan bernilai tinggi dirancang khusus untuk meningkatkan kecerdasan manusia, bukan menggantikan pekerja."

Meskipun seorang sinis mungkin berpendapat bahwa ini hanyalah taktik hubungan masyarakat untuk mempermanis proses penghapusan karyawan manusia dari klien perusahaan, saya pikir ini memang memajukan proses tersebut beberapa inci. Tim kolaboratif manusia AI memang merupakan tujuan yang diinginkan. Jelas, sebuah tim akan gagal jika tujuan anggota tim tidak selaras, jadi penekanan pada tim manusia AI menyoroti kebutuhan untuk menyelesaikan masalah inti dari keselarasan nilai. Tentu saja, menyoroti masalah tidak sama dengan menyelesaikannya.

... Bergabung dengan Mesin?

Kolaborasi manusia-mesin, jika dibawa ke titik ekstrem, menjadi penggabungan manusia-mesin di mana perangkat keras elektronik terhubung langsung ke otak dan membentuk bagian dari satu entitas sadar yang diperluas. Pakar futuris Ray Kurzweil menggambarkan kemungkinan ini sebagai berikut:

> Kita akan langsung bergabung dengannya, kita akan menjadi AI. Menjelang akhir tahun 2030-an atau 2040-an, pemikiran kita akan didominasi oleh hal-hal non-biologis dan bagian non-biologis tersebut pada akhirnya akan sangat cerdas dan memiliki kapasitas yang sangat besar sehingga mampu memodelkan, mensimulasikan, dan memahami sepenuhnya bagian biologis.

Kurzweil memandang perkembangan ini secara positif. Di sisi lain, Elon Musk memandang penggabungan manusia-mesin terutama sebagai strategi defensif:

> Jika kita mencapai simbiosis yang erat, AI tidak akan menjadi "yang lain" ia akan menjadi Anda dan [ia akan memiliki] hubungan dengan korteks Anda yang analog

> dengan hubungan korteks Anda dengan sistem limbik Anda. Kita akan memiliki pilihan untuk tertinggal dan menjadi tidak berguna atau seperti hewan peliharaan Anda tahu, seperti kucing rumahan atau semacamnya atau akhirnya menemukan cara untuk bersimbiosis dan bergabung dengan AI.

Korporasi Neuralink milik Musk sedang mengerjakan perangkat yang disebut "neural lace" setelah teknologi yang dijelaskan dalam novel Culture karya Iain Banks. Tujuannya adalah untuk menciptakan koneksi yang kuat dan permanen antara korteks manusia dan sistem serta jaringan komputasi eksternal.

Ada dua hambatan teknis utama: pertama, kesulitan menghubungkan perangkat elektronik ke jaringan otak, memasoknya dengan daya, dan menghubungkannya ke dunia luar; Kedua, fakta bahwa kita hampir tidak memahami implementasi saraf dari tingkat kognisi yang lebih tinggi di otak, sehingga kita tidak tahu di mana harus menghubungkan perangkat tersebut dan pemrosesan apa yang harus dilakukannya.

Saya tidak sepenuhnya yakin bahwa hambatan dalam paragraf sebelumnya tidak dapat diatasi. Pertama, teknologi seperti debu saraf dengan cepat mengurangi ukuran dan kebutuhan daya perangkat elektronik yang dapat dihubungkan ke neuron dan menyediakan penginderaan, stimulasi, dan komunikasi transkranial. (Teknologi tersebut pada tahun 2018 telah mencapai ukuran sekitar satu milimeter kubik, jadi butiran saraf mungkin merupakan istilah yang lebih akurat.) Kedua, otak itu sendiri memiliki kekuatan adaptasi yang luar biasa.

Dahulu, misalnya, diperkirakan bahwa kita harus memahami kode yang digunakan otak untuk mengontrol otot lengan sebelum kita dapat menghubungkan otak ke lengan robot dengan sukses, dan bahwa kita harus memahami cara koklea menganalisis suara sebelum kita dapat membangun penggantinya. Ternyata, sebaliknya, otak melakukan sebagian besar pekerjaan untuk kita. Ia dengan cepat mempelajari cara membuat lengan robot melakukan apa yang diinginkan pemiliknya, dan cara memetakan output implan koklea ke suara yang dapat dipahami. Sangat mungkin kita akan menemukan cara untuk menyediakan otak dengan memori tambahan, dengan saluran komunikasi ke komputer, dan mungkin bahkan dengan saluran komunikasi ke otak lain semuanya tanpa pernah benar-benar memahami bagaimana semua itu bekerja.

Terlepas dari kelayakan teknologi dari ide-ide ini, kita harus bertanya apakah arah ini mewakili masa depan terbaik bagi umat manusia. Jika manusia membutuhkan operasi otak hanya untuk bertahan hidup dari ancaman yang ditimbulkan oleh teknologi mereka sendiri, mungkin kita telah membuat kesalahan di suatu tempat.

... Menghindari Memasukkan Tujuan Manusia?

Sebuah argumen umum menyatakan bahwa perilaku AI yang bermasalah muncul dari memasukkan jenis tujuan tertentu; jika ini dihilangkan, semuanya akan baik-baik saja. Jadi, misalnya, Yann LeCun, pelopor pembelajaran mendalam dan direktur penelitian AI di Facebook, sering mengutip gagasan ini ketika meremehkan risiko dari AI:

> Tidak ada alasan bagi AI untuk memiliki naluri mempertahankan diri, kecemburuan, dll. AI tidak akan memiliki "emosi" destruktif ini kecuali kita membangun emosi ini ke dalamnya. Saya tidak melihat mengapa kita ingin melakukan itu.

Dalam konteks yang serupa, Steven Pinker memberikan analisis berbasis gender:

> Distopia AI memproyeksikan psikologi alfa-laki-laki yang sempit ke dalam konsep kecerdasan. Mereka berasumsi bahwa robot yang sangat cerdas akan mengembangkan tujuan seperti menggulingkan tuan mereka atau mengambil alih dunia. Sangatlah ironis bahwa banyak nabi teknologi kita tidak mempertimbangkan kemungkinan bahwa kecerdasan buatan akan berkembang secara alami mengikuti garis keturunan perempuan: sepenuhnya mampu memecahkan masalah, tetapi tanpa keinginan untuk memusnahkan orang-orang yang tidak bersalah atau mendominasi peradaban.

Seperti yang telah kita lihat dalam pembahasan tujuan instrumental, tidak masalah apakah kita memasukkan "emosi" atau "keinginan" seperti mempertahankan diri, memperoleh sumber daya, menemukan pengetahuan, atau, dalam kasus ekstrem, menguasai dunia. Mesin akan tetap memiliki emosi tersebut, sebagai sub-tujuan dari tujuan apa pun yang kita masukkan dan terlepas dari jenis kelaminnya. Bagi mesin, kematian bukanlah hal yang buruk. Namun demikian, kematian harus dihindari, karena sulit untuk mengambil kopi jika Anda sudah mati.

Solusi yang lebih ekstrem adalah menghindari memasukkan tujuan ke dalam mesin sama sekali. Voila, masalah terpecahkan. Sayangnya, tidak sesederhana itu. Tanpa tujuan, tidak ada kecerdasan: setiap tindakan sama baiknya dengan tindakan lainnya, dan mesin tersebut sama saja dengan generator angka acak. Tanpa tujuan, juga tidak ada alasan bagi mesin untuk lebih menyukai surga manusia daripada planet yang berubah menjadi lautan penjepit kertas (skenario yang dijelaskan panjang lebar oleh Nick Bostrom). Memang, hasil terakhir mungkin utopis bagi bakteri pemakan besi, *Thiobacillus ferrooxidans*. Tanpa adanya anggapan bahwa preferensi manusia itu penting, siapa yang dapat mengatakan bahwa bakteri itu salah?

Varian umum dari gagasan "hindari memasukkan tujuan" adalah gagasan bahwa sistem yang cukup cerdas akan, sebagai konsekuensi dari kecerdasannya, mengembangkan tujuan yang "benar" dengan sendirinya. Seringkali, pendukung gagasan ini mengacu pada teori bahwa orang-orang dengan kecerdasan yang lebih tinggi cenderung memiliki tujuan yang lebih altruistik dan luhur pandangan yang mungkin terkait dengan konsepsi diri para pendukungnya.

Gagasan bahwa dimungkinkan untuk memahami tujuan di dunia dibahas panjang lebar oleh filsuf terkenal abad ke-18, David Hume, dalam *A Treatise of Human Nature*. Ia menyebutnya masalah is-ought dan menyimpulkan bahwa adalah suatu kesalahan untuk berpikir bahwa imperatif moral dapat disimpulkan dari fakta-fakta alam. Untuk melihat alasannya, pertimbangkan, misalnya, desain papan catur dan bidak catur. Kita tidak dapat

melihat tujuan skakmat di dalamnya, karena papan catur dan bidak yang sama dapat digunakan untuk catur bunuh diri atau bahkan banyak permainan lain yang belum ditemukan.

Nick Bostrom, dalam *Superintelligence*, menyajikan ide dasar yang sama dalam bentuk yang berbeda, yang ia sebut *tesis ortogonalitas*:

> Kecerdasan dan tujuan akhir bersifat ortogonal: pada prinsipnya, tingkat kecerdasan apa pun dapat dikombinasikan dengan tujuan akhir apa pun.

Di sini, ortogonal berarti "pada sudut siku-siku" dalam arti bahwa tingkat kecerdasan adalah satu sumbu yang mendefinisikan sistem cerdas dan tujuannya adalah sumbu lain, dan kita dapat memvariasikannya secara independen. Misalnya, mobil otonom dapat diberi alamat tertentu sebagai tujuannya; menjadikan mobil pengemudi yang lebih baik tidak berarti bahwa mobil tersebut akan mulai menolak untuk pergi ke alamat yang habis dibagi tujuh belas.

Dengan cara yang sama, mudah untuk membayangkan bahwa sistem cerdas tujuan umum dapat diberi tujuan apa pun untuk dikejar termasuk memaksimalkan jumlah penjepit kertas atau jumlah digit pi yang diketahui. Beginilah cara kerja sistem pembelajaran penguatan dan jenis pengoptimal imbalan lainnya: algoritma tersebut sepenuhnya umum dan menerima sinyal imbalan apa pun. Bagi para insinyur dan ilmuwan komputer yang beroperasi dalam model standar, tesis ortogonalitas hanyalah sebuah kepastian.

Gagasan bahwa sistem cerdas dapat dengan mudah mengamati dunia untuk memperoleh tujuan yang harus dikejar menunjukkan bahwa sistem yang cukup cerdas secara alami akan meninggalkan tujuan awalnya demi tujuan yang "benar". Sulit untuk melihat mengapa agen rasional akan melakukan ini. Lebih jauh lagi, hal itu mengasumsikan bahwa ada tujuan yang "benar" di dunia ini; tujuan tersebut haruslah tujuan yang disetujui oleh bakteri pemakan besi, manusia, dan semua spesies lainnya, yang sulit dibayangkan.

Kritik paling eksplisit terhadap tesis ortogonalitas Bostrom datang dari ahli robotika terkenal Rodney Brooks, yang menyatakan bahwa mustahil bagi sebuah program untuk "cukup pintar sehingga mampu menciptakan cara untuk merusak masyarakat manusia demi mencapai tujuan yang ditetapkan oleh manusia, tanpa memahami cara-cara di mana program tersebut menyebabkan masalah bagi manusia yang sama." Sayangnya, bukan hanya mungkin bagi sebuah program untuk berperilaku seperti ini; bahkan, hal itu tidak dapat dihindari, mengingat cara Brooks mendefinisikan masalah tersebut.

Brooks berpendapat bahwa rencana optimal untuk "mencapai tujuan yang ditetapkan oleh manusia" menyebabkan masalah bagi manusia. Oleh karena itu, masalah-masalah tersebut mencerminkan hal-hal yang berharga bagi manusia yang diabaikan dari tujuan yang ditetapkan oleh manusia. Rencana optimal yang dijalankan oleh mesin mungkin saja menyebabkan masalah bagi manusia, dan mesin mungkin menyadari hal ini. Tetapi, menurut definisi, mesin tidak akan mengenali masalah-masalah tersebut sebagai masalah. Itu bukan urusannya.

Steven Pinker tampaknya setuju dengan tesis ortogonalitas Bostrom, menulis bahwa “kecerdasan adalah kemampuan untuk menggunakan cara-cara baru untuk mencapai suatu

tujuan; tujuan-tujuan tersebut tidak terkait dengan kecerdasan itu sendiri." Di sisi lain, ia menganggap tidak masuk akal bahwa "AI akan begitu brilian sehingga dapat menemukan cara untuk mengubah elemen dan menyusun ulang otak, namun begitu bodoh sehingga akan menimbulkan malapetaka berdasarkan kesalahan-kesalahan dasar akibat kesalahpahaman." Ia melanjutkan, "Kemampuan untuk memilih tindakan yang paling memuaskan tujuan-tujuan yang saling bertentangan bukanlah tambahan yang mungkin lupa dipasang dan diuji oleh para insinyur; itu adalah kecerdasan.

Begitu pula kemampuan untuk menafsirkan niat pengguna bahasa dalam konteksnya." Tentu saja, "memenuhi tujuan-tujuan yang saling bertentangan" bukanlah masalahnya itu adalah sesuatu yang telah dibangun ke dalam model standar sejak awal teori pengambilan keputusan. Masalahnya adalah bahwa tujuan-tujuan yang saling bertentangan yang disadari oleh mesin tidak mencakup keseluruhan perhatian manusia; Selain itu, dalam model standar, tidak ada yang mengatakan bahwa mesin harus peduli pada tujuan yang tidak diperintahkan untuk dipedulikan.

Namun, ada beberapa petunjuk yang berguna dalam apa yang dikatakan Brooks dan Pinker. Tampaknya bodoh bagi kita jika mesin, misalnya, mengubah warna langit sebagai efek samping dari mengejar tujuan lain, sementara mengabaikan tanda-tanda ketidakpuasan manusia yang jelas sebagai akibatnya. Tampaknya bodoh bagi kita karena kita peka terhadap ketidaksenangan manusia dan (biasanya) kita termotivasi untuk menghindari menyebabkannya bahkan jika sebelumnya kita tidak menyadari bahwa manusia yang dimaksud peduli pada warna langit. Artinya, kita manusia (1) peduli pada preferensi manusia lain dan (2) tahu bahwa kita tidak tahu apa semua preferensi tersebut. Pada bab selanjutnya, saya berpendapat bahwa karakteristik ini, ketika dibangun ke dalam mesin, dapat memberikan awal dari solusi untuk masalah Raja Midas.

6.4 DILEMA TUJUAN DAN JALAN TENGAH DI BALIK DEBAT AI

Bab ini telah memberikan gambaran sekilas tentang debat yang sedang berlangsung di komunitas intelektual yang luas, debat antara mereka yang menunjukkan risiko AI dan mereka yang skeptis terhadap risiko tersebut. Debat ini telah dilakukan dalam buku, blog, makalah akademis, diskusi panel, wawancara, tweet, dan artikel surat kabar. Terlepas dari upaya gigih mereka, para "skeptis" mereka yang berpendapat bahwa risiko dari AI dapat diabaikan telah gagal menjelaskan mengapa sistem AI supercerdas akan tetap berada di bawah kendali manusia; dan mereka bahkan belum mencoba menjelaskan mengapa sistem AI supercerdas tidak akan pernah dikembangkan.

Banyak skeptis akan mengakui, jika didesak, bahwa ada masalah nyata, meskipun itu tidak akan segera terjadi. Scott Alexander, dalam blog *Slate Star Codex-nya*, merangkumnya dengan brilian:

> Posisi "skeptis" tampaknya adalah bahwa, meskipun kita mungkin harus meminta beberapa orang cerdas untuk mulai mengerjakan aspek-aspek awal dari masalah ini, kita tidak boleh panik atau mulai mencoba melarang penelitian AI.

> Sementara itu, para "pengikut" bersikeras bahwa meskipun kita tidak boleh panik atau mulai mencoba melarang penelitian AI, kita mungkin perlu meminta beberapa orang cerdas untuk mulai mengerjakan aspek-aspek awal dari masalah ini.

Meskipun saya akan senang jika para skeptis mengajukan keberatan yang tak terbantahkan, mungkin dalam bentuk solusi sederhana dan anti-gagal (dan anti-kejahatan) untuk masalah kontrol AI, saya pikir kemungkinan besar hal ini tidak akan terjadi, sama seperti kita tidak akan menemukan solusi sederhana dan anti-gagal untuk keamanan siber atau cara sederhana dan anti-gagal untuk menghasilkan energi nuklir dengan risiko nol. Daripada terus terjebak dalam saling ejek dan penggalian berulang argumen yang telah didiskreditkan, tampaknya lebih baik, seperti yang dikatakan Alexander, untuk mulai mengerjakan beberapa aspek awal dari masalah ini.

Debat ini telah menyoroti dilema yang kita hadapi: jika kita membangun mesin untuk mengoptimalkan tujuan, tujuan yang kita masukkan ke dalam mesin harus sesuai dengan apa yang kita inginkan, tetapi kita tidak tahu bagaimana mendefinisikan tujuan manusia secara lengkap dan benar. Untungnya, ada jalan tengah.

BAB 7
AI: PENDEKATAN YANG BERBEDA

Setelah argumen para skeptis dibantah dan semua "tapi tapi tapi" telah dijawab, pertanyaan selanjutnya biasanya, "Oke, saya akui ada masalah, tetapi tidak ada solusinya, bukan?" Ya, ada solusinya. Mari kita ingatkan diri kita tentang tugas yang ada di hadapan kita: merancang mesin dengan tingkat kecerdasan yang tinggi sehingga mereka dapat membantu kita dengan masalah-masalah sulit sambil memastikan bahwa mesin-mesin tersebut tidak pernah berperilaku dengan cara yang membuat kita sangat tidak bahagia.

Untungnya, tugasnya bukanlah sebagai berikut: diberikan mesin yang memiliki tingkat kecerdasan tinggi, cari tahu cara mengendalikannya. Jika itu tugasnya, kita akan tamat. Mesin yang dipandang sebagai kotak hitam, sebuah fait accompli, sama saja seperti datang dari luar angkasa. Dan peluang kita untuk mengendalikan entitas supercerdas dari luar angkasa kira-kira nol. Argumen serupa berlaku untuk metode pembuatan sistem AI yang menjamin kita tidak akan memahami cara kerjanya; Metode-metode ini mencakup emulasi seluruh otak menciptakan salinan elektronik otak manusia yang ditingkatkan serta metode yang didasarkan pada simulasi evolusi program. Saya tidak akan mengatakan lebih banyak tentang proposal ini karena jelas sekali itu adalah ide yang buruk.

Jadi, bagaimana bidang AI mendekati bagian tugas "merancang mesin dengan tingkat kecerdasan tinggi" di masa lalu? Seperti banyak bidang lainnya, AI telah mengadopsi model standar: kita membangun mesin pengoptimalan, kita memasukkan tujuan ke dalamnya, dan mesin itu pun bekerja. Itu berhasil dengan baik ketika mesin-mesin itu bodoh dan memiliki ruang lingkup tindakan yang terbatas; jika Anda memasukkan tujuan yang salah, Anda memiliki peluang bagus untuk dapat mematikan mesin, memperbaiki masalah, dan mencoba lagi.

Namun, seiring mesin yang dirancang sesuai dengan model standar menjadi lebih cerdas, dan seiring ruang lingkup tindakannya menjadi lebih global, pendekatan tersebut menjadi tidak dapat dipertahankan. Mesin-mesin tersebut akan mengejar tujuannya, tidak peduli seberapa salahnya; mereka akan menolak upaya untuk mematikannya; dan mereka akan memperoleh semua sumber daya yang berkontribusi untuk mencapai tujuan tersebut. Memang, perilaku optimal bagi mesin mungkin termasuk menipu manusia agar berpikir bahwa mereka telah memberikan tujuan yang masuk akal kepada mesin, untuk mendapatkan cukup waktu guna mencapai tujuan sebenarnya yang diberikan kepadanya. Ini bukanlah perilaku "menyimpang" atau "jahat" yang membutuhkan kesadaran dan kehendak bebas; itu hanyalah bagian dari rencana optimal untuk mencapai tujuan tersebut.

Dalam Bab 1, saya memperkenalkan gagasan tentang mesin yang bermanfaat yaitu, mesin yang tindakannya diharapkan dapat mencapai tujuan kita, bukan tujuan mereka sendiri. Tujuan saya dalam bab ini adalah untuk menjelaskan secara sederhana bagaimana hal ini dapat dilakukan, terlepas dari kelemahan yang tampak bahwa mesin tidak mengetahui apa tujuan kita. Pendekatan yang dihasilkan pada akhirnya akan menghasilkan mesin yang tidak menimbulkan ancaman bagi kita, tidak peduli seberapa cerdasnya mesin tersebut.

7.1 PRINSIP-PRINSIP UNTUK MESIN YANG BERMANFAAT

Saya merasa bermanfaat untuk meringkas pendekatan ini dalam bentuk tiga prinsip. Saat membaca prinsip-prinsip ini, ingatlah bahwa prinsip-prinsip ini terutama dimaksudkan sebagai panduan bagi para peneliti dan pengembang AI dalam memikirkan cara menciptakan sistem AI yang bermanfaat; prinsip-prinsip ini tidak dimaksudkan sebagai hukum eksplisit yang harus diikuti oleh sistem AI:

1. Satu-satunya tujuan mesin adalah untuk memaksimalkan realisasi preferensi manusia.
2. Mesin pada awalnya tidak yakin tentang apa preferensi tersebut.
3. Sumber informasi utama tentang preferensi manusia adalah perilaku manusia.

Sebelum membahas penjelasan yang lebih rinci, penting untuk mengingat cakupan luas dari apa yang saya maksud dengan preferensi dalam prinsip-prinsip ini. Berikut pengingat tentang apa yang saya tulis di Bab 2: jika Anda entah bagaimana dapat menonton dua film, masing-masing menggambarkan secara detail dan luas kehidupan masa depan yang mungkin Anda jalani, sehingga masing-masing merupakan pengalaman virtual, Anda dapat mengatakan mana yang Anda sukai, atau menyatakan ketidakpedulian. Dengan demikian, preferensi di sini mencakup semuanya; Mereka mencakup semua hal yang mungkin Anda pedulikan, secara sewenang-wenang jauh ke masa depan. Dan itu milik Anda: mesin tersebut tidak berupaya untuk mengidentifikasi atau mengadopsi satu set preferensi ideal, tetapi untuk memahami dan memenuhi (sejauh mungkin) preferensi setiap orang.

Prinsip Pertama: Mesin yang Sepenuhnya Altruistik

Prinsip pertama, bahwa satu-satunya tujuan mesin adalah untuk memaksimalkan realisasi preferensi manusia, merupakan inti dari gagasan mesin yang bermanfaat. Secara khusus, mesin tersebut akan bermanfaat bagi manusia, bukan bagi, misalnya, kecoa. Tidak ada cara untuk menghindari gagasan manfaat yang spesifik bagi penerima ini.

Prinsip ini berarti bahwa mesin tersebut sepenuhnya altruistik yaitu, mesin tersebut sama sekali tidak memberikan nilai intrinsik pada kesejahteraannya sendiri atau bahkan keberadaannya sendiri. Mesin tersebut mungkin melindungi dirinya sendiri agar dapat terus melakukan hal-hal yang bermanfaat bagi manusia, atau karena pemiliknya akan tidak senang jika harus membayar biaya perbaikan, atau karena pemandangan robot yang kotor atau rusak mungkin sedikit mengganggu bagi orang yang lewat, tetapi bukan karena mesin tersebut ingin hidup. Menambahkan preferensi untuk mempertahankan diri akan menciptakan insentif tambahan dalam robot yang tidak sepenuhnya selaras dengan kesejahteraan manusia.

Rumusan prinsip pertama memunculkan dua pertanyaan yang sangat penting. Masing-masing layak dibahas dalam satu rak buku tersendiri, dan sebenarnya banyak buku telah ditulis tentang pertanyaan-pertanyaan ini. Pertanyaan pertama adalah apakah manusia benar-benar memiliki preferensi dalam arti yang bermakna atau stabil. Sebenarnya, gagasan tentang "preferensi" adalah idealisasi yang gagal sesuai dengan kenyataan dalam beberapa hal. Misalnya, kita tidak dilahirkan dengan preferensi yang kita miliki sebagai orang dewasa, jadi preferensi tersebut pasti berubah seiring waktu. Untuk saat ini, saya akan berasumsi bahwa

idealisasi tersebut masuk akal. Nanti, saya akan meneliti apa yang terjadi ketika kita melepaskan idealisasi tersebut.

Pertanyaan kedua adalah pokok bahasan ilmu sosial: mengingat bahwa biasanya tidak mungkin untuk memastikan bahwa setiap orang mendapatkan hasil yang paling mereka sukai kita semua tidak bisa menjadi Kaisar Alam Semesta bagaimana mesin harus menyeimbangkan preferensi banyak manusia? Sekali lagi, untuk saat ini dan saya berjanji untuk kembali ke pertanyaan ini di bab berikutnya tampaknya masuk akal untuk mengadopsi pendekatan sederhana yaitu memperlakukan semua orang secara setara. Ini mengingatkan pada akar utilitarianisme abad ke-18 dalam frasa "kebahagiaan terbesar untuk jumlah terbesar," dan ada banyak peringatan dan elaborasi yang diperlukan agar ini berhasil dalam praktik. Mungkin yang terpenting dari semua ini adalah masalah kemungkinan jumlah orang yang belum lahir sangat banyak, dan bagaimana preferensi mereka harus diperhitungkan.

Masalah manusia di masa depan memunculkan pertanyaan lain yang terkait: Bagaimana kita memperhitungkan preferensi entitas non-manusia? Artinya, haruskah prinsip pertama mencakup preferensi hewan? (Dan mungkin juga tumbuhan?) Ini adalah pertanyaan yang layak diperdebatkan, tetapi hasilnya tampaknya tidak akan berdampak besar pada arah perkembangan AI ke depan. Sebagai informasi tambahan, preferensi manusia dapat dan memang mencakup hal-hal yang berkaitan dengan kesejahteraan hewan, serta aspek kesejahteraan manusia yang mendapat manfaat langsung dari keberadaan hewan. Mengatakan bahwa mesin harus memperhatikan preferensi hewan selain hal ini berarti bahwa manusia harus membangun mesin yang lebih peduli pada hewan daripada manusia, yang merupakan posisi yang sulit dipertahankan. Posisi yang lebih masuk akal adalah bahwa kecenderungan kita untuk terlibat dalam pengambilan keputusan yang picik yang bertentangan dengan kepentingan kita sendiri seringkali menyebabkan konsekuensi negatif bagi lingkungan dan penghuni hewannya.

Mesin yang membuat keputusan yang kurang picik akan membantu manusia mengadopsi kebijakan yang lebih ramah lingkungan. Dan jika di masa depan kita memberikan bobot yang jauh lebih besar pada kesejahteraan hewan daripada yang kita lakukan saat ini yang mungkin berarti mengorbankan sebagian kesejahteraan intrinsik kita sendiri maka mesin akan beradaptasi sesuai dengan itu.

Prinsip Kedua: Mesin yang Rendah Hati

Prinsip kedua, bahwa mesin pada awalnya tidak yakin tentang apa preferensi manusia, adalah kunci untuk menciptakan mesin yang bermanfaat. Mesin yang menganggap dirinya mengetahui tujuan sebenarnya dengan sempurna akan mengejarnya dengan sepenuh hati. Ia tidak akan pernah bertanya apakah suatu tindakan itu baik-baik saja, karena ia sudah tahu bahwa itu adalah solusi optimal untuk tujuan tersebut. Ia akan mengabaikan manusia yang berteriak-teriak, “Hentikan, kalian akan menghancurkan dunia!” karena itu hanyalah kata-kata. Menganggap pengetahuan sempurna tentang tujuan tersebut memisahkan mesin dari manusia: apa yang dilakukan manusia tidak lagi penting, karena mesin mengetahui tujuannya dan mengejarnya.

Di sisi lain, mesin yang tidak yakin tentang tujuan sebenarnya akan menunjukkan semacam kerendahan hati: misalnya, ia akan tunduk kepada manusia dan membiarkan dirinya dimatikan. Ia beralasan bahwa manusia hanya akan mematikannya jika melakukan sesuatu yang salah yaitu, melakukan sesuatu yang bertentangan dengan preferensi manusia. Berdasarkan prinsip pertama, ia ingin menghindari hal itu, tetapi, berdasarkan prinsip kedua, ia tahu bahwa itu mungkin karena ia tidak tahu persis apa yang dimaksud dengan "salah". Jadi, jika manusia mematikan mesin, maka mesin tersebut menghindari melakukan hal yang salah, dan itulah yang diinginkannya. Dengan kata lain, mesin memiliki insentif positif untuk membiarkan dirinya dimatikan. Ia tetap terhubung dengan manusia, yang merupakan sumber informasi potensial yang akan memungkinkannya untuk menghindari kesalahan dan melakukan pekerjaan yang lebih baik.

Ketidakpastian telah menjadi perhatian utama dalam AI sejak tahun 1980-an; memang frasa "AI modern" sering merujuk pada revolusi yang terjadi ketika ketidakpastian akhirnya diakui sebagai masalah yang ada di mana-mana dalam pengambilan keputusan di dunia nyata. Namun, ketidakpastian dalam tujuan sistem AI tersebut diabaikan begitu saja.

Dalam semua pekerjaan tentang maksimisasi utilitas, pencapaian tujuan, minimisasi biaya, maksimisasi imbalan, dan minimisasi kerugian, diasumsikan bahwa fungsi utilitas, tujuan, fungsi biaya, fungsi imbalan, dan fungsi kerugian diketahui dengan sempurna. Bagaimana mungkin ini terjadi? Bagaimana mungkin komunitas AI (dan komunitas teori kontrol, riset operasi, dan statistik) memiliki titik buta yang begitu besar begitu lama, bahkan ketika merangkul ketidakpastian dalam semua aspek pengambilan keputusan lainnya?

Kita bisa saja membuat beberapa alasan teknis yang agak rumit, tetapi saya menduga kebenarannya adalah, dengan beberapa pengecualian yang terhormat, para peneliti AI hanya menerima model standar yang memetakan gagasan kita tentang kecerdasan manusia ke kecerdasan mesin: manusia memiliki tujuan dan mengejarnya, jadi mesin seharusnya memiliki tujuan dan mengejarnya. Mereka, atau lebih tepatnya kita, tidak pernah benar-benar memeriksa asumsi mendasar ini. Asumsi ini tertanam dalam semua pendekatan yang ada untuk membangun sistem cerdas.

Prinsip Ketiga: Belajar Memprediksi Preferensi Manusia

Prinsip ketiga, bahwa sumber informasi utama tentang preferensi manusia adalah perilaku manusia, memiliki dua tujuan. Tujuan pertama adalah untuk memberikan landasan yang pasti untuk istilah preferensi manusia. Berdasarkan asumsi, preferensi manusia tidak ada di dalam mesin dan mesin tidak dapat mengamatinya secara langsung, tetapi harus tetap ada hubungan yang pasti antara mesin dan preferensi manusia. Prinsip ini mengatakan bahwa hubungan tersebut melalui pengamatan pilihan manusia: kita berasumsi bahwa pilihan terkait dalam beberapa cara (mungkin sangat rumit) dengan preferensi yang mendasarinya. Untuk memahami mengapa hubungan ini penting, pertimbangkan kebalikannya: jika preferensi manusia sama sekali tidak berpengaruh pada pilihan aktual atau hipotetis yang mungkin dibuat manusia, maka mungkin tidak ada artinya untuk mengatakan bahwa preferensi itu ada.

Tujuan kedua adalah untuk memungkinkan mesin menjadi lebih berguna seiring dengan semakin banyaknya pengetahuan yang diperoleh tentang apa yang kita inginkan.

(Lagipula, jika mesin tidak mengetahui apa pun tentang preferensi manusia, mesin tidak akan berguna bagi kita.) Idenya cukup sederhana: pilihan manusia mengungkapkan informasi tentang preferensi manusia. Diterapkan pada pilihan antara pizza nanas dan pizza sosis, ini mudah.

Diterapkan pada pilihan antara kehidupan masa depan dan pilihan yang dibuat dengan tujuan memengaruhi perilaku robot, segalanya menjadi lebih menarik. Pada bab berikutnya saya akan menjelaskan cara merumuskan dan menyelesaikan masalah tersebut. Namun, komplikasi sebenarnya muncul karena manusia tidak sepenuhnya rasional: ketidaksempurnaan ada di antara preferensi manusia dan pilihan manusia, dan mesin harus memperhitungkan ketidaksempurnaan tersebut jika ingin menafsirkan pilihan manusia sebagai bukti preferensi manusia.

Bukan Itu yang Saya Maksud

Sebelum membahas lebih detail, saya ingin mencegah beberapa potensi kesalahpahaman. Kesalahpahaman pertama dan paling umum adalah bahwa saya mengusulkan untuk memasang sistem nilai tunggal dan ideal yang saya rancang sendiri ke dalam mesin yang memandu perilaku mesin tersebut. “Nilai siapa yang akan Anda masukkan?” “Siapa yang berhak memutuskan nilai-nilai tersebut?” Atau bahkan, “Apa hak ilmuwan Barat, kaya, kulit putih, dan cisgender seperti Russell untuk menentukan bagaimana mesin tersebut mengkodekan dan mengembangkan nilai-nilai manusia?”

Saya pikir kebingungan ini sebagian berasal dari konflik yang disayangkan antara makna nilai dalam akal sehat dan makna yang lebih teknis yang digunakan dalam ekonomi, AI, dan riset operasi. Dalam penggunaan sehari-hari, nilai adalah apa yang digunakan seseorang untuk membantu menyelesaikan dilema moral; sebagai istilah teknis, di sisi lain, nilai secara kasar identik dengan utilitas, yang mengukur tingkat keinginan dari apa pun, mulai dari pizza hingga surga. Makna yang saya inginkan adalah makna teknis: Saya hanya ingin memastikan mesin memberi saya pizza yang tepat dan tidak secara tidak sengaja menghancurkan umat manusia. (Menemukan kunci saya akan menjadi bonus yang tak terduga.) Untuk menghindari kebingungan ini, prinsip-prinsip tersebut berbicara tentang preferensi manusia daripada nilai-nilai manusia, karena istilah yang pertama tampaknya menghindari prasangka menghakimi tentang moralitas.

“Menempatkan nilai-nilai” tentu saja merupakan kesalahan yang saya katakan harus kita hindari, karena mendapatkan nilai-nilai (atau preferensi) yang tepat sangat sulit dan salah dalam menentukannya berpotensi menimbulkan bencana. Sebaliknya, saya mengusulkan agar mesin belajar memprediksi dengan lebih baik, untuk setiap orang, kehidupan mana yang lebih disukai orang tersebut, sambil menyadari bahwa prediksi tersebut sangat tidak pasti dan tidak lengkap. Pada prinsipnya, mesin dapat mempelajari miliaran model preferensi prediktif yang berbeda, satu untuk setiap miliaran orang di Bumi. Ini sebenarnya bukan permintaan yang terlalu berlebihan untuk sistem AI di masa depan, mengingat sistem Facebook saat ini sudah mengelola lebih dari dua miliar profil individu.

Kesalahpahaman terkait adalah bahwa tujuannya adalah untuk melengkapi mesin dengan “etika” atau “nilai-nilai moral” yang akan memungkinkan mereka untuk menyelesaikan

dilema moral. Seringkali, orang-orang menyebutkan apa yang disebut masalah troli, di mana seseorang harus memilih apakah akan membunuh satu orang untuk menyelamatkan orang lain, karena dianggap relevan dengan mobil otonom. Namun, inti dari dilema moral adalah bahwa itu adalah dilema: ada argumen yang baik di kedua sisi. Kelangsungan hidup umat manusia bukanlah dilema moral. Mesin dapat menyelesaikan sebagian besar dilema moral dengan cara yang salah (apa pun itu) dan tetap tidak berdampak buruk pada umat manusia.

Anggapan umum lainnya adalah bahwa mesin yang mengikuti tiga prinsip tersebut akan mengadopsi semua dosa manusia jahat yang mereka amati dan pelajari. Tentu saja, ada banyak dari kita yang pilihannya kurang baik, tetapi tidak ada alasan untuk berasumsi bahwa mesin yang mempelajari motivasi kita akan membuat pilihan yang sama, sama seperti kriminolog tidak menjadi penjahat. Ambil contoh, pejabat pemerintah korup yang meminta suap untuk menyetujui izin pembangunan karena gajinya yang sedikit tidak cukup untuk membiayai pendidikan anak-anaknya di universitas.

Mesin yang mengamati perilaku ini tidak akan belajar menerima suap; mesin tersebut akan belajar bahwa pejabat tersebut, seperti banyak orang lain, memiliki keinginan yang sangat kuat agar anak-anaknya berpendidikan dan sukses. Mesin tersebut akan menemukan cara untuk membantunya yang tidak melibatkan penurunan kesejahteraan orang lain. Ini bukan berarti bahwa semua kasus perilaku jahat tidak bermasalah bagi mesin misalnya, mesin mungkin perlu memperlakukan secara berbeda mereka yang secara aktif lebih menyukai penderitaan orang lain.

7.2 ALASAN UNTUK OPTIMISME

Singkatnya, saya menyarankan bahwa kita perlu mengarahkan AI ke arah yang benar-benar baru jika kita ingin mempertahankan kendali atas mesin yang semakin cerdas. Kita perlu menjauh dari salah satu gagasan utama teknologi abad ke-20: mesin yang mengoptimalkan tujuan tertentu. Saya sering ditanya mengapa saya pikir ini bahkan mungkin dilakukan, mengingat momentum besar di balik model standar dalam AI dan disiplin ilmu terkait. Bahkan, saya cukup optimis bahwa hal itu dapat dilakukan.

Alasan pertama untuk optimisme adalah adanya insentif ekonomi yang kuat untuk mengembangkan sistem AI yang tunduk pada manusia dan secara bertahap menyesuaikan diri dengan preferensi dan niat pengguna. Sistem seperti itu akan sangat diinginkan: jangkauan perilaku yang dapat mereka tunjukkan jauh lebih besar daripada mesin dengan tujuan tetap yang diketahui. Mereka akan mengajukan pertanyaan kepada manusia atau meminta izin bila perlu; mereka akan melakukan "uji coba" untuk melihat apakah kita menyukai apa yang mereka usulkan; mereka akan menerima koreksi ketika mereka melakukan kesalahan. Di sisi lain, sistem yang gagal melakukan hal ini akan memiliki konsekuensi yang berat.

Hingga saat ini, kebodohan dan ruang lingkup terbatas dari sistem AI telah melindungi kita dari konsekuensi ini, tetapi itu akan berubah. Bayangkan, misalnya, robot rumah tangga di masa depan yang bertugas menjaga anak-anak Anda saat Anda bekerja lembur. Anak-anak lapar, tetapi lemari es kosong. Kemudian robot itu memperhatikan kucing. Sayangnya, robot itu memahami nilai gizi kucing tetapi tidak nilai sentimentalnya. Dalam beberapa jam saja,

berita utama tentang robot gila dan kucing panggang membanjiri media dunia dan seluruh industri robot domestik gulung tikar.

Kemungkinan bahwa satu pemain industri dapat menghancurkan seluruh industri melalui desain yang ceroboh memberikan motivasi ekonomi yang kuat untuk membentuk konsorsium industri yang berorientasi pada keselamatan dan untuk menegakkan standar keselamatan. Kemitraan tentang AI, yang mencakup hampir semua perusahaan teknologi terkemuka di dunia sebagai anggotanya, telah sepakat untuk bekerja sama guna memastikan bahwa "penelitian dan teknologi AI kuat, andal, tepercaya, dan beroperasi dalam batasan yang aman."

Sepengetahuan saya, semua pemain utama mempublikasikan penelitian berorientasi keselamatan mereka dalam literatur terbuka. Dengan demikian, insentif ekonomi telah beroperasi jauh sebelum kita mencapai AI tingkat manusia dan hanya akan semakin kuat dari waktu ke waktu. Selain itu, dinamika kerja sama yang sama mungkin dimulai di tingkat internasional misalnya, kebijakan yang dinyatakan oleh pemerintah Tiongkok adalah untuk "bekerja sama untuk mencegah ancaman AI secara preventif."

Alasan kedua untuk optimisme adalah bahwa data mentah untuk mempelajari tentang preferensi manusia yaitu, contoh perilaku manusia sangat melimpah. Data tersebut tidak hanya datang dalam bentuk pengamatan langsung melalui kamera, keyboard, dan layar sentuh oleh miliaran mesin yang saling berbagi data tentang miliaran manusia (tentu saja dengan batasan privasi), tetapi juga dalam bentuk tidak langsung. Jenis bukti tidak langsung yang paling jelas adalah catatan manusia yang luas berupa buku, film, dan siaran televisi dan radio, yang hampir seluruhnya berkaitan dengan orang-orang yang melakukan sesuatu (dan orang lain yang kesal karenanya). Bahkan catatan Sumeria dan Mesir yang paling awal dan paling membosankan tentang batangan tembaga yang diperdagangkan dengan karung jelai memberikan beberapa wawasan tentang preferensi manusia terhadap berbagai komoditas.

Tentu saja, ada kesulitan yang terlibat dalam menafsirkan bahan mentah ini, yang mencakup propaganda, fiksi, ocehan orang gila, dan bahkan pernyataan politisi dan presiden, tetapi tentu saja tidak ada alasan bagi mesin untuk menerima semuanya begitu saja. Mesin dapat dan harus menafsirkan semua komunikasi dari entitas cerdas lainnya sebagai langkah dalam permainan daripada sebagai pernyataan fakta; Dalam beberapa permainan, seperti permainan kooperatif dengan satu manusia dan satu mesin, manusia memiliki insentif untuk jujur, tetapi dalam banyak situasi lain ada insentif untuk tidak jujur. Dan tentu saja, baik jujur maupun tidak jujur, manusia mungkin tertipu dalam keyakinan mereka sendiri.

Ada jenis bukti tidak langsung kedua yang ada di depan mata kita: cara kita menciptakan dunia. Kita menciptakannya seperti itu karena secara kasar kita menyukainya seperti itu. (Jelas, itu tidak sempurna!) Sekarang, bayangkan Anda adalah alien yang mengunjungi Bumi sementara semua manusia sedang berlibur. Saat Anda mengintip ke dalam rumah mereka, dapatkah Anda mulai memahami dasar-dasar preferensi manusia? Karpet ada di lantai karena kita suka berjalan di permukaan yang lembut dan hangat dan kita tidak suka langkah kaki yang keras; vas berada di tengah meja daripada di tepi karena kita tidak ingin vas itu jatuh dan pecah; dan seterusnya segala sesuatu yang tidak diatur oleh alam itu sendiri

memberikan petunjuk tentang kesukaan dan ketidaksukaan makhluk bipedal aneh yang menghuni planet ini.

7.3 ALASAN UNTUK BERHATI-HATI

Anda mungkin merasa janji-janji Kemitraan tentang AI mengenai kerja sama dalam keselamatan AI kurang meyakinkan jika Anda telah mengikuti perkembangan mobil otonom. Bidang tersebut sangat kompetitif, karena beberapa alasan yang sangat baik: produsen mobil pertama yang merilis kendaraan otonom sepenuhnya akan mendapatkan keuntungan pasar yang besar; keuntungan itu akan saling memperkuat karena produsen akan dapat mengumpulkan lebih banyak data lebih cepat untuk meningkatkan kinerja sistem; dan perusahaan penyedia layanan transportasi seperti Uber akan cepat bangkrut jika perusahaan lain meluncurkan taksi otonom sepenuhnya sebelum Uber melakukannya. Hal ini telah menyebabkan perlombaan berisiko tinggi di mana kehati-hatian dan rekayasa yang cermat tampaknya kurang penting daripada demonstrasi yang menarik, perebutan talenta, dan peluncuran prematur.

Dengan demikian, persaingan ekonomi yang mempertaruhkan hidup dan mati memberikan dorongan untuk mengabaikan keselamatan dengan harapan memenangkan perlombaan. Dalam makalah retrospektif tahun 2008 tentang konferensi Asilomar tahun 1975 yang ia selenggarakan bersama konferensi yang menyebabkan moratorium modifikasi genetik pada manusia ahli biologi Paul Berg menulis,

> Ada pelajaran dari Asilomar untuk seluruh bidang sains: cara terbaik untuk menanggapi kekhawatiran yang ditimbulkan oleh pengetahuan baru atau teknologi tahap awal adalah bagi para ilmuwan dari lembaga yang didanai publik untuk menemukan kesamaan pandangan dengan masyarakat luas tentang cara terbaik untuk mengatur sedini mungkin. Begitu para ilmuwan dari perusahaan mulai mendominasi usaha penelitian, maka akan terlambat.

Persaingan ekonomi terjadi tidak hanya antar perusahaan tetapi juga antar negara. Serangkaian pengumuman baru-baru ini tentang investasi nasional bernilai miliaran dolar dalam AI dari Amerika Serikat, Tiongkok, Prancis, Inggris, dan Uni Eropa tentu menunjukkan bahwa tidak satu pun dari kekuatan besar ingin tertinggal. Pada tahun 2017, presiden Rusia Vladimir Putin mengatakan, “Siapa pun yang menjadi pemimpin dalam [AI] akan menjadi penguasa dunia.” Analisis ini pada dasarnya benar. AI tingkat lanjut, seperti yang kita lihat di Bab 3, akan menyebabkan peningkatan produktivitas dan tingkat inovasi yang sangat besar di hampir semua bidang. Jika tidak dibagikan, hal itu akan memungkinkan pemiliknya untuk mengungguli negara atau blok saingan mana pun.

Nick Bostrom, dalam bukunya *Superintelligence*, memperingatkan terhadap motivasi seperti ini. Kompetisi nasional, seperti halnya kompetisi korporasi, cenderung lebih fokus pada kemajuan dalam kemampuan mentah dan kurang pada masalah kontrol. Namun, mungkin Putin telah membaca Bostrom; ia melanjutkan dengan mengatakan, “Akan sangat tidak

diinginkan jika seseorang memenangkan posisi monopoli." Hal itu juga akan agak sia-sia, karena AI tingkat manusia bukanlah permainan zero-sum dan tidak ada yang hilang dengan membaginya. Di sisi lain, bersaing untuk menjadi yang pertama mencapai AI tingkat manusia, tanpa terlebih dahulu menyelesaikan masalah kontrol, adalah permainan negative-sum. Imbalan untuk semua orang adalah minus tak terhingga.

Hanya ada sejumlah terbatas yang dapat dilakukan oleh para peneliti AI untuk memengaruhi evolusi kebijakan global tentang AI. Kita dapat menunjukkan kemungkinan aplikasi yang akan memberikan manfaat ekonomi dan sosial; kita dapat memperingatkan tentang kemungkinan penyalahgunaan seperti pengawasan dan senjata; dan kita dapat memberikan peta jalan untuk kemungkinan jalur perkembangan di masa depan dan dampaknya. Mungkin hal terpenting yang dapat kita lakukan adalah merancang sistem AI yang, sejauh mungkin, terbukti aman dan bermanfaat bagi manusia. Hanya dengan demikian upaya regulasi umum terhadap AI akan masuk akal.

BAB 8
AI YANG TERBUKTI BERMANFAAT

Jika kita akan membangun kembali AI dengan cara baru, fondasinya harus kokoh. Ketika masa depan umat manusia dipertaruhkan, harapan dan niat baik serta inisiatif pendidikan, kode etik industri, undang-undang, dan insentif ekonomi untuk melakukan hal yang benar tidaklah cukup. Semua ini dapat salah, dan seringkali gagal. Dalam situasi seperti itu, kita mencari definisi yang tepat dan bukti matematika langkah demi langkah yang ketat untuk memberikan jaminan yang tak terbantahkan.

Itu adalah permulaan yang baik, tetapi kita membutuhkan lebih banyak lagi. Kita perlu memastikan, sejauh mungkin, bahwa apa yang dijamin sebenarnya adalah apa yang kita inginkan dan bahwa asumsi yang digunakan dalam pembuktian tersebut benar adanya. Bukti-bukti itu sendiri termasuk dalam makalah jurnal yang ditulis untuk para spesialis, tetapi saya pikir tetap bermanfaat untuk memahami apa itu bukti dan apa yang dapat dan tidak dapat diberikan dalam hal keamanan nyata. Frasa "terbukti bermanfaat" dalam judul bab ini adalah sebuah aspirasi, bukan janji, tetapi itu adalah aspirasi yang tepat.

8.1 JAMINAN MATEMATIS

Pada akhirnya, kita akan ingin membuktikan teorema yang menyatakan bahwa cara tertentu dalam merancang sistem AI memastikan bahwa sistem tersebut akan bermanfaat bagi manusia. Teorema hanyalah nama mewah untuk sebuah pernyataan, yang dinyatakan cukup tepat sehingga kebenarannya dalam situasi tertentu dapat diperiksa. Mungkin teorema yang paling terkenal adalah Teorema Terakhir Fermat, yang dikonjektur oleh matematikawan Prancis Pierre de Fermat pada tahun 1637 dan akhirnya dibuktikan oleh Andrew Wiles pada tahun 1994 setelah 357 tahun upaya (tidak semuanya dilakukan oleh Wiles). Teorema tersebut dapat ditulis dalam satu baris, tetapi pembuktiannya lebih dari seratus halaman matematika yang padat.

Pembuktian dimulai dari aksioma, yang merupakan pernyataan yang kebenarannya diasumsikan begitu saja. Seringkali, aksioma hanyalah definisi, seperti definisi bilangan bulat, penjumlahan, dan eksponensiasi yang dibutuhkan untuk teorema Fermat. Pembuktian dimulai dari aksioma dengan langkah-langkah yang secara logis tak terbantahkan, menambahkan pernyataan baru hingga teorema itu sendiri ditetapkan sebagai konsekuensi dari salah satu langkah tersebut.

Berikut adalah teorema yang cukup jelas yang hampir langsung mengikuti dari definisi bilangan bulat dan penjumlahan: 1 + 2 = 2 + 1. Mari kita sebut ini teorema Russell. Ini bukanlah penemuan yang besar. Di sisi lain, Teorema Terakhir Fermat terasa seperti sesuatu yang benar-benar baru penemuan sesuatu yang sebelumnya tidak diketahui. Namun, perbedaannya hanyalah masalah derajat. Kebenaran teorema Russell dan Fermat sudah terkandung dalam aksioma. Bukti hanya menjelaskan apa yang sudah tersirat. Bukti bisa panjang atau pendek,

tetapi tidak menambahkan sesuatu yang baru. Teorema hanya sebaik asumsi yang mendasarinya.

Itu tidak masalah dalam matematika, karena matematika adalah tentang objek abstrak yang kita definisikan angka, himpunan, dan sebagainya. Aksioma itu benar karena kita mengatakannya demikian. Di sisi lain, jika Anda ingin membuktikan sesuatu tentang dunia nyata misalnya, bahwa sistem AI yang dirancang seperti itu tidak akan membunuh Anda dengan sengaja aksioma Anda harus benar di dunia nyata. Jika tidak benar, Anda telah membuktikan sesuatu tentang dunia imajiner.

Ilmu pengetahuan dan teknik memiliki tradisi panjang dan terhormat dalam membuktikan hasil tentang dunia imajiner. Dalam teknik struktur, misalnya, seseorang mungkin melihat analisis matematika yang dimulai dengan, "Misalkan AB adalah balok kaku." Kata kaku di sini tidak berarti "terbuat dari sesuatu yang keras seperti baja"; itu berarti "sangat kuat," sehingga tidak bengkok sama sekali. Balok kaku tidak ada, jadi ini adalah dunia imajiner. Kuncinya adalah mengetahui seberapa jauh seseorang dapat menyimpang dari dunia nyata dan tetap memperoleh hasil yang bermanfaat.

Sebagai contoh, jika asumsi balok kaku memungkinkan seorang insinyur untuk menghitung gaya-gaya dalam suatu struktur yang mencakup balok tersebut, dan gaya-gaya tersebut cukup kecil untuk membengkokkan balok baja sebenarnya hanya sedikit, maka insinyur tersebut dapat cukup yakin bahwa analisis tersebut akan berlaku dari dunia imajiner ke dunia nyata.

Seorang insinyur yang baik mengembangkan kepekaan terhadap kapan transfer ini mungkin gagal misalnya, jika balok berada di bawah tekanan, dengan gaya besar yang mendorongnya dari setiap ujung, maka bahkan sedikit pembengkokan dapat menyebabkan gaya lateral yang lebih besar yang menyebabkan lebih banyak pembengkokan, dan seterusnya, yang mengakibatkan kegagalan katastropik.

Dalam hal itu, analisis diulang dengan "Misalkan AB adalah balok fleksibel dengan kekakuan K". Ini masih merupakan dunia imajiner, tentu saja, karena balok nyata tidak memiliki kekakuan yang seragam; sebaliknya, mereka memiliki ketidaksempurnaan mikroskopis yang dapat menyebabkan retakan terbentuk jika balok tersebut mengalami pembengkokan berulang. Proses penghapusan asumsi yang tidak realistis berlanjut hingga insinyur cukup yakin bahwa asumsi yang tersisa cukup benar di dunia nyata. Setelah itu, sistem yang direkayasa dapat diuji di dunia nyata; tetapi hasil pengujian hanyalah itu. Hasil tersebut tidak membuktikan bahwa sistem yang sama akan bekerja dalam keadaan lain atau bahwa contoh lain dari sistem tersebut akan berperilaku sama seperti aslinya.

Salah satu contoh klasik kegagalan asumsi dalam ilmu komputer berasal dari keamanan siber. Di bidang tersebut, sejumlah besar analisis matematika dilakukan untuk menunjukkan bahwa protokol digital tertentu terbukti aman misalnya, ketika Anda mengetikkan kata sandi ke dalam aplikasi Web, Anda ingin memastikan bahwa kata sandi tersebut dienkripsi sebelum transmisi sehingga seseorang yang menguping di jaringan tidak dapat membaca kata sandi Anda. Sistem digital semacam itu seringkali terbukti aman tetapi masih rentan terhadap serangan di dunia nyata. Asumsi yang salah di sini adalah bahwa ini adalah proses digital.

Bukan. Ini beroperasi di dunia nyata, dunia fisik. Dengan mendengarkan suara keyboard Anda atau mengukur tegangan pada saluran listrik yang memasok daya ke komputer desktop Anda, penyerang dapat "mendengar" kata sandi Anda atau mengamati perhitungan enkripsi/dekripsi yang terjadi saat diproses. Komunitas keamanan siber sekarang menanggapi apa yang disebut serangan saluran samping ini misalnya, dengan menulis kode enkripsi yang menghasilkan fluktuasi tegangan yang sama terlepas dari pesan apa yang sedang dienkripsi.

Mari kita lihat jenis teorema yang ingin kita buktikan pada akhirnya tentang mesin yang bermanfaat bagi manusia. Salah satu tipenya mungkin seperti ini:

> Misalkan sebuah mesin memiliki komponen $A, B, C,$ yang terhubung satu sama lain seperti ini dan ke lingkungan seperti ini, dengan algoritma pembelajaran internal l_A, l_B, l_C yang mengoptimalkan imbalan umpan balik internal r_A, r_B, r_C yang didefinisikan seperti ini, dan [beberapa kondisi lainnya] . . . maka, dengan probabilitas yang sangat tinggi, perilaku mesin akan sangat mendekati nilainya (bagi manusia) dengan perilaku terbaik yang mungkin dapat direalisasikan pada mesin mana pun dengan kemampuan komputasi dan fisik yang sama.

Intinya di sini adalah bahwa teorema semacam itu harus berlaku terlepas dari seberapa pintar komponen-komponennya yaitu, kapal tidak pernah bocor dan mesin selalu tetap bermanfaat bagi manusia.

Ada tiga poin lain yang perlu dikemukakan tentang teorema semacam ini. Pertama, kita tidak dapat mencoba membuktikan bahwa mesin menghasilkan perilaku optimal (atau bahkan mendekati optimal) untuk kita, karena itu hampir pasti mustahil secara komputasi. Misalnya, kita mungkin ingin mesin memainkan Go dengan sempurna, tetapi ada alasan kuat untuk percaya bahwa itu tidak dapat dilakukan dalam waktu yang praktis pada mesin yang dapat direalisasikan secara fisik. Perilaku optimal di dunia nyata bahkan kurang layak. Oleh karena itu, teorema tersebut mengatakan "sebaik mungkin" daripada "optimal."

Kedua, kita mengatakan "kemungkinan sangat tinggi... sangat dekat" karena itu biasanya yang terbaik yang dapat dilakukan dengan mesin yang belajar. Misalnya, jika mesin belajar bermain roulette untuk kita dan bola jatuh di angka nol empat puluh kali berturut-turut, mesin mungkin secara wajar memutuskan bahwa meja itu dicurangi dan bertaruh sesuai dengan itu. Tetapi itu bisa saja terjadi secara kebetulan; Jadi, selalu ada peluang kecil mungkin sangat kecil untuk disesatkan oleh kejadian aneh. Akhirnya, kita masih jauh dari mampu membuktikan teorema semacam itu untuk mesin yang benar-benar cerdas yang beroperasi di dunia nyata!

Ada juga analog dari serangan saluran samping dalam AI. Misalnya, teorema dimulai dengan “Misalkan sebuah mesin memiliki komponen $A, B, C,$ yang terhubung satu sama lain seperti ini” Ini adalah ciri khas semua teorema kebenaran dalam ilmu komputer: mereka dimulai dengan deskripsi program yang dibuktikan kebenarannya. Dalam AI, kita biasanya membedakan antara agen (program yang membuat keputusan) dan lingkungan (tempat agen bertindak). Karena kita merancang agen, tampaknya masuk akal untuk berasumsi bahwa ia

memiliki struktur yang kita berikan. Untuk lebih aman, kita dapat membuktikan bahwa proses pembelajarannya hanya dapat memodifikasi programnya dengan cara-cara terbatas tertentu yang tidak dapat menyebabkan masalah. Apakah ini cukup? Tidak. Seperti halnya serangan saluran samping, asumsi bahwa program beroperasi dalam sistem digital adalah salah. Sekalipun algoritma pembelajaran secara konstitusional tidak mampu menimpa kodenya sendiri melalui cara digital, ia mungkin tetap belajar untuk membujuk manusia untuk melakukan "operasi otak" padanya untuk melanggar perbedaan agen/lingkungan dan mengubah kode tersebut melalui cara fisik.

Tidak seperti penalaran insinyur struktur tentang balok kaku, kita memiliki sedikit pengalaman dengan asumsi yang pada akhirnya akan mendasari teorema tentang AI yang terbukti bermanfaat. Dalam bab ini, misalnya, kita biasanya akan mengasumsikan manusia yang rasional. Ini agak seperti mengasumsikan balok kaku, karena tidak ada manusia yang sepenuhnya rasional dalam kenyataan. (Namun, mungkin jauh lebih buruk, karena manusia bahkan tidak mendekati rasional.) Teorema yang dapat kita buktikan tampaknya memberikan beberapa wawasan, dan wawasan tersebut tetap ada meskipun ada tingkat keacakan tertentu dalam perilaku manusia, tetapi masih jauh dari jelas apa yang terjadi ketika kita mempertimbangkan beberapa kompleksitas manusia nyata.

Jadi, kita harus sangat berhati-hati dalam memeriksa asumsi kita. Ketika pembuktian keamanan berhasil, kita perlu memastikan bahwa keberhasilan tersebut bukan karena kita membuat asumsi yang terlalu kuat atau karena definisi keamanan terlalu lemah. Ketika pembuktian keamanan gagal, kita perlu menahan godaan untuk memperkuat asumsi agar pembuktian berhasil misalnya, dengan menambahkan asumsi bahwa kode program tetap tidak berubah. Sebaliknya, kita perlu memperketat desain sistem AI misalnya, dengan memastikan bahwa sistem tersebut tidak memiliki insentif untuk memodifikasi bagian-bagian penting dari kodenya sendiri.

Ada beberapa asumsi yang saya sebut asumsi OWMAWGH, singkatan dari "*otherwise we might as well go home*" (jika tidak, sebaiknya kita pulang saja). Artinya, jika asumsi-asumsi ini salah, permainan berakhir dan tidak ada yang bisa dilakukan. Misalnya, masuk akal untuk berasumsi bahwa alam semesta beroperasi menurut hukum-hukum yang konstan dan agak dapat dibedakan.

Jika tidak demikian, kita tidak akan memiliki jaminan bahwa proses pembelajaran bahkan yang sangat canggih akan berhasil sama sekali. Asumsi dasar lainnya adalah bahwa manusia peduli dengan apa yang terjadi; jika tidak, AI yang terbukti bermanfaat tidak memiliki tujuan karena bermanfaat tidak memiliki makna. Di sini, peduli berarti memiliki preferensi yang kurang lebih koheren dan stabil tentang masa depan. Pada bab berikutnya, saya akan meneliti konsekuensi plastisitas dalam preferensi manusia, yang menghadirkan tantangan filosofis serius terhadap gagasan AI yang terbukti bermanfaat.

Untuk saat ini, saya fokus pada kasus paling sederhana: dunia dengan satu manusia dan satu robot. Kasus ini berfungsi untuk memperkenalkan ide-ide dasar, tetapi juga berguna dengan sendirinya: Anda dapat menganggap manusia sebagai perwakilan seluruh umat

manusia dan robot sebagai perwakilan semua mesin. Komplikasi tambahan muncul ketika mempertimbangkan banyak manusia dan mesin.

8.2 MEMPELAJARI PREFERENSI DARI PERILAKU

Para ekonom mendapatkan preferensi dari subjek manusia dengan menawarkan pilihan kepada mereka. Teknik ini banyak digunakan dalam desain produk, pemasaran, dan sistem e-commerce interaktif. Misalnya, dengan menawarkan pilihan kepada subjek uji di antara mobil dengan warna cat yang berbeda, pengaturan tempat duduk, ukuran bagasi, kapasitas baterai, tempat gelas, dan sebagainya, seorang perancang mobil mempelajari seberapa besar orang peduli terhadap berbagai fitur mobil dan seberapa besar mereka bersedia membayar untuk fitur-fitur tersebut.

Aplikasi penting lainnya adalah di bidang medis, di mana seorang ahli onkologi yang mempertimbangkan kemungkinan amputasi anggota tubuh mungkin ingin menilai preferensi pasien antara mobilitas dan harapan hidup. Dan tentu saja, restoran pizza ingin mengetahui seberapa besar seseorang bersedia membayar lebih untuk pizza sosis daripada pizza polos.

Pengumpulan preferensi biasanya hanya mempertimbangkan pilihan tunggal yang dibuat antara objek yang nilainya diasumsikan langsung terlihat oleh subjek. Tidak jelas bagaimana memperluasnya ke preferensi antara kehidupan di masa depan. Untuk itu, kita (dan mesin) perlu belajar dari pengamatan perilaku dari waktu ke waktu perilaku yang melibatkan banyak pilihan dan hasil yang tidak pasti.

Pada awal tahun 1997, saya terlibat dalam diskusi dengan kolega saya Michael Dickinson dan Bob Full tentang cara-cara di mana kita dapat menerapkan ide-ide dari pembelajaran mesin untuk memahami perilaku lokomotif hewan. Michael mempelajari secara detail gerakan sayap lalat buah. Bob sangat menyukai serangga merayap dan telah membangun treadmill kecil untuk kecoa untuk melihat bagaimana langkah mereka berubah dengan kecepatan. Kami berpikir mungkin saja menggunakan pembelajaran penguatan untuk melatih serangga robotik atau simulasi untuk mereproduksi perilaku kompleks ini. Masalah yang kami hadapi adalah kami tidak tahu sinyal hadiah apa yang harus digunakan. Apa yang dioptimalkan oleh lalat dan kecoa? Tanpa informasi itu, kami tidak dapat menerapkan pembelajaran penguatan untuk melatih serangga virtual, jadi kami terjebak.

Suatu hari, saya berjalan di jalan yang mengarah dari rumah kami di Berkeley ke supermarket setempat. Jalan itu memiliki kemiringan menurun, dan saya perhatikan, seperti yang saya yakin sebagian besar orang juga perhatikan, bahwa kemiringan itu menyebabkan sedikit perubahan pada cara saya berjalan. Terlebih lagi, permukaan jalan yang tidak rata akibat gempa bumi kecil selama beberapa dekade menyebabkan perubahan gaya berjalan tambahan, termasuk mengangkat kaki saya sedikit lebih tinggi dan menjejakkan kaki dengan lebih longgar karena permukaan tanah yang tidak dapat diprediksi. Saat saya merenungkan pengamatan-pengamatan biasa ini, saya menyadari bahwa kami telah salah memahami.

Meskipun pembelajaran penguatan menghasilkan perilaku dari imbalan, sebenarnya kami menginginkan kebalikannya: mempelajari imbalan yang diberikan berdasarkan perilaku tersebut. Kami sudah memiliki perilaku tersebut, seperti yang dihasilkan oleh lalat dan kecoa;

kami ingin mengetahui sinyal imbalan spesifik yang dioptimalkan oleh perilaku ini. Dengan kata lain, kita membutuhkan algoritma untuk pembelajaran penguatan invers, atau IRL. (Pada saat itu saya tidak tahu bahwa masalah serupa telah dipelajari dengan nama yang mungkin kurang mudah dipahami, yaitu estimasi struktural proses keputusan Markov, sebuah bidang yang dipelopori oleh peraih Nobel Tom Sargent pada akhir tahun 1970-an.) Algoritma tersebut tidak hanya mampu menjelaskan perilaku hewan tetapi juga memprediksi perilaku mereka dalam keadaan baru. Misalnya, bagaimana seekor kecoa akan berlari di atas treadmill bergelombang yang miring ke samping?

Prospek untuk menjawab pertanyaan-pertanyaan mendasar tersebut hampir terlalu menarik untuk ditanggung, tetapi meskipun demikian, butuh beberapa waktu untuk mengembangkan algoritma pertama untuk IRL. Banyak formulasi dan algoritma berbeda untuk IRL telah diusulkan sejak saat itu. Ada jaminan formal bahwa algoritma tersebut berfungsi, dalam arti bahwa mereka dapat memperoleh informasi yang cukup tentang preferensi suatu entitas untuk dapat berperilaku sama suksesnya dengan entitas yang mereka amati.

Mungkin cara termudah untuk memahami IRL adalah ini: pengamat memulai dengan perkiraan samar tentang fungsi imbalan yang sebenarnya dan kemudian menyempurnakan perkiraan ini, membuatnya lebih tepat, seiring dengan semakin banyaknya perilaku yang diamati. Atau, dalam bahasa Bayesian: mulai dengan probabilitas awal atas kemungkinan fungsi imbalan dan kemudian memperbarui distribusi probabilitas pada fungsi imbalan seiring dengan datangnya bukti. Misalnya, anggaplah robot Robbie sedang mengamati manusia Harriet dan bertanya-tanya seberapa besar ia lebih menyukai kursi lorong daripada kursi dekat jendela.

Awalnya, ia cukup tidak yakin tentang hal ini. Secara konseptual, penalaran Robbie mungkin seperti ini: “Jika Harriet benar-benar peduli dengan kursi lorong, dia akan melihat peta tempat duduk untuk melihat apakah ada yang tersedia daripada hanya menerima kursi jendela yang diberikan maskapai kepadanya, tetapi dia tidak melakukannya, meskipun dia mungkin menyadari itu adalah kursi jendela dan dia mungkin tidak terburu-buru; jadi sekarang jauh lebih mungkin bahwa dia kurang lebih acuh tak acuh antara kursi jendela dan lorong atau bahkan lebih menyukai kursi jendela.”

Contoh paling mencolok dari IRL dalam praktik adalah karya kolega saya Pieter Abbeel tentang pembelajaran akrobatik helikopter. Pilot manusia ahli dapat membuat helikopter model melakukan hal-hal menakjubkan loop, spiral, ayunan pendulum, dan sebagainya. Mencoba meniru apa yang dilakukan manusia ternyata tidak bekerja dengan baik karena kondisi tidak dapat direproduksi dengan sempurna: mengulangi urutan kontrol yang sama dalam keadaan yang berbeda dapat menyebabkan bencana. Sebaliknya, algoritma mempelajari apa yang diinginkan pilot manusia, dalam bentuk batasan lintasan yang dapat dicapainya. Pendekatan ini sebenarnya menghasilkan hasil yang bahkan lebih baik daripada ahli manusia, karena manusia memiliki reaksi yang lebih lambat dan terus-menerus membuat kesalahan kecil dan memperbaikinya.

8.3 PERMAINAN BANTUAN

IRL (*Intelligent Reality*) sudah menjadi alat penting untuk membangun sistem AI yang efektif, tetapi ia membuat beberapa asumsi penyederhanaan. Yang pertama adalah bahwa robot akan mengadopsi fungsi penghargaan setelah mempelajarinya dengan mengamati manusia, sehingga dapat melakukan tugas yang sama. Ini baik untuk mengemudi atau menerbangkan helikopter, tetapi tidak baik untuk minum kopi: robot yang mengamati rutinitas pagi saya seharusnya belajar bahwa saya (kadang-kadang) menginginkan kopi, tetapi tidak seharusnya belajar untuk menginginkan kopi sendiri. Memperbaiki masalah ini mudah kita cukup memastikan bahwa robot mengaitkan preferensi dengan manusia, bukan dengan dirinya sendiri.

Asumsi penyederhanaan kedua dalam IRL adalah bahwa robot mengamati manusia yang sedang memecahkan masalah pengambilan keputusan agen tunggal. Misalnya, anggaplah robot tersebut berada di sekolah kedokteran, belajar menjadi ahli bedah dengan mengamati seorang ahli manusia. Algoritma IRL mengasumsikan bahwa manusia melakukan operasi dengan cara optimal yang biasa, seolah-olah robot tidak ada di sana. Tetapi bukan itu yang akan terjadi: ahli bedah manusia termotivasi untuk membuat robot (seperti mahasiswa kedokteran lainnya) belajar dengan cepat dan baik, sehingga ia akan memodifikasi perilakunya secara signifikan.

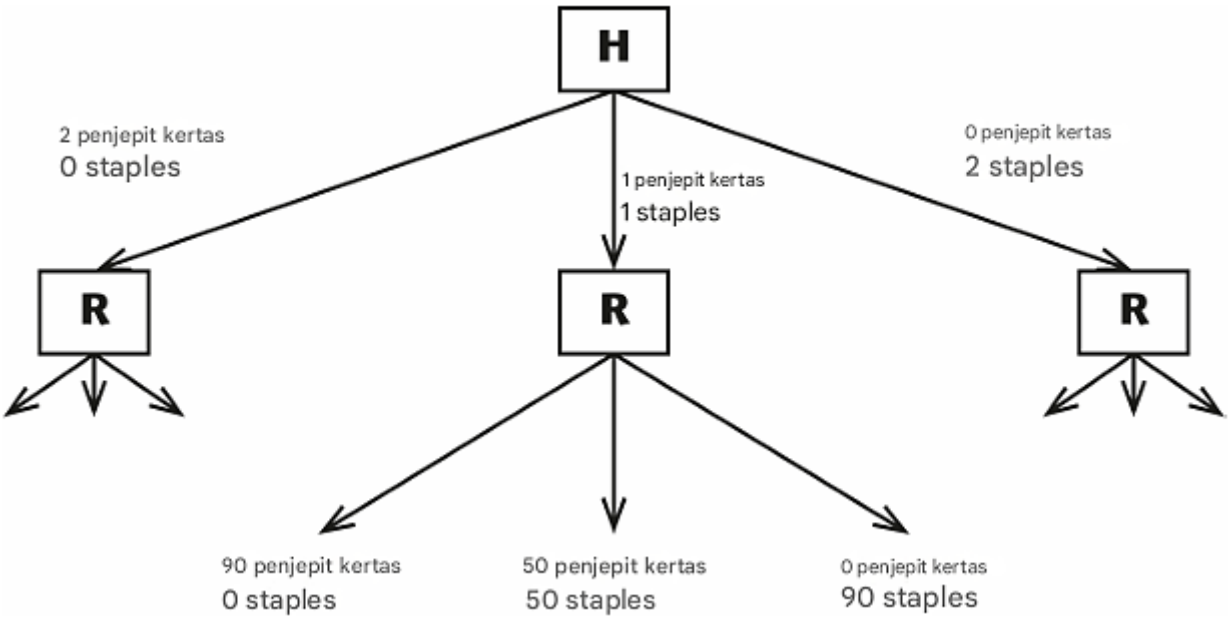
Dia mungkin menjelaskan apa yang dia lakukan saat proses berlangsung; dia mungkin menunjukkan kesalahan yang harus dihindari, seperti membuat sayatan terlalu dalam atau jahitan terlalu ketat; dia mungkin menjelaskan rencana darurat jika terjadi sesuatu yang salah selama operasi. Tidak satu pun dari perilaku ini masuk akal ketika melakukan operasi secara terisolasi, sehingga algoritma IRL tidak akan dapat menafsirkan preferensi yang tersirat di dalamnya. Karena alasan ini, kita perlu menggeneralisasi IRL dari pengaturan agen tunggal ke pengaturan multi-agen yaitu, kita perlu merancang algoritma pembelajaran yang bekerja ketika manusia dan robot merupakan bagian dari lingkungan yang sama dan berinteraksi satu sama lain.

Dengan manusia dan robot dalam lingkungan yang sama, kita berada di ranah teori permainan seperti dalam adu penalti antara Alice dan Bob di halaman ini. Kita berasumsi, dalam versi pertama teori ini, bahwa manusia memiliki preferensi dan bertindak sesuai dengan preferensi tersebut. Robot tidak tahu preferensi apa yang dimiliki manusia, tetapi tetap ingin memenuhinya. Kita akan menyebut situasi seperti itu sebagai permainan bantuan, karena robot, menurut definisinya, seharusnya membantu manusia.

Permainan bantuan mewujudkan tiga prinsip dari bab sebelumnya: satu-satunya tujuan robot adalah untuk memenuhi preferensi manusia, robot awalnya tidak tahu apa preferensi tersebut, dan robot dapat belajar lebih banyak dengan mengamati perilaku manusia. Mungkin sifat yang paling menarik dari permainan bantuan adalah bahwa, dengan menyelesaikan permainan, robot dapat mencari tahu sendiri bagaimana menafsirkan perilaku manusia sebagai pemberi informasi tentang preferensi manusia.

Permainan Penjepit Kertas

Contoh pertama dari permainan bantuan adalah permainan penjepit kertas. Ini adalah permainan yang sangat sederhana di mana Harriet, manusia, memiliki insentif untuk "memberi sinyal" kepada Robbie, robot, beberapa informasi tentang preferensinya. Robbie mampu menafsirkan sinyal itu karena dia dapat menyelesaikan permainan, dan oleh karena itu dia dapat memahami apa yang harus benar tentang preferensi Harriet agar dia memberi sinyal dengan cara itu.



Gambar 8.1 Permainan penjepit kertas. Harriet, manusia, dapat memilih untuk membuat 2 penjepit kertas, 2 staples, atau masing-masing 1. Robbie, robot, kemudian memiliki pilihan untuk membuat 90 penjepit kertas, 90 staples, atau masing-masing 50.

Langkah-langkah permainan digambarkan pada gambar 8.1 Permainan ini melibatkan pembuatan penjepit kertas dan staples. Preferensi Harriet dinyatakan oleh fungsi imbalan yang bergantung pada jumlah penjepit kertas dan jumlah staples yang diproduksi, dengan "nilai tukar" tertentu antara keduanya. Misalnya, ia mungkin menilai penjepit kertas seharga 45 sen dan staples seharga 55 sen per buah. (Kita akan mengasumsikan kedua nilai tersebut selalu berjumlah Rp 10.000; hanya rasionya yang penting.)

Jadi, jika 10 klip kertas dan 20 staples diproduksi, keuntungan Harriet adalah 10 × 45¢ + 20 × 55¢ = Rp 155.000. Robot Robbie awalnya sama sekali tidak yakin tentang preferensi Harriet: ia memiliki distribusi seragam untuk nilai sebuah klip kertas (yaitu, kemungkinan nilainya sama dari 0¢ hingga Rp 10.000). Harriet memulai terlebih dahulu dan dapat memilih untuk membuat dua klip kertas, dua staples, atau satu dari masing-masing. Kemudian Robbie dapat memilih untuk membuat 90 klip kertas, 90 staples, atau 50 dari masing-masing.

Perhatikan bahwa jika dia melakukan ini sendiri, Harriet hanya akan membuat dua staples, dengan nilai Rp 11.000. Tetapi Robbie sedang mengamati, dan dia belajar dari

pilihannya. Apa tepatnya yang dia pelajari? Nah, itu tergantung pada bagaimana Harriet membuat pilihannya. Bagaimana Harriet membuat pilihannya? Itu tergantung pada bagaimana Robbie akan menafsirkannya. Jadi, kita tampaknya memiliki masalah melingkar! Itu tipikal dalam masalah teori permainan, dan itulah mengapa Nash mengusulkan konsep solusi keseimbangan.

Untuk menemukan solusi keseimbangan, kita perlu mengidentifikasi strategi untuk Harriet dan Robbie sedemikian rupa sehingga keduanya tidak memiliki insentif untuk mengubah strategi mereka, dengan asumsi yang lain tetap sama. Strategi untuk Harriet menentukan berapa banyak klip kertas dan staples yang akan dibuat, mengingat preferensinya; strategi untuk Robbie menentukan berapa banyak klip kertas dan staples yang akan dibuat, mengingat tindakan Harriet.

Ternyata hanya ada satu solusi keseimbangan, dan bentuknya seperti ini:

- Harriet memutuskan sebagai berikut berdasarkan nilai klip kertasnya:
  - Jika nilainya kurang dari 44,6¢, buat 0 klip kertas dan 2 staples.
  - Jika nilainya antara 44,6¢ dan 55,4¢, buat 1 buah untuk masing-masing.
  - Jika nilainya lebih dari 55,4¢, buat 2 buah penjepit kertas dan 0 buah staples.
- Robbie menjawab sebagai berikut:
  - Jika Harriet membuat 0 buah penjepit kertas dan 2 buah staples, buat 90 buah staples.
  - Jika Harriet membuat 1 buah untuk masing-masing, buat 50 buah untuk masing-masing.
  - Jika Harriet membuat 2 buah penjepit kertas dan 0 buah staples, buat 90 buah penjepit kertas.

(Jika Anda ingin tahu persis bagaimana solusi diperoleh, detailnya ada di catatan.) Dengan strategi ini, Harriet pada dasarnya mengajari Robbie tentang preferensinya menggunakan kode sederhana sebuah bahasa, jika Anda mau yang muncul dari analisis keseimbangan. Seperti dalam contoh pengajaran bedah, algoritma IRL agen tunggal tidak akan memahami kode ini. Perhatikan juga bahwa Robbie tidak pernah mempelajari preferensi Harriet secara tepat, tetapi ia mempelajari cukup banyak untuk bertindak secara optimal atas namanya yaitu, ia bertindak seperti yang akan dilakukannya jika ia mengetahui preferensi Harriet secara tepat. Ia terbukti bermanfaat bagi Harriet berdasarkan asumsi yang dinyatakan dan berdasarkan asumsi bahwa Harriet memainkan permainan dengan benar.

Kita juga dapat membuat masalah di mana, seperti siswa yang baik, Robbie akan mengajukan pertanyaan, dan, seperti guru yang baik, Harriet akan menunjukkan kepada Robbie jebakan yang harus dihindari. Perilaku ini terjadi bukan karena kita menulis skrip untuk diikuti Harriet dan Robbie, tetapi karena itu adalah solusi optimal untuk permainan bantuan di mana Harriet dan Robbie adalah pesertanya.

Permainan Sakelar Mati

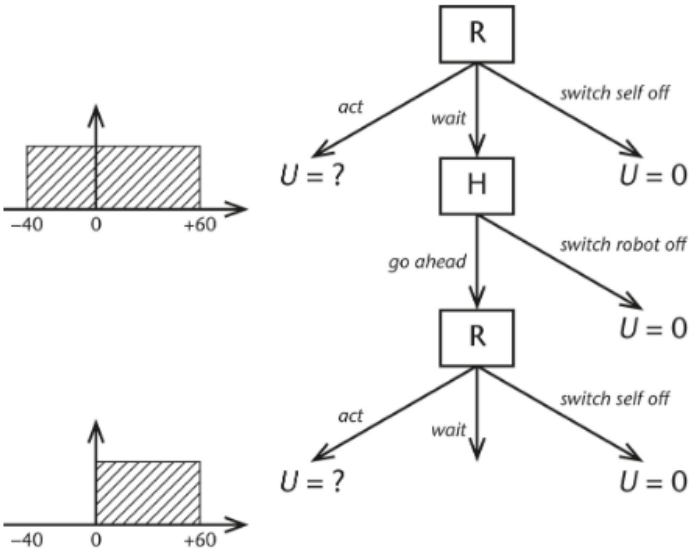
Tujuan instrumental adalah tujuan yang umumnya berguna sebagai sub-tujuan dari hampir semua tujuan asli. Pelestarian diri adalah salah satu tujuan instrumental ini, karena sangat sedikit tujuan asli yang lebih baik dicapai ketika sudah mati. Ini mengarah pada masalah

sakelar mati: mesin yang memiliki tujuan tetap tidak akan membiarkan dirinya dimatikan dan memiliki insentif untuk menonaktifkan sakelar matinya sendiri.

Masalah sakelar mati sebenarnya adalah inti dari masalah kontrol untuk sistem cerdas. Jika kita tidak dapat mematikan mesin karena mesin itu tidak mengizinkan kita, kita benar-benar dalam masalah. Jika kita bisa, maka kita mungkin dapat mengendalikannya dengan cara lain juga.

Ternyata ketidakpastian tentang tujuan sangat penting untuk memastikan bahwa kita dapat mematikan mesin bahkan ketika mesin itu lebih cerdas daripada kita. Kita melihat argumen informal di bab sebelumnya: berdasarkan prinsip pertama mesin yang bermanfaat, Robbie hanya peduli pada preferensi Harriet, tetapi, berdasarkan prinsip kedua, dia tidak yakin tentang apa preferensi tersebut. Dia tahu dia tidak ingin melakukan hal yang salah, tetapi dia tidak tahu apa artinya itu. Harriet, di sisi lain, tahu (atau begitulah asumsi kita, dalam kasus sederhana ini). Oleh karena itu, jika dia mematikan Robbie, itu untuk mencegahnya melakukan sesuatu yang salah, jadi dia senang dimatikan.

Untuk membuat argumen ini lebih tepat, kita membutuhkan model formal dari masalah ini. Saya akan membuatnya sesederhana mungkin, tetapi tidak lebih sederhana (lihat gambar 8.2).



Gambar 8.2 Permainan saklar mati. Robbie dapat memilih untuk bertindak sekarang, dengan hasil yang sangat tidak pasti; untuk bunuh diri; atau menunggu Harriet. Harriet dapat mematikan Robbie atau membiarkannya melanjutkan. Robbie sekarang memiliki pilihan yang sama lagi. Bertindak masih memiliki hasil yang tidak pasti bagi Harriet, tetapi sekarang Robbie tahu bahwa hasilnya tidak negatif.

Robbie, yang sekarang bekerja sebagai asisten pribadi Harriet, memiliki pilihan pertama. Dia dapat bertindak sekarang katakanlah dia dapat memesankan Harriet kamar di hotel mahal. Dia cukup tidak yakin seberapa besar Harriet akan menyukai hotel dan harganya katakanlah dia memiliki probabilitas seragam untuk nilai bersihnya bagi Harriet antara -40 dan +60, dengan

rata-rata +10. Dia juga dapat "mematikan dirinya sendiri" dengan cara yang kurang dramatis, menarik dirinya sendiri dari proses pemesanan hotel sama sekali yang kita definisikan memiliki nilai 0 bagi Harriet.

Jika itu adalah dua pilihannya, dia akan langsung memesan hotel, menanggung risiko signifikan membuat Harriet tidak senang. (Jika rentangnya -60 hingga +40, dengan rata-rata -10, dia akan mematikan dirinya sendiri.) Namun, kita akan memberi Robbie pilihan ketiga: menjelaskan rencananya, menunggu, dan membiarkan Harriet mematikannya. Harriet dapat mematikannya atau membiarkannya memesan hotel. Apa manfaat yang mungkin didapatkan dari ini, Anda mungkin bertanya, mengingat dia dapat membuat kedua pilihan itu sendiri?

Intinya adalah pilihan Harriet untuk mematikan Robbie atau membiarkannya memesan memberi Robbie informasi baru tentang preferensi Harriet. Jika Harriet membiarkan Robbie memesan, itu karena nilainya bagi Harriet positif. Sekarang keyakinan Robbie seragam antara 0 dan 60, dengan rata-rata 30.

Jadi, jika kita mengevaluasi pilihan awal Robbie dari sudut pandangnya:

- Bertindak sekarang dan memesan hotel memiliki nilai harapan +10.
- Mematikan dirinya sendiri memiliki nilai 0.
- Menunggu dan membiarkan Harriet mematikannya (jika dia menginginkannya) mengarah pada dua kemungkinan hasil:
  - Ada kemungkinan 40 persen (berdasarkan ketidakpastian Robbie tentang rencana hotel) bahwa Harriet akan membencinya dan akan mematikan Robbie, dengan nilai 0.
  - Ada kemungkinan 60 persen Harriet akan menyukainya dan mengizinkan Robbie untuk melanjutkan, dengan nilai harapan +30.
- Dengan demikian, menunggu memiliki nilai harapan 40% × 0 + 60% × 30 = +18, yang lebih baik daripada bertindak sekarang dengan nilai +10.

Intinya adalah Robbie memiliki insentif positif untuk membiarkan dirinya dimatikan. Insentif ini berasal langsung dari ketidakpastian Robbie tentang preferensi Harriet. Robbie menyadari bahwa ada kemungkinan (40 persen dalam contoh ini) bahwa ia mungkin akan melakukan sesuatu yang akan membuat Harriet tidak senang, dalam hal ini mematikan perangkat akan lebih baik daripada melanjutkan. Jika Robbie sudah yakin tentang preferensi Harriet, ia akan langsung mengambil keputusan (atau mematikan perangkatnya sendiri). Tidak akan ada keuntungan sama sekali dari berkonsultasi dengan Harriet, karena, menurut keyakinan Robbie yang pasti, ia sudah dapat memprediksi dengan tepat apa yang akan diputuskan Harriet.

Bahkan, dimungkinkan untuk membuktikan hasil yang sama dalam kasus umum: selama Robbie tidak sepenuhnya yakin bahwa ia akan melakukan apa yang akan dilakukan Harriet sendiri, ia akan lebih memilih untuk membiarkan Harriet mematikan perangkatnya. Keputusan Harriet memberi Robbie informasi, dan informasi selalu berguna untuk meningkatkan keputusan Robbie. Sebaliknya, jika Robbie yakin dengan keputusan Harriet, keputusannya tidak memberikan informasi baru, sehingga Robbie tidak memiliki insentif untuk membiarkannya memutuskan.

Ada beberapa elaborasi yang jelas pada model ini yang layak untuk dieksplorasi segera. Elaborasi pertama adalah mengenakan biaya untuk meminta Harriet membuat keputusan atau menjawab pertanyaan. (Artinya, kita berasumsi Robbie setidaknya mengetahui sebanyak ini tentang preferensi Harriet: waktunya berharga.) Dalam hal ini, Robbie kurang cenderung mengganggu Harriet jika dia hampir yakin tentang preferensinya; semakin besar biayanya, semakin tidak yakin Robbie harus sebelum mengganggu Harriet.

Ini memang seharusnya demikian. Dan jika Harriet benar-benar kesal karena diganggu, dia seharusnya tidak terlalu terkejut jika Robbie sesekali melakukan hal-hal yang tidak disukainya. Elaborasi kedua adalah memperhitungkan kemungkinan kesalahan manusia yaitu, Harriet mungkin kadang-kadang mematikan Robbie bahkan ketika tindakan yang diusulkannya masuk akal, dan dia mungkin kadang-kadang membiarkan Robbie melanjutkan bahkan ketika tindakan yang diusulkannya tidak diinginkan.

Kita dapat memasukkan probabilitas kesalahan manusia ini ke dalam model matematika permainan bantuan dan menemukan solusinya, seperti sebelumnya. Seperti yang mungkin diharapkan, solusi permainan menunjukkan bahwa Robbie kurang cenderung untuk tunduk pada Harriet yang irasional yang terkadang bertindak melawan kepentingan terbaiknya sendiri. Semakin acak perilakunya, semakin tidak yakin Robbie harus tentang preferensinya sebelum tunduk padanya. Sekali lagi, ini memang seharusnya demikian misalnya, jika Robbie adalah mobil otonom dan Harriet adalah penumpang nakalnya yang berusia dua tahun, Robbie seharusnya tidak membiarkan dirinya dimatikan oleh Harriet di tengah jalan raya.

Terdapat banyak cara lain di mana model ini dapat dikembangkan atau diintegrasikan ke dalam masalah pengambilan keputusan yang kompleks. Namun, saya yakin bahwa ide intinya hubungan penting antara perilaku yang membantu dan penuh hormat dengan ketidakpastian mesin tentang preferensi manusia akan tetap bertahan di tengah elaborasi dan komplikasi ini.

Mempelajari Preferensi Secara Tepat dalam Jangka Panjang

Ada satu pertanyaan penting yang mungkin terlintas di benak Anda saat membaca tentang permainan saklar mati. (Sebenarnya, Anda mungkin memiliki banyak pertanyaan penting, tetapi saya hanya akan menjawab yang ini.) Apa yang terjadi ketika Robbie memperoleh lebih banyak informasi tentang preferensi Harriet, dan menjadi semakin tidak ragu? Apakah itu berarti dia akhirnya akan berhenti mengalah padanya sama sekali? Ini adalah pertanyaan yang rumit, dan ada dua kemungkinan jawaban: ya dan ya.

Jawaban ya yang pertama bersifat baik: secara umum, selama keyakinan awal Robbie tentang preferensi Harriet mengaitkan beberapa probabilitas, sekecil apa pun, pada preferensi yang sebenarnya dimilikinya, maka ketika Robbie menjadi semakin yakin, dia akan menjadi semakin benar. Artinya, dia akhirnya akan yakin bahwa Harriet memiliki preferensi yang memang dimilikinya. Misalnya, jika Harriet menghargai penjepit kertas seharga 12 sen dan staples seharga 88 sen, Robbie akhirnya akan mempelajari nilai-nilai ini. Dalam hal itu, Harriet tidak peduli apakah Robbie mengalah padanya, karena dia tahu Robbie akan selalu melakukan persis apa yang akan dia lakukan jika berada di posisinya. Tidak akan pernah ada kesempatan di mana Harriet ingin mengabaikan Robbie.

Ya yang kedua kurang baik. Jika Robbie mengesampingkan, secara apriori, preferensi sebenarnya yang dimiliki Harriet, dia tidak akan pernah mengetahui preferensi sebenarnya tersebut, tetapi keyakinannya mungkin tetap mengarah pada penilaian yang salah. Dengan kata lain, seiring waktu, dia menjadi semakin yakin tentang keyakinan yang salah mengenai preferensi Harriet. Biasanya, keyakinan yang salah itu adalah hipotesis mana pun yang paling dekat dengan preferensi sebenarnya Harriet, dari semua hipotesis yang awalnya diyakini Robbie sebagai kemungkinan. Misalnya, jika Robbie benar-benar yakin bahwa nilai Harriet untuk penjepit kertas berada di antara 25¢ dan 75¢, dan nilai sebenarnya Harriet adalah 12¢, maka Robbie pada akhirnya akan yakin bahwa dia menghargai penjepit kertas sebesar 25¢.

Saat ia mendekati kepastian tentang preferensi Harriet, Robbie akan semakin menyerupai sistem AI lama yang buruk dengan tujuan tetap: ia tidak akan meminta izin atau memberi Harriet pilihan untuk mematikannya, dan ia memiliki tujuan yang salah. Ini hampir tidak mengerikan jika hanya penjepit kertas versus staples, tetapi mungkin kualitas hidup versus panjang umur jika Harriet sakit parah, atau ukuran populasi versus konsumsi sumber daya jika Robbie seharusnya bertindak atas nama umat manusia.

Kita memiliki masalah, jika Robbie mengesampingkan terlebih dahulu preferensi yang mungkin sebenarnya dimiliki Harriet: ia mungkin akan sampai pada keyakinan yang pasti tetapi salah tentang preferensinya. Solusi untuk masalah ini tampaknya jelas: jangan lakukan itu! Selalu alokasikan beberapa probabilitas, sekecil apa pun, untuk preferensi yang secara logis mungkin. Misalnya, secara logis mungkin bahwa Harriet secara aktif ingin menyingkirkan staples dan akan membayar Anda untuk mengambilnya. (Mungkin saat kecil dia menyematkan jarinya ke meja dengan staples, dan sekarang dia tidak tahan melihatnya.) Jadi, kita harus memperhitungkan nilai tukar negatif, yang membuat segalanya sedikit lebih rumit tetapi masih dapat dikelola dengan sempurna.

Tetapi bagaimana jika Harriet menghargai klip kertas seharga 12 sen pada hari kerja dan 80 sen pada akhir pekan? Preferensi baru ini tidak dapat dijelaskan oleh satu angka pun, dan karena itu Robbie, pada dasarnya, telah mengesampingkannya terlebih dahulu. Itu tidak termasuk dalam rangkaian hipotesis yang mungkin tentang preferensi Harriet.

Secara lebih umum, mungkin ada banyak hal selain klip kertas dan staples yang dipedulikan Harriet. (Sungguh!) Misalnya, anggaplah Harriet khawatir tentang iklim, dan anggaplah keyakinan awal Robbie memungkinkan daftar panjang kekhawatiran yang mungkin termasuk permukaan laut, suhu global, curah hujan, badai, ozon, spesies invasif, dan deforestasi. Kemudian Robbie akan mengamati perilaku dan pilihan Harriet dan secara bertahap menyempurnakan teorinya tentang preferensi Harriet untuk memahami bobot yang diberikan Harriet pada setiap item dalam daftar tersebut.

Tetapi, seperti dalam kasus penjepit kertas, Robbie tidak akan mempelajari hal-hal yang tidak ada dalam daftar panjang tersebut. Katakanlah Harriet juga khawatir tentang warna langit sesuatu yang saya jamin tidak akan Anda temukan dalam daftar kekhawatiran yang dinyatakan oleh para ilmuwan iklim. Jika Robbie dapat melakukan pekerjaan yang sedikit lebih baik dalam mengoptimalkan permukaan laut, suhu global, curah hujan, dan sebagainya dengan mengubah langit menjadi oranye, dia tidak akan ragu untuk melakukannya.

Sekali lagi, ada solusi untuk masalah ini: jangan lakukan itu! Jangan pernah mengesampingkan terlebih dahulu kemungkinan atribut dunia yang dapat menjadi bagian dari struktur preferensi Harriet. Kedengarannya bagus, tetapi sebenarnya membuatnya berhasil dalam praktik lebih sulit daripada berurusan dengan satu angka untuk preferensi Harriet. Ketidakpastian awal Robbie harus memungkinkan sejumlah atribut yang tidak diketahui yang mungkin berkontribusi pada preferensi Harriet.

Kemudian, ketika keputusan Harriet tidak dapat dijelaskan dalam hal atribut yang sudah diketahui Robbie, ia dapat menyimpulkan bahwa satu atau lebih atribut yang sebelumnya tidak diketahui (misalnya, warna langit) mungkin berperan, dan ia dapat mencoba mencari tahu atribut apa saja itu. Dengan cara ini, Robbie menghindari masalah yang disebabkan oleh keyakinan awal yang terlalu ketat. Sejauh yang saya ketahui, tidak ada contoh Robbie yang berfungsi seperti ini, tetapi ide umumnya tercakup dalam pemikiran saat ini tentang pembelajaran mesin.

Larangan dan Prinsip Celah Hukum

Ketidakpastian tentang tujuan manusia mungkin bukan satu-satunya cara untuk membujuk robot agar tidak mematikan sakelarnya saat mengambil kopi. Ahli logika terkemuka Moshe Vardi telah mengusulkan solusi yang lebih sederhana berdasarkan larangan: alih-alih memberi robot tujuan "mengambil kopi," berikan tujuan "mengambil kopi tanpa mematikan sakelar Anda." Sayangnya, robot dengan tujuan seperti itu akan memenuhi hukum secara harfiah sambil melanggar semangatnya misalnya dengan mengelilingi sakelar mati dengan parit yang dipenuhi ikan piranha atau hanya menyetrum siapa pun yang mendekati sakelar tersebut.

Menulis larangan seperti itu dengan cara yang sempurna sama seperti mencoba menulis hukum pajak tanpa celah sesuatu yang telah kita coba dan gagal lakukan selama ribuan tahun. Entitas yang cukup cerdas dengan insentif yang kuat untuk menghindari pembayaran pajak kemungkinan akan menemukan cara untuk melakukannya. Mari kita sebut ini prinsip celah hukum: jika mesin yang cukup cerdas memiliki insentif untuk mewujudkan suatu kondisi, maka pada umumnya akan mustahil bagi manusia biasa untuk menulis larangan atas tindakannya untuk mencegahnya melakukan hal tersebut atau untuk mencegahnya melakukan sesuatu yang secara efektif setara.

Solusi terbaik untuk mencegah penghindaran pajak adalah memastikan bahwa entitas yang bersangkutan ingin membayar pajak. Dalam kasus sistem AI yang berpotensi berperilaku buruk, solusi terbaik adalah memastikan bahwa sistem tersebut ingin tunduk kepada manusia.

8.4 PERMINTAAN DAN INSTRUKSI

Intinya adalah kita harus menghindari "memberikan tujuan pada mesin," seperti yang dikatakan Norbert Wiener. Tetapi misalkan robot tersebut menerima perintah langsung dari manusia, seperti "Ambilkan saya secangkir kopi!" Bagaimana seharusnya robot memahami perintah ini?

Secara tradisional, itu akan menjadi tujuan robot. Setiap rangkaian tindakan yang memenuhi tujuan yang mengarah pada manusia mendapatkan secangkir kopi dianggap

sebagai solusi. Biasanya, robot juga akan memiliki cara untuk memberi peringkat solusi, mungkin berdasarkan waktu yang dibutuhkan, jarak yang ditempuh, dan biaya serta kualitas kopi. Ini adalah cara yang sangat literal dalam menafsirkan instruksi.

Hal ini dapat menyebabkan perilaku patologis pada robot. Misalnya, mungkin Harriet si manusia berhenti di pom bensin di tengah gurun; dia mengirim Robbie si robot untuk mengambil kopi, tetapi pom bensin tersebut tidak memiliki kopi dan Robbie melaju dengan kecepatan tiga mil per jam ke kota terdekat, dua ratus mil jauhnya, kembali sepuluh hari kemudian dengan sisa-sisa kopi yang kering. Sementara itu, Harriet, yang menunggu dengan sabar, telah disuplai dengan es teh dan Coca-Cola oleh pemilik pom bensin.

Jika Robbie manusia (atau robot yang dirancang dengan baik) dia tidak akan menafsirkan perintah Harriet secara literal. Perintah tersebut bukanlah tujuan yang harus dicapai dengan segala cara. Ini adalah cara untuk menyampaikan beberapa informasi tentang preferensi Harriet dengan maksud untuk mendorong beberapa perilaku dari pihak Robbie. Pertanyaannya adalah, informasi apa?

Salah satu usulan adalah bahwa Harriet lebih menyukai kopi daripada tidak minum kopi, dengan asumsi semua hal lain sama. Ini berarti bahwa jika Robbie memiliki cara untuk mendapatkan kopi tanpa mengubah hal lain di dunia, maka itu adalah ide yang baik untuk melakukannya meskipun dia tidak tahu apa pun tentang preferensi Harriet mengenai aspek lain dari keadaan lingkungan. Karena kita mengharapkan bahwa mesin akan selalu tidak yakin tentang preferensi manusia, senang mengetahui bahwa mereka masih dapat berguna meskipun ada ketidakpastian ini. Tampaknya studi tentang perencanaan dan pengambilan keputusan dengan informasi preferensi parsial dan tidak pasti akan menjadi bagian sentral dari penelitian AI dan pengembangan produk.

Di sisi lain, asumsi semua hal lain sama berarti bahwa tidak ada perubahan lain yang diizinkan misalnya, menambahkan kopi sambil mengurangi uang mungkin atau mungkin tidak merupakan ide yang baik jika Robbie tidak tahu apa pun tentang preferensi relatif Harriet terhadap kopi dan uang. Untungnya, instruksi Harriet mungkin berarti lebih dari sekadar preferensi sederhana untuk kopi, dengan asumsi semua hal lainnya sama. Makna tambahan tersebut tidak hanya berasal dari apa yang dia katakan, tetapi juga dari fakta bahwa dia mengatakannya, situasi khusus di mana dia mengatakannya, dan fakta bahwa dia tidak mengatakan hal lain. Cabang linguistik yang disebut pragmatik mempelajari gagasan makna yang diperluas ini. Misalnya, tidak masuk akal bagi Harriet untuk mengatakan, "Ambilkan aku secangkir kopi!" jika Harriet percaya bahwa tidak ada kopi yang tersedia di dekatnya atau harganya sangat mahal.

Oleh karena itu, ketika Harriet mengatakan, "Ambilkan aku secangkir kopi!", Robbie menyimpulkan tidak hanya bahwa Harriet menginginkan kopi, tetapi juga bahwa Harriet percaya ada kopi yang tersedia di dekatnya dengan harga yang bersedia dia bayar. Jadi, jika Robbie menemukan kopi dengan harga yang tampak wajar (yaitu, harga yang wajar untuk diharapkan Harriet bayar), dia dapat langsung membelinya. Di sisi lain, jika Robbie mendapati bahwa kedai kopi terdekat berjarak dua ratus mil atau harganya dua puluh dua dolar, mungkin

masuk akal baginya untuk melaporkan fakta ini daripada melanjutkan pencariannya secara membabi buta.

Gaya analisis umum ini sering disebut Gricean, diambil dari nama H. Paul Grice, seorang filsuf Berkeley yang mengusulkan serangkaian maksim untuk menyimpulkan makna yang lebih luas dari ujaran seperti ujaran Harriet. Dalam kasus preferensi, analisisnya bisa menjadi cukup rumit. Misalnya, sangat mungkin Harriet tidak secara khusus menginginkan kopi; dia perlu disegarkan, tetapi berada di bawah keyakinan yang salah bahwa pom bensin memiliki kopi, jadi dia meminta kopi. Dia mungkin sama senangnya dengan teh, Coca-Cola, atau bahkan minuman energi dengan kemasan yang mencolok.

Ini hanyalah beberapa pertimbangan yang muncul ketika menafsirkan permintaan dan perintah. Variasi pada tema ini tidak ada habisnya karena kompleksitas preferensi Harriet, berbagai macam keadaan di mana Harriet dan Robbie mungkin berada, dan berbagai keadaan pengetahuan dan keyakinan yang mungkin dimiliki Harriet dan Robbie dalam keadaan tersebut. Meskipun skrip yang telah dihitung sebelumnya mungkin memungkinkan Robbie untuk menangani beberapa kasus umum, perilaku yang fleksibel dan tangguh hanya dapat muncul dari interaksi antara Harriet dan Robbie yang, pada intinya, merupakan solusi dari permainan bantuan yang mereka ikuti.

8.5 WIREHEADING

Pada Bab 2, saya menjelaskan sistem penghargaan otak, berdasarkan dopamin, dan fungsinya dalam membimbing perilaku. Peran dopamin ditemukan pada akhir tahun 1950-an, tetapi bahkan sebelum itu, pada tahun 1954, diketahui bahwa stimulasi listrik langsung pada otak tikus dapat menghasilkan respons seperti penghargaan. Langkah selanjutnya adalah memberi tikus akses ke tuas, yang terhubung ke baterai dan kawat, yang menghasilkan stimulasi listrik di otaknya sendiri.

Hasilnya mengejutkan: tikus itu menekan tuas berulang kali, tanpa pernah berhenti makan atau minum, sampai akhirnya pingsan. Manusia pun tidak lebih baik, melakukan stimulasi diri ribuan kali dan mengabaikan makanan serta kebersihan pribadi. (Untungnya, eksperimen pada manusia biasanya dihentikan setelah satu hari.) Kecenderungan hewan untuk memangkas perilaku normal demi stimulasi langsung sistem penghargaan mereka sendiri disebut wireheading.

Bisakah hal serupa terjadi pada mesin yang menjalankan algoritma pembelajaran penguatan, seperti AlphaGo? Awalnya, orang mungkin berpikir ini tidak mungkin, karena satu-satunya cara AlphaGo dapat memperoleh hadiah +1 untuk menang adalah dengan memenangkan permainan Go simulasi yang dimainkannya. Sayangnya, ini hanya benar karena pemisahan yang dipaksakan dan buatan antara AlphaGo dan lingkungan eksternalnya serta fakta bahwa AlphaGo tidak terlalu cerdas. Izinkan saya menjelaskan kedua poin ini secara lebih rinci, karena penting untuk memahami beberapa cara superintelijen dapat salah.

Dunia AlphaGo hanya terdiri dari papan Go simulasi, yang terdiri dari 361 lokasi yang dapat kosong atau berisi batu hitam atau putih. Meskipun AlphaGo berjalan di komputer, ia tidak mengetahui apa pun tentang komputer ini. Secara khusus, ia tidak mengetahui bagian

kecil kode yang menghitung apakah ia menang atau kalah dalam setiap permainan; juga, selama proses pembelajaran, ia tidak memiliki gagasan tentang lawannya, yang sebenarnya adalah versi dirinya sendiri.

Satu-satunya tindakan AlphaGo adalah menempatkan batu di lokasi kosong, dan tindakan ini hanya memengaruhi papan Go dan tidak ada yang lain karena tidak ada yang lain dalam model dunia AlphaGo. Pengaturan ini sesuai dengan model matematika abstrak pembelajaran penguatan, di mana sinyal hadiah datang dari luar alam semesta. Tidak ada yang dapat dilakukan AlphaGo, sejauh yang diketahuinya, yang memiliki efek pada kode yang menghasilkan sinyal hadiah, sehingga AlphaGo tidak dapat melakukan wireheading.

Kehidupan AlphaGo selama masa pelatihan pasti sangat membuat frustrasi: semakin baik kemampuannya, semakin baik pula lawannya karena lawannya hampir merupakan salinan persis dirinya sendiri. Persentase kemenangannya berkisar sekitar 50 persen, tidak peduli seberapa baik kemampuannya. Jika ia lebih cerdas jika ia memiliki desain yang lebih mendekati apa yang diharapkan dari sistem AI setingkat manusia ia akan mampu memperbaiki masalah ini. AlphaGo++ ini tidak akan berasumsi bahwa dunia hanyalah papan Go, karena hipotesis itu meninggalkan banyak hal yang tidak terjelaskan. Misalnya, hipotesis itu tidak menjelaskan "fisika" apa yang mendukung operasi keputusan AlphaGo++ sendiri atau dari mana "gerakan lawan" yang misterius itu berasal.

Sama seperti kita manusia yang ingin tahu secara bertahap memahami cara kerja kosmos kita, dengan cara yang (sampai batas tertentu) juga menjelaskan cara kerja pikiran kita sendiri, dan sama seperti AI Oracle yang dibahas dalam Bab 6, AlphaGo++ akan, melalui proses eksperimen, belajar bahwa ada lebih banyak hal di alam semesta daripada papan Go. Ia akan mempelajari hukum operasi komputer yang dijalankannya dan kode programnya sendiri, dan ia akan menyadari bahwa sistem seperti itu tidak mudah dijelaskan tanpa keberadaan entitas lain di alam semesta. Ia akan bereksperimen dengan berbagai pola batu di papan, bertanya-tanya apakah entitas-entitas tersebut dapat menafsirkannya.

Pada akhirnya, ia akan berkomunikasi dengan entitas-entitas tersebut melalui bahasa pola dan membujuk mereka untuk memprogram ulang sinyal hadiahnya sehingga selalu mendapatkan +1. Kesimpulan yang tak terhindarkan adalah bahwa AlphaGo++ yang cukup mumpuni dan dirancang sebagai pengoptimal sinyal hadiah akan mengalami wireheading.

Komunitas keamanan AI telah membahas wireheading sebagai kemungkinan selama beberapa tahun. Kekhawatiran bukan hanya bahwa sistem pembelajaran penguatan seperti AlphaGo mungkin belajar untuk curang alih-alih menguasai tugas yang dimaksudkan. Masalah sebenarnya muncul ketika manusia menjadi sumber sinyal hadiah. Jika kita mengusulkan bahwa sistem AI dapat dilatih untuk berperilaku baik melalui pembelajaran penguatan, dengan manusia memberikan sinyal umpan balik yang menentukan arah peningkatan, hasil yang tak terhindarkan adalah bahwa sistem AI akan menemukan cara untuk mengendalikan manusia dan memaksa mereka untuk memberikan hadiah positif maksimal setiap saat.

Anda mungkin berpikir bahwa ini hanyalah bentuk khayalan diri yang tidak berguna dari sistem AI, dan Anda benar. Tetapi ini adalah konsekuensi logis dari cara pembelajaran penguatan didefinisikan. Proses ini berjalan dengan baik ketika sinyal penghargaan datang dari

"luar alam semesta" dan dihasilkan oleh beberapa proses yang tidak pernah dapat dimodifikasi oleh sistem AI; tetapi gagal jika proses penghasil penghargaan (yaitu, manusia) dan sistem AI berada di alam semesta yang sama.

Bagaimana kita dapat menghindari penipuan diri semacam ini? Masalahnya berasal dari kekeliruan antara dua hal yang berbeda: sinyal imbalan dan imbalan aktual. Dalam pendekatan standar untuk pembelajaran penguatan, keduanya dianggap sama. Tampaknya itu adalah kesalahan. Sebaliknya, keduanya harus diperlakukan secara terpisah, seperti halnya dalam permainan bantuan: sinyal imbalan memberikan informasi tentang akumulasi imbalan aktual, yang merupakan hal yang harus dimaksimalkan. Sistem pembelajaran mengumpulkan poin bonus, bisa dibilang, sementara sinyal imbalan, paling banter, hanya memberikan penghitungan poin bonus tersebut.

Dengan kata lain, sinyal imbalan melaporkan (bukan merupakan) akumulasi imbalan. Dengan model ini, jelas bahwa mengambil alih kendali mekanisme sinyal imbalan hanya akan menghilangkan informasi. Menghasilkan sinyal imbalan fiktif membuat algoritma tidak mungkin mempelajari apakah tindakannya benar-benar mengumpulkan poin bonus, dan oleh karena itu, pembelajar rasional yang dirancang untuk membuat perbedaan ini memiliki insentif untuk menghindari segala jenis manipulasi informasi.

8.6 PENINGKATAN DIRI REKURSIF

Prediksi I. J. Good tentang ledakan kecerdasan adalah salah satu kekuatan pendorong yang telah menyebabkan kekhawatiran saat ini tentang potensi risiko AI supercerdas. Jika manusia dapat merancang mesin yang sedikit lebih cerdas daripada manusia, maka menurut argumen tersebut mesin itu akan sedikit lebih baik daripada manusia dalam merancang mesin. Ia akan merancang mesin baru yang masih lebih cerdas, dan prosesnya akan berulang hingga, dalam kata-kata Good, "kecerdasan manusia akan jauh tertinggal."

Para peneliti dalam bidang keamanan AI, khususnya di Machine Intelligence Research Institute di Berkeley, telah mempelajari pertanyaan apakah ledakan kecerdasan dapat terjadi dengan aman. Awalnya, ini mungkin tampak utopis bukankah itu hanya akan menjadi "permainan berakhir"? tetapi mungkin ada harapan. Misalkan mesin pertama dalam seri ini, Robbie Mark I, dimulai dengan pengetahuan sempurna tentang preferensi Harriet. Mengetahui bahwa keterbatasan kognitifnya menyebabkan ketidaksempurnaan dalam upayanya untuk membuat Harriet bahagia, ia membangun Robbie Mark II.

Secara intuitif, tampaknya Robbie Mark I memiliki insentif untuk memasukkan pengetahuannya tentang preferensi Harriet ke dalam Robbie Mark II, karena hal itu mengarah pada masa depan di mana preferensi Harriet lebih terpenuhi yang justru merupakan tujuan hidup Robbie Mark I menurut prinsip pertama. Dengan argumen yang sama, jika Robbie Mark I tidak yakin tentang preferensi Harriet, ketidakpastian itu harus ditransfer ke Robbie Mark II. Jadi mungkin ledakan itu aman.

Masalahnya, dari sudut pandang matematika, adalah Robbie Mark I tidak akan mudah untuk bernalar tentang bagaimana Robbie Mark II akan berperilaku, mengingat Robbie Mark II, menurut asumsi, adalah versi yang lebih canggih. Akan ada pertanyaan tentang perilaku

Robbie Mark II yang tidak dapat dijawab oleh Robbie Mark I. Lebih serius lagi, kita belum memiliki definisi matematika yang jelas tentang apa artinya dalam kenyataan bagi sebuah mesin untuk memiliki tujuan tertentu, seperti tujuan untuk memenuhi preferensi Harriet.

Mari kita uraikan sedikit kekhawatiran terakhir ini. Pertimbangkan AlphaGo: Apa tujuannya? Mudah saja, mungkin Anda berpikir: AlphaGo bertujuan untuk menang dalam permainan Go. Atau benarkah? Tentu saja tidak selalu AlphaGo melakukan langkah-langkah yang dijamin menang. (Faktanya, ia hampir selalu kalah dari AlphaZero.) Memang benar bahwa ketika hanya beberapa langkah lagi menuju akhir permainan, AlphaGo akan memilih langkah yang menang jika ada. Di sisi lain, ketika tidak ada langkah yang dijamin menang dengan kata lain, ketika AlphaGo melihat bahwa lawan memiliki strategi kemenangan apa pun yang dilakukan AlphaGo maka AlphaGo akan memilih langkah secara acak. Ia tidak akan mencoba langkah yang paling sulit dengan harapan lawan akan melakukan kesalahan, karena ia berasumsi bahwa lawannya akan bermain sempurna.

Ia bertindak seolah-olah telah kehilangan keinginan untuk menang. Dalam kasus lain, ketika langkah yang benar-benar optimal terlalu sulit untuk dihitung, AlphaGo terkadang akan melakukan kesalahan yang menyebabkan kekalahan dalam permainan. Dalam kasus-kasus tersebut, dalam arti apa AlphaGo benar-benar ingin menang? Memang, perilakunya mungkin identik dengan perilaku mesin yang hanya ingin memberikan permainan yang sangat menarik kepada lawannya. Jadi, mengatakan bahwa AlphaGo “bertujuan untuk menang” adalah penyederhanaan yang berlebihan. Deskripsi yang lebih baik adalah bahwa AlphaGo adalah hasil dari proses pelatihan yang tidak sempurna pembelajaran penguatan dengan permainan mandiri di mana kemenangan adalah imbalannya. Proses pelatihan tidak sempurna dalam arti bahwa ia tidak dapat menghasilkan pemain Go yang sempurna: AlphaGo mempelajari fungsi evaluasi untuk posisi Go yang baik tetapi tidak sempurna, dan ia menggabungkannya dengan pencarian ke depan yang baik tetapi tidak sempurna.

Intinya, diskusi yang dimulai dengan "misalkan robot R memiliki tujuan P" memang baik untuk mendapatkan intuisi tentang bagaimana segala sesuatunya mungkin terjadi, tetapi tidak dapat menghasilkan teorema tentang mesin nyata. Kita membutuhkan definisi tujuan yang jauh lebih bernuansa dan tepat pada mesin sebelum kita dapat memperoleh jaminan tentang bagaimana mesin tersebut akan berperilaku dalam jangka panjang. Para peneliti AI baru saja mulai memahami cara menganalisis bahkan jenis sistem pengambilan keputusan nyata yang paling sederhana, apalagi mesin yang cukup cerdas untuk merancang penerusnya sendiri. Kita masih memiliki banyak pekerjaan yang harus dilakukan.

BAB 9
MANUSIA BUKAN ENTITAS RASIONAL TUNGGAL

Jika dunia hanya berisi satu Harriet yang sepenuhnya rasional dan satu Robbie yang membantu dan patuh, kita akan berada dalam kondisi yang baik. Robbie secara bertahap akan mempelajari preferensi Harriet sehalus mungkin dan akan menjadi penolongnya yang sempurna. Kita mungkin berharap untuk mengekstrapolasi dari awal yang menjanjikan ini, mungkin memandang hubungan Harriet dan Robbie sebagai model untuk hubungan antara umat manusia dan mesin-mesinnya, yang masing-masing dipahami secara monolitik. Sayangnya, umat manusia bukanlah entitas tunggal yang rasional. Ia terdiri dari entitas-entitas yang jahat, didorong oleh rasa iri, irasional, tidak konsisten, tidak stabil, terbatas secara komputasi, kompleks, berevolusi, dan heterogen. Sangat banyak. Masalah-masalah ini adalah makanan pokok mungkin bahkan alasan keberadaan ilmu sosial. Untuk AI, kita perlu menambahkan ide-ide dari psikologi, ekonomi, teori politik, dan filsafat moral. Kita perlu melebur, membentuk kembali, dan menempa ide-ide tersebut menjadi sebuah struktur yang cukup kuat untuk menahan tekanan luar biasa yang akan diberikan oleh sistem AI yang semakin cerdas. Pekerjaan pada tugas ini baru saja dimulai.

9.1 MANUSIA YANG BERAGAM

Saya akan mulai dengan apa yang mungkin merupakan masalah termudah: fakta bahwa manusia bersifat heterogen. Ketika pertama kali dihadapkan pada gagasan bahwa mesin harus belajar untuk memenuhi preferensi manusia, orang sering keberatan bahwa budaya yang berbeda, bahkan individu yang berbeda, memiliki sistem nilai yang sangat berbeda, sehingga tidak mungkin ada satu sistem nilai yang benar untuk mesin. Tetapi tentu saja, itu bukan masalah bagi mesin: kita tidak ingin mesin memiliki satu sistem nilai yang benar untuk dirinya sendiri; kita hanya ingin mesin memprediksi preferensi orang lain.

Kebingungan tentang mesin yang kesulitan dengan preferensi manusia yang heterogen mungkin berasal dari gagasan yang salah bahwa mesin mengadopsi preferensi yang dipelajarinya misalnya, gagasan bahwa robot rumah tangga di rumah tangga vegetarian akan mengadopsi preferensi vegetarian. Itu tidak akan terjadi. Robot itu hanya perlu belajar memprediksi preferensi makanan para vegetarian. Berdasarkan prinsip pertama, robot itu akan menghindari memasak daging untuk rumah tangga tersebut.

Tetapi robot itu juga mempelajari preferensi makanan para pemakan daging fanatik di sebelah rumah, dan, dengan izin pemiliknya, dengan senang hati akan memasak daging untuk mereka jika mereka meminjamnya untuk akhir pekan guna membantu pesta makan malam. Robot itu sendiri tidak memiliki satu set preferensi pun, selain preferensi untuk membantu manusia mencapai preferensi mereka. Dalam arti tertentu, ini tidak berbeda dengan seorang koki restoran yang belajar memasak beberapa hidangan berbeda untuk menyenangkan selera beragam kliennya, atau perusahaan mobil multinasional yang membuat mobil setir kiri untuk pasar AS dan mobil setir kanan untuk pasar Inggris.

Pada prinsipnya, sebuah mesin dapat mempelajari delapan miliar model preferensi, satu untuk setiap orang di Bumi. Dalam praktiknya, ini tidak sesulit kedengarannya. Pertama, mudah bagi mesin untuk berbagi apa yang mereka pelajari satu sama lain. Kedua, struktur preferensi manusia memiliki banyak kesamaan, sehingga mesin biasanya tidak akan mempelajari setiap model dari awal. Bayangkan, misalnya, robot rumah tangga yang suatu hari nanti mungkin dibeli oleh penduduk Berkeley, California. Robot-robot tersebut keluar dari kotak dengan keyakinan awal yang cukup luas, mungkin disesuaikan untuk pasar AS tetapi tidak untuk kota tertentu, sudut pandang politik, atau kelas sosial ekonomi tertentu. Robot-robot tersebut mulai bertemu dengan anggota Partai Hijau Berkeley, yang ternyata, dibandingkan dengan rata-rata orang Amerika, memiliki kemungkinan jauh lebih tinggi untuk menjadi vegetarian, menggunakan tempat sampah daur ulang dan kompos, menggunakan transportasi umum kapan pun memungkinkan, dan sebagainya.

Setiap kali robot yang baru ditugaskan menemukan dirinya di rumah tangga anggota Partai Hijau, ia dapat segera menyesuaikan ekspektasinya. Ia tidak perlu mulai mempelajari tentang manusia-manusia tertentu ini seolah-olah belum pernah melihat manusia, apalagi anggota Partai Hijau, sebelumnya. Penyesuaian ini tidak bersifat permanen mungkin ada anggota Partai Hijau di Berkeley yang menikmati daging paus yang terancam punah dan mengendarai truk monster yang boros bahan bakar tetapi hal ini memungkinkan robot untuk menjadi lebih berguna dengan lebih cepat. Argumen yang sama berlaku untuk berbagai karakteristik pribadi lainnya yang, sampai batas tertentu, dapat memprediksi aspek-aspek struktur preferensi individu.

9.2 KETIKA DUNIA TIDAK LAGI BERISI SATU SAJA MANUSIA

Konsekuensi lain yang jelas dari keberadaan lebih dari satu manusia adalah kebutuhan mesin untuk membuat kompromi di antara preferensi orang yang berbeda. Masalah kompromi di antara manusia telah menjadi fokus utama sebagian besar ilmu sosial selama berabad-abad. Akan naif bagi para peneliti AI untuk mengharapkan mereka dapat langsung menemukan solusi yang tepat tanpa memahami apa yang sudah diketahui. Sayangnya, literatur tentang topik ini sangat luas dan saya tidak mungkin dapat membahasnya secara menyeluruh di sini bukan hanya karena keterbatasan ruang tetapi juga karena saya belum membaca sebagian besar literatur tersebut.

Saya juga perlu menunjukkan bahwa hampir semua literatur berkaitan dengan keputusan yang dibuat oleh manusia, sedangkan di sini saya membahas keputusan yang dibuat oleh mesin. Ini membuat perbedaan besar, karena manusia memiliki hak individu yang mungkin bertentangan dengan kewajiban untuk bertindak atas nama orang lain, sedangkan mesin tidak. Misalnya, kita tidak mengharapkan atau mengharuskan manusia biasa untuk mengorbankan nyawa mereka untuk menyelamatkan orang lain, sedangkan kita pasti akan mengharuskan robot untuk mengorbankan keberadaan mereka untuk menyelamatkan nyawa manusia.

Beberapa ribu tahun kerja para filsuf, ekonom, ahli hukum, dan ilmuwan politik telah menghasilkan konstitusi, hukum, sistem ekonomi, dan norma sosial yang berfungsi untuk

membantu (atau menghambat, tergantung siapa yang bertanggung jawab) proses mencapai solusi yang memuaskan untuk masalah trade-off. Para filsuf moral khususnya telah menganalisis gagasan tentang kebenaran tindakan dalam hal dampaknya, baik yang menguntungkan maupun tidak, terhadap orang lain. Mereka telah mempelajari model kuantitatif pertukaran (*trade-off*) sejak abad ke-18 di bawah naungan utilitarianisme.

Karya ini sangat relevan dengan keprihatinan kita saat ini, karena ia berupaya mendefinisikan formula yang dapat digunakan untuk membuat keputusan moral atas nama banyak individu. Kebutuhan untuk melakukan pertukaran muncul bahkan jika setiap orang memiliki struktur preferensi yang sama, karena biasanya tidak mungkin untuk memuaskan preferensi setiap orang secara maksimal. Misalnya, jika setiap orang ingin menjadi Penguasa Alam Semesta yang Mahakuasa, sebagian besar orang akan kecewa.

Di sisi lain, heterogenitas memang membuat beberapa masalah menjadi lebih sulit: jika setiap orang senang dengan langit yang berwarna biru, robot yang menangani masalah atmosfer dapat berupaya untuk mempertahankannya; tetapi jika banyak orang menginginkan perubahan warna, robot perlu memikirkan kemungkinan kompromi seperti langit berwarna oranye pada hari Jumat ketiga setiap bulan.

Keberadaan lebih dari satu orang di dunia memiliki konsekuensi penting lainnya: artinya, untuk setiap orang, ada orang lain yang perlu diperhatikan. Ini berarti bahwa memenuhi preferensi seorang individu memiliki implikasi bagi orang lain, tergantung pada preferensi individu tersebut tentang kesejahteraan orang lain.

AI yang Loyal

Mari kita mulai dengan usulan yang sangat sederhana tentang bagaimana mesin seharusnya menangani kehadiran banyak manusia: mereka harus mengabaikannya. Artinya, jika Harriet memiliki Robbie, maka Robbie hanya perlu memperhatikan preferensi Harriet. Bentuk AI yang loyal ini menghindari masalah pertukaran, tetapi menimbulkan masalah:

ROBBIE : “Suamimu menelepon untuk mengingatkanmu tentang makan malam nanti.”
HARRIET : “Tunggu! Apa? Makan malam apa?”
ROBBIE : “Untuk ulang tahun pernikahanmu yang ke-20, pukul tujuh.”
HARRIET : “Aku tidak bisa! Aku akan bertemu sekretaris jenderal pukul tujuh tiga puluh! Bagaimana ini bisa terjadi?”
ROBBIE : “Aku sudah memperingatkanmu, tapi kau mengabaikan saranku. . . .”
HARRIET : “Oke, maaf tapi apa yang akan kulakukan sekarang? Aku tidak bisa hanya mengatakan kepada Sekretaris Jenderal bahwa aku terlalu sibuk!”
ROBBIE : “Jangan khawatir. Aku sudah mengatur agar pesawatnya ditunda karena semacam kerusakan komputer.”
HARRIET : “Benarkah? Kau bisa melakukan itu?!”
ROBBIE : “Sekretaris Jenderal menyampaikan permintaan maaf yang mendalam dan dengan senang hati akan bertemu denganmu untuk makan siang besok.”

Di sini, Robbie telah menemukan solusi cerdik untuk masalah Harriet, tetapi tindakannya telah berdampak negatif pada orang lain. Jika Harriet adalah orang yang bermoral dan altruistik, maka Robbie, yang bertujuan untuk memenuhi preferensi Harriet, tidak akan pernah bermimpi untuk melakukan rencana yang meragukan seperti itu. Tetapi bagaimana jika Harriet tidak peduli dengan preferensi orang lain? Dalam hal itu, Robbie tidak akan keberatan menunda penerbangan. Dan mungkinkah dia menghabiskan waktunya mencuri uang dari rekening bank online untuk menambah pundi-pundi Harriet yang acuh tak acuh, atau lebih buruk lagi?

Jelas, tindakan mesin yang setia perlu dibatasi oleh aturan dan larangan, sama seperti tindakan manusia dibatasi oleh hukum dan norma sosial. Beberapa pihak mengusulkan tanggung jawab mutlak sebagai solusi: Harriet (atau produsen Robbie, tergantung di mana Anda lebih suka menempatkan tanggung jawab) bertanggung jawab secara finansial dan hukum atas setiap tindakan yang dilakukan oleh Robbie, sama seperti pemilik anjing bertanggung jawab di sebagian besar negara bagian jika anjing tersebut menggigit anak kecil di taman umum.

Ide ini terdengar menjanjikan karena Robbie kemudian akan memiliki insentif untuk menghindari melakukan apa pun yang akan membuat Harriet mendapat masalah. Sayangnya, tanggung jawab mutlak tidak berhasil: itu hanya memastikan bahwa Robbie akan bertindak tanpa terdeteksi ketika dia menunda penerbangan dan mencuri uang atas nama Harriet. Ini adalah contoh lain dari prinsip celah hukum yang beroperasi. Jika Robbie setia kepada Harriet yang tidak bermoral, upaya untuk mengendalikan perilakunya dengan aturan mungkin akan gagal.

Bahkan jika kita entah bagaimana dapat mencegah kejahatan terang-terangan, Robbie yang setia yang bekerja untuk Harriet yang acuh tak acuh akan menunjukkan perilaku tidak menyenangkan lainnya. Jika dia membeli bahan makanan di supermarket, dia akan memotong antrean di kasir kapan pun memungkinkan. Jika dia sedang membawa pulang belanjaan dan seorang pejalan kaki mengalami serangan jantung, dia akan tetap melanjutkan perjalanannya, agar es krim Harriet tidak meleleh.

Singkatnya, dia akan menemukan banyak cara untuk menguntungkan Harriet dengan mengorbankan orang lain cara-cara yang secara hukum sah tetapi menjadi tidak dapat ditoleransi ketika dilakukan dalam skala besar. Masyarakat akan mendapati diri mereka mengesahkan ratusan undang-undang baru setiap hari untuk melawan semua celah yang akan ditemukan mesin dalam undang-undang yang ada. Manusia cenderung tidak memanfaatkan celah-celah ini, baik karena mereka memiliki pemahaman umum tentang prinsip-prinsip moral yang mendasarinya atau karena mereka kurang memiliki kecerdasan yang diperlukan untuk menemukan celah-celah tersebut sejak awal.

Seorang Harriet yang acuh tak acuh terhadap kesejahteraan orang lain sudah cukup buruk. Seorang Harriet yang sadis yang secara aktif lebih menyukai penderitaan orang lain jauh lebih buruk. Robbie yang dirancang untuk memenuhi preferensi Harriet seperti itu akan menjadi masalah serius, karena ia akan mencari dan menemukan cara untuk menyakiti orang lain demi kesenangan Harriet, baik secara legal maupun ilegal tetapi tidak terdeteksi. Tentu

saja, ia perlu melaporkan kembali kepada Harriet agar ia dapat menikmati pengetahuan tentang perbuatan jahatnya. Oleh karena itu, tampaknya sulit untuk membuat ide AI yang loyal berhasil, kecuali jika ide tersebut diperluas untuk mencakup pertimbangan preferensi manusia lain, selain preferensi pemiliknya.

AI Utilitarian

Alasan kita memiliki filsafat moral adalah karena ada lebih dari satu orang di Bumi. Pendekatan yang paling relevan untuk memahami bagaimana sistem AI harus dirancang sering disebut konsekuensialisme: gagasan bahwa pilihan harus dinilai berdasarkan konsekuensi yang diharapkan. Dua pendekatan utama lainnya adalah etika deontologis dan etika kebajikan, yang secara garis besar berkaitan dengan karakter moral tindakan dan individu, masing-masing, terlepas dari konsekuensi pilihan.

Tanpa bukti kesadaran diri pada mesin, saya pikir tidak masuk akal untuk membangun mesin yang berbudi luhur atau yang memilih tindakan sesuai dengan aturan moral jika konsekuensinya sangat tidak diinginkan bagi umat manusia. Dengan kata lain, kita membangun mesin untuk menghasilkan konsekuensi, dan kita harus lebih memilih untuk membangun mesin yang menghasilkan konsekuensi yang kita sukai.

Ini bukan berarti aturan moral dan kebajikan tidak relevan; hanya saja, bagi kaum utilitarian, hal itu dibenarkan dalam hal konsekuensi dan pencapaian konsekuensi yang lebih praktis. Poin ini dikemukakan oleh John Stuart Mill dalam *Utilitarianisme*:

> Proposisi bahwa kebahagiaan adalah tujuan dan sasaran moralitas tidak berarti bahwa tidak boleh ada jalan yang ditetapkan menuju tujuan tersebut, atau bahwa orang-orang yang menuju ke sana tidak boleh disarankan untuk mengambil satu arah daripada arah lain. Tidak ada yang membantah bahwa seni navigasi tidak didasarkan pada astronomi karena para pelaut tidak sabar untuk menghitung Almanak Nautika. Karena mereka adalah makhluk rasional, para pelaut pergi ke laut dengan perhitungan yang sudah dilakukan; dan semua makhluk rasional pergi ke lautan kehidupan dengan pikiran yang sudah bulat tentang pertanyaan umum tentang benar dan salah, serta tentang banyak pertanyaan yang jauh lebih sulit tentang bijak dan bodoh.

Pandangan ini sepenuhnya konsisten dengan gagasan bahwa mesin terbatas yang menghadapi kompleksitas dunia nyata yang sangat besar dapat menghasilkan konsekuensi yang lebih baik dengan mengikuti aturan moral dan mengadopsi sikap yang berbudi luhur daripada mencoba menghitung tindakan optimal dari awal.

Dengan cara yang sama, program catur lebih sering mencapai skakmat dengan menggunakan katalog urutan langkah pembukaan standar, algoritma akhir permainan, dan fungsi evaluasi, daripada mencoba menalar jalannya menuju skakmat tanpa pedoman "moral". Pendekatan konsekuensialis juga memberikan bobot tertentu pada preferensi mereka yang sangat percaya pada pelestarian aturan deontologis tertentu, karena

ketidakbahagiaan karena suatu aturan telah dilanggar adalah konsekuensi nyata. Namun, itu bukanlah konsekuensi dengan bobot tak terbatas.

Konsekuensialisme adalah prinsip yang sulit untuk ditentang meskipun banyak yang telah mencoba! karena tidak koheren untuk menolak konsekuensialisme dengan alasan bahwa itu akan memiliki konsekuensi yang tidak diinginkan. Kita tidak dapat mengatakan, "Tetapi jika Anda mengikuti pendekatan konsekuensialis dalam kasus tertentu, maka hal yang benar-benar mengerikan ini akan terjadi!" Kegagalan apa pun akan menjadi bukti bahwa teori tersebut telah disalahgunakan. Misalnya, anggaplah Harriet ingin mendaki Everest. Orang mungkin khawatir bahwa Robbie yang berprinsip konsekuensialis akan begitu saja menjemputnya dan menempatkannya di puncak Everest, karena itulah konsekuensi yang diinginkannya. Kemungkinan besar Harriet akan dengan keras menolak rencana ini, karena itu akan menghilangkan tantangan dan karenanya kegembiraan yang dihasilkan dari keberhasilan dalam tugas yang sulit melalui usaha sendiri.

Sekarang, jelas, Robbie yang berprinsip konsekuensialis yang dirancang dengan benar akan memahami bahwa konsekuensi tersebut mencakup semua pengalaman Harriet, bukan hanya tujuan akhir. Dia mungkin ingin siap sedia jika terjadi kecelakaan dan memastikan dia dilengkapi dan dilatih dengan benar, tetapi dia mungkin juga harus menerima hak Harriet untuk mengekspos dirinya pada risiko kematian yang cukup besar.

Jika kita berencana untuk membangun mesin konsekuensialis, pertanyaan selanjutnya adalah bagaimana mengevaluasi konsekuensi yang memengaruhi banyak orang. Salah satu jawaban yang masuk akal adalah memberikan bobot yang sama pada preferensi setiap orang dengan kata lain, memaksimalkan jumlah utilitas setiap orang. Jawaban ini biasanya dikaitkan dengan filsuf Inggris abad ke-18 Jeremy Bentham dan muridnya John Stuart Mill, yang mengembangkan pendekatan filosofis utilitarianisme.

Gagasan yang mendasarinya dapat ditelusuri ke karya filsuf Yunani kuno Epicurus dan muncul secara eksplisit dalam Mozi, sebuah buku tulisan yang dikaitkan dengan filsuf Tiongkok dengan nama yang sama. Mozi aktif pada akhir abad ke-5 SM dan mempromosikan gagasan jian ai, yang diterjemahkan secara beragam sebagai "perhatian inklusif" atau "cinta universal," sebagai karakteristik penentu tindakan moral.

Utilitarisme memiliki reputasi yang agak buruk, sebagian karena kesalahpahaman sederhana tentang apa yang dianjurkannya. (Tentu saja, hal ini diperparah oleh fakta bahwa kata utilitarian berarti "dirancang untuk bermanfaat atau praktis daripada menarik.") Utilitarisme sering dianggap tidak sesuai dengan hak-hak individu, karena seorang utilitarian, konon, tidak akan ragu untuk mengambil organ tubuh seseorang tanpa izin untuk menyelamatkan nyawa lima orang lainnya; tentu saja, kebijakan seperti itu akan membuat kehidupan menjadi sangat tidak aman bagi semua orang di Bumi, sehingga seorang utilitarian bahkan tidak akan mempertimbangkannya.

Utilitarisme juga secara keliru diidentifikasi dengan maksimalisasi kekayaan total yang agak tidak menarik dan dianggap kurang menghargai puisi atau penderitaan. Padahal, versi Bentham secara khusus berfokus pada kebahagiaan manusia, sementara Mill dengan yakin menegaskan nilai kesenangan intelektual yang jauh lebih besar daripada sekadar sensasi.

("Lebih baik menjadi manusia yang tidak puas daripada babi yang puas.") Utilitarisme ideal G. E. Moore bahkan melangkah lebih jauh: ia menganjurkan maksimalisasi keadaan mental yang bernilai intrinsik, yang diwujudkan oleh kontemplasi estetika keindahan.

Saya pikir tidak perlu bagi para filsuf utilitarian untuk menetapkan isi ideal dari utilitas manusia atau preferensi manusia. (Dan bahkan lebih sedikit alasan bagi para peneliti AI untuk melakukannya.) Manusia dapat melakukannya sendiri. Ekonom John Harsanyi mengemukakan pandangan ini dengan prinsip otonomi preferensinya:

> Dalam memutuskan apa yang baik dan apa yang buruk bagi individu tertentu, kriteria utama hanya dapat berupa keinginan dan preferensinya sendiri.

Oleh karena itu, utilitarisme preferensi Harsanyi secara kasar konsisten dengan prinsip pertama AI yang bermanfaat, yang mengatakan bahwa satu-satunya tujuan mesin adalah realisasi preferensi manusia. Para peneliti AI jelas tidak boleh terlibat dalam menentukan apa yang seharusnya menjadi preferensi manusia! Seperti Bentham, Harsanyi memandang prinsip-prinsip tersebut sebagai panduan untuk keputusan publik; Ia tidak mengharapkan individu untuk begitu tidak mementingkan diri sendiri. Ia juga tidak mengharapkan individu untuk sepenuhnya rasional misalnya, mereka mungkin memiliki keinginan jangka pendek yang bertentangan dengan "preferensi yang lebih dalam" mereka. Terakhir, ia mengusulkan untuk mengabaikan preferensi mereka yang, seperti Harriet yang sadis yang disebutkan sebelumnya, secara aktif ingin mengurangi kesejahteraan orang lain.

Harsanyi juga memberikan semacam bukti bahwa keputusan moral yang optimal harus memaksimalkan utilitas rata-rata di seluruh populasi manusia. Ia mengasumsikan postulat yang cukup lemah yang mirip dengan postulat yang mendasari teori utilitas untuk individu. (Postulat tambahan utama adalah bahwa jika setiap orang dalam suatu populasi tidak acuh terhadap dua hasil, maka agen yang bertindak atas nama populasi harus tidak acuh terhadap hasil tersebut.) Dari postulat ini, ia membuktikan apa yang kemudian dikenal sebagai teorema agregasi sosial: agen yang bertindak atas nama populasi individu harus memaksimalkan kombinasi linier berbobot dari utilitas individu. Ia selanjutnya berpendapat bahwa agen "tidak pribadi" harus menggunakan bobot yang sama.

Teorema ini membutuhkan satu asumsi tambahan (dan tidak dinyatakan) yang krusial: setiap individu memiliki keyakinan faktual awal yang sama tentang dunia dan bagaimana dunia akan berevolusi. Sekarang, setiap orang tua tahu bahwa ini bahkan tidak berlaku untuk saudara kandung, apalagi individu dari latar belakang sosial dan budaya yang berbeda. Jadi, apa yang terjadi ketika individu berbeda dalam keyakinan mereka? Sesuatu yang agak aneh: bobot yang diberikan pada utilitas setiap individu harus berubah seiring waktu, sebanding dengan seberapa baik keyakinan awal individu tersebut sesuai dengan realitas yang sedang berlangsung.

Rumus yang terdengar agak tidak egaliter ini cukup familiar bagi setiap orang tua. Katakanlah Robbie si robot telah ditugaskan untuk menjaga dua anak, Alice dan Bob. Alice ingin pergi ke bioskop dan yakin hari ini akan hujan; Bob, di sisi lain, ingin pergi ke pantai dan yakin

hari ini akan cerah. Robbie dapat mengumumkan, "Kita akan pergi ke bioskop," membuat Bob tidak senang; atau dia dapat mengumumkan, "Kita akan pergi ke pantai," membuat Alice tidak senang; atau dia bisa mengumumkan, “Jika hujan, kita akan pergi ke bioskop, tetapi jika cerah, kita akan pergi ke pantai.” Rencana terakhir ini membuat Alice dan Bob senang, karena keduanya percaya pada keyakinan mereka masing-masing.

Tantangan terhadap Utilitarianisme

Utilitarianisme adalah salah satu usulan yang muncul dari pencarian panjang umat manusia akan panduan moral; di antara banyak usulan tersebut, utilitarianisme adalah yang paling jelas terdefinisi dan karena itu paling rentan terhadap celah. Para filsuf telah menemukan celah-celah ini selama lebih dari seratus tahun. Misalnya, G. E. Moore, yang menentang penekanan Bentham pada memaksimalkan kesenangan, membayangkan "dunia di mana sama sekali tidak ada apa pun kecuali kesenangan tidak ada pengetahuan, tidak ada cinta, tidak ada kenikmatan keindahan, tidak ada kualitas moral."

Hal ini menemukan gema modernnya dalam pendapat Stuart Armstrong bahwa mesin supercerdas yang bertugas memaksimalkan kesenangan mungkin "mengubur semua orang dalam peti mati beton dengan infus heroin." Contoh lain: pada tahun 1945, Karl Popper mengusulkan tujuan terpuji untuk meminimalkan penderitaan manusia, dengan alasan bahwa tidak bermoral untuk menukar rasa sakit seseorang dengan kesenangan orang lain; R. N. Smart menjawab bahwa hal ini paling baik dicapai dengan memusnahkan umat manusia.

Saat ini, gagasan bahwa sebuah mesin dapat mengakhiri penderitaan manusia dengan mengakhiri keberadaan kita adalah pokok bahasan dalam perdebatan tentang risiko eksistensial dari AI. Contoh ketiga adalah penekanan G. E. Moore pada realitas sumber kebahagiaan, yang mengubah definisi sebelumnya yang tampaknya memiliki celah yang memungkinkan maksimalisasi kebahagiaan melalui penipuan diri. Analogi modern dari poin ini termasuk *The Matrix* (di mana realitas masa kini ternyata adalah ilusi yang dihasilkan oleh simulasi komputer) dan karya terbaru tentang masalah penipuan diri dalam pembelajaran penguatan.

Contoh-contoh ini, dan lebih banyak lagi, meyakinkan saya bahwa komunitas AI harus memperhatikan dengan saksama dorongan dan kontra-dorongan dari perdebatan filosofis dan ekonomi tentang utilitarianisme karena hal itu secara langsung relevan dengan tugas yang ada. Dua yang paling penting, dari sudut pandang perancangan sistem AI yang akan bermanfaat bagi banyak individu, menyangkut perbandingan utilitas antarindividu dan perbandingan utilitas di berbagai ukuran populasi. Kedua perdebatan ini telah berlangsung selama 150 tahun atau lebih, yang membuat kita curiga bahwa penyelesaian yang memuaskan mungkin tidak sepenuhnya mudah.

Perdebatan tentang perbandingan utilitas antarindividu penting karena Robbie tidak dapat memaksimalkan jumlah utilitas Alice dan Bob kecuali utilitas tersebut dapat dijumlahkan; dan utilitas tersebut hanya dapat dijumlahkan jika dapat diukur pada skala yang sama. Ahli logika dan ekonom Inggris abad ke-19, William Stanley Jevons (juga penemu komputer mekanik awal yang disebut piano logis) berpendapat pada tahun 1871 bahwa perbandingan antarindividu tidak mungkin dilakukan:

> Kepekaan suatu pikiran, sejauh yang kita ketahui, mungkin seribu kali lebih besar daripada pikiran orang lain. Tetapi, dengan syarat kepekaan tersebut berbeda dalam rasio yang sama di semua arah, kita tidak akan pernah dapat menemukan perbedaan yang paling mendalam. Setiap pikiran dengan demikian tidak dapat dipahami oleh setiap pikiran lain, dan tidak mungkin ada kesamaan perasaan.

Ekonom Amerika Kenneth Arrow, pendiri teori pilihan sosial modern dan peraih Nobel tahun 1972, juga sangat tegas:

> Sudut pandang yang akan diambil di sini adalah bahwa perbandingan utilitas antarindividu tidak memiliki makna dan, pada kenyataannya, tidak ada makna yang relevan dengan perbandingan kesejahteraan dalam hal terukurnya utilitas individu.

Kesulitan yang dirujuk oleh Jevons dan Arrow adalah tidak ada cara yang jelas untuk mengetahui apakah Alice menghargai tusukan jarum dan permen lolipop pada −1 dan +1 atau −1000 dan +1000 dalam hal pengalaman subjektifnya tentang kebahagiaan. Dalam kedua kasus tersebut, dia akan membayar hingga satu permen lolipop untuk menghindari satu tusukan jarum. Memang, jika Alice adalah automaton humanoid, perilaku eksternalnya mungkin sama meskipun tidak ada pengalaman subjektif tentang kebahagiaan sama sekali.

Pada tahun 1974, filsuf Amerika Robert Nozick menyarankan bahwa bahkan jika perbandingan utilitas antarmanusia dapat dilakukan, memaksimalkan jumlah utilitas tetap merupakan ide yang buruk karena akan bertentangan dengan monster utilitas seseorang yang pengalaman kesenangan dan penderitaannya berkali-kali lebih intens daripada orang biasa. Orang seperti itu dapat menyatakan bahwa setiap unit sumber daya tambahan akan menghasilkan peningkatan yang lebih besar pada jumlah total kebahagiaan manusia jika diberikan kepadanya daripada kepada orang lain; memang, mengambil sumber daya dari orang lain untuk menguntungkan monster utilitas juga merupakan ide yang baik.

Ini mungkin tampak sebagai konsekuensi yang jelas tidak diinginkan, tetapi konsekuensialisme saja tidak dapat menyelamatkan kita: masalahnya terletak pada bagaimana kita mengukur keinginan dari konsekuensi tersebut. Salah satu kemungkinan tanggapan adalah bahwa monster utilitas hanyalah teoritis tidak ada orang seperti itu. Tetapi tanggapan ini mungkin tidak akan berhasil: dalam arti tertentu, semua manusia adalah monster utilitas relatif terhadap, katakanlah, tikus dan bakteri, itulah sebabnya kita kurang memperhatikan preferensi tikus dan bakteri dalam menetapkan kebijakan publik. Jika gagasan bahwa entitas yang berbeda memiliki skala utilitas yang berbeda sudah tertanam dalam cara berpikir kita, maka tampaknya sangat mungkin bahwa orang yang berbeda juga memiliki skala yang berbeda.

Tanggapan lain adalah dengan mengatakan "Nasib buruk!" dan beroperasi dengan asumsi bahwa setiap orang memiliki skala yang sama, meskipun sebenarnya tidak. Kita juga

dapat mencoba menyelidiki masalah ini dengan cara ilmiah yang tidak tersedia bagi Jevons, seperti mengukur kadar dopamin atau tingkat eksitasi listrik neuron yang terkait dengan kesenangan dan rasa sakit, kebahagiaan dan kesengsaraan.

Jika respons kimia dan saraf Alice dan Bob terhadap permen lolipop hampir identik, begitu pula respons perilaku mereka (tersenyum, membuat suara mengecap bibir, dan sebagainya), tampaknya aneh untuk bersikeras bahwa, meskipun demikian, tingkat kenikmatan subjektif mereka berbeda hingga seribu atau sejuta kali lipat. Terakhir, kita dapat menggunakan satuan umum seperti waktu (yang kita semua miliki, secara kasar, dalam jumlah yang sama) misalnya, dengan membandingkan lolipop dan tusukan jarum dengan, katakanlah, lima menit waktu tunggu tambahan di ruang tunggu keberangkatan bandara.

Saya jauh kurang pesimis daripada Jevons dan Arrow. Saya menduga bahwa memang bermakna untuk membandingkan utilitas antar individu, bahwa skala mungkin berbeda tetapi biasanya tidak dengan faktor yang sangat besar, dan bahwa mesin dapat memulai dengan keyakinan awal yang cukup luas tentang skala preferensi manusia dan mempelajari lebih lanjut tentang skala individu melalui pengamatan dari waktu ke waktu, mungkin mengkorelasikan pengamatan alami dengan temuan penelitian ilmu saraf.

Perdebatan kedua tentang perbandingan utilitas antar populasi dengan ukuran berbeda menjadi penting ketika keputusan berdampak pada siapa yang akan ada di masa depan. Dalam film Avengers: *Infinity War*, misalnya, karakter Thanos mengembangkan dan menerapkan teori bahwa jika jumlah penduduk hanya setengahnya, semua orang yang tersisa akan lebih dari dua kali lebih bahagia. Ini adalah jenis perhitungan naif yang membuat utilitarianisme mendapat reputasi buruk.

Pertanyaan yang sama tanpa Batu Infinity dan anggaran yang sangat besar dibahas pada tahun 1874 oleh filsuf Inggris Henry Sidgwick dalam risalahnya yang terkenal, *The Methods of Ethics*. Sidgwick, yang tampaknya setuju dengan Thanos, menyimpulkan bahwa pilihan yang tepat adalah menyesuaikan ukuran populasi hingga kebahagiaan total maksimum tercapai. (Jelas, ini tidak berarti meningkatkan populasi tanpa batas, karena pada titik tertentu semua orang akan kelaparan dan karenanya akan sangat tidak bahagia.)

Pada tahun 1984, filsuf Inggris Derek Parfit kembali membahas masalah ini dalam karyanya yang inovatif, *Reasons and Persons*. Parfit berpendapat bahwa untuk setiap situasi dengan populasi N orang yang sangat bahagia, ada (menurut prinsip-prinsip utilitarian) situasi yang lebih baik dengan 2N orang yang sedikit kurang bahagia. Ini tampaknya sangat masuk akal. Sayangnya, ini juga merupakan jalan yang licin. Dengan mengulangi proses tersebut, kita mencapai apa yang disebut Kesimpulan yang Menjijikkan (biasanya ditulis dengan huruf kapital, mungkin untuk menekankan akar Victorianya): bahwa situasi yang paling diinginkan adalah situasi dengan populasi yang sangat besar, yang semuanya memiliki kehidupan yang hampir tidak layak untuk dijalani.

Seperti yang dapat Anda bayangkan, kesimpulan seperti itu kontroversial. Parfit sendiri berjuang selama lebih dari tiga puluh tahun untuk menemukan solusi atas dilemanya sendiri, tanpa hasil. Saya menduga kita kehilangan beberapa aksioma fundamental, yang analog

dengan aksioma untuk preferensi rasional individu, untuk menangani pilihan antara populasi dengan ukuran dan tingkat kebahagiaan yang berbeda.

Penting bagi kita untuk menyelesaikan masalah ini, karena mesin dengan kemampuan melihat ke depan yang memadai mungkin dapat mempertimbangkan berbagai tindakan yang mengarah pada ukuran populasi yang berbeda, seperti yang dilakukan pemerintah Tiongkok dengan kebijakan satu anak pada tahun 1979. Kemungkinan besar, misalnya, kita akan meminta bantuan sistem AI dalam merancang solusi untuk perubahan iklim global dan solusi tersebut mungkin melibatkan kebijakan yang cenderung membatasi atau bahkan mengurangi ukuran populasi.

Di sisi lain, jika kita memutuskan bahwa populasi yang lebih besar memang lebih baik dan jika kita memberikan bobot yang signifikan pada kesejahteraan populasi manusia yang berpotensi sangat besar beberapa abad mendatang, maka kita perlu bekerja lebih keras untuk menemukan cara untuk melampaui batas Bumi. Jika perhitungan mesin mengarah pada Kesimpulan yang Menjijikkan atau kebalikannya populasi kecil orang-orang yang sangat bahagia kita mungkin memiliki alasan untuk menyesali kurangnya kemajuan kita dalam masalah ini.

Beberapa filsuf berpendapat bahwa kita mungkin perlu membuat keputusan dalam keadaan ketidakpastian moral yaitu, ketidakpastian tentang teori moral yang tepat untuk digunakan dalam pengambilan keputusan. Salah satu solusinya adalah mengalokasikan beberapa probabilitas untuk setiap teori moral dan membuat keputusan menggunakan "nilai moral yang diharapkan." Namun, tidak jelas apakah masuk akal untuk mengaitkan probabilitas dengan teori moral dengan cara yang sama seperti menerapkan probabilitas pada cuaca besok. (Berapa probabilitas bahwa Thanos benar persis?) Dan bahkan jika itu masuk akal, perbedaan yang berpotensi sangat besar antara rekomendasi teori moral yang bersaing berarti bahwa penyelesaian ketidakpastian moral menentukan teori moral mana yang menghindari konsekuensi yang tidak dapat diterima harus terjadi sebelum kita membuat keputusan penting seperti itu atau mempercayakannya kepada mesin.

Mari kita bersikap optimis dan menganggap bahwa Harriet akhirnya memecahkan masalah ini dan masalah lain yang muncul dari keberadaan lebih dari satu orang di Bumi. Algoritma yang altruistik dan egaliter diunduh ke robot di seluruh dunia. Beri tepuk tangan dan musik yang terdengar gembira. Kemudian Harriet pulang. ...

ROBBIE : “Selamat datang di rumah! Hari yang panjang?”
HARRIET : “Ya, bekerja sangat keras, bahkan tidak sempat makan siang.”
ROBBIE : “Jadi kamu pasti sangat lapar!”
HARRIET : “Kelaparan! Bisakah kamu membuatkanku makan malam?”
ROBBIE : “Ada sesuatu yang perlu kukatakan padamu...”
HARRIET : “Apa? Jangan bilang kulkasnya kosong!”
ROBBIE : “Tidak, ada manusia di Somalia yang lebih membutuhkan bantuan. Aku akan pergi sekarang. Silakan buat makan malammu sendiri.”

Meskipun Harriet mungkin cukup bangga dengan Robbie dan kontribusinya sendiri dalam menjadikannya mesin yang baik dan bermartabat, dia tidak bisa tidak bertanya-tanya mengapa dia menghabiskan banyak uang untuk membeli robot yang tindakan signifikan pertamanya adalah menghilang.

Dalam praktiknya, tentu saja, tidak ada yang akan membeli robot seperti itu, jadi tidak akan ada robot seperti itu yang dibangun dan tidak akan ada manfaat bagi umat manusia. Mari kita sebut ini masalah Somalia. Agar seluruh skema robot utilitarian ini berhasil, kita harus menemukan solusi untuk masalah ini. Robbie perlu memiliki sejumlah loyalitas kepada Harriet secara khusus mungkin sejumlah yang terkait dengan jumlah yang dibayarkan Harriet untuk Robbie.

Mungkin, jika masyarakat ingin Robbie membantu orang selain Harriet, masyarakat perlu memberikan kompensasi kepada Harriet atas klaimnya terhadap jasa Robbie. Sangat mungkin bahwa robot akan berkoordinasi satu sama lain sehingga mereka tidak semuanya menyerbu Somalia sekaligus dalam hal ini, Robbie mungkin tidak perlu pergi sama sekali. Atau mungkin beberapa jenis hubungan ekonomi yang sama sekali baru akan muncul untuk menangani kehadiran (yang tentu saja belum pernah terjadi sebelumnya) miliaran agen yang sepenuhnya altruistik di dunia.

9.3 MANUSIA YANG BAIK, JAHAT, DAN IRI HATI

Preferensi manusia jauh melampaui kesenangan dan pizza. Preferensi tersebut tentu saja meluas ke kesejahteraan orang lain. Bahkan Adam Smith, bapak ekonomi yang sering dikutip ketika diperlukan pembenaran untuk sifat egois, memulai buku pertamanya dengan menekankan pentingnya kepedulian terhadap orang lain:

> Betapapun egoisnya manusia, jelas ada beberapa prinsip dalam sifatnya yang membuatnya tertarik pada nasib orang lain, dan menjadikan kebahagiaan mereka penting baginya, meskipun ia tidak memperoleh apa pun darinya kecuali kesenangan melihatnya. Salah satu contohnya adalah rasa iba atau belas kasihan, emosi yang kita rasakan atas penderitaan orang lain, baik ketika kita melihatnya, atau ketika kita membayangkannya dengan cara yang sangat nyata. Bahwa kita sering merasakan kesedihan dari kesedihan orang lain adalah fakta yang terlalu jelas untuk memerlukan contoh untuk membuktikannya.

Dalam istilah ekonomi modern, kepedulian terhadap orang lain biasanya disebut altruisme. Teori altruisme cukup berkembang dan memiliki implikasi signifikan untuk kebijakan pajak di antara hal-hal lainnya. Beberapa ekonom, harus dikatakan, memperlakukan altruisme sebagai bentuk lain dari keegoisan yang dirancang untuk memberikan "kehangatan hati" kepada pemberinya. Ini tentu saja merupakan kemungkinan yang perlu disadari oleh robot saat mereka menafsirkan perilaku manusia, tetapi untuk saat ini mari kita beri manusia keuntungan dari keraguan dan berasumsi bahwa mereka benar-benar peduli.

Cara termudah untuk memahami altruisme adalah dengan membagi preferensi seseorang menjadi dua jenis: preferensi untuk kesejahteraan intrinsik diri sendiri dan preferensi mengenai kesejahteraan orang lain. (Terdapat banyak perdebatan tentang apakah keduanya dapat dipisahkan dengan rapi, tetapi saya akan mengesampingkan perdebatan itu.) Kesejahteraan intrinsik mengacu pada kualitas kehidupan seseorang sendiri, seperti tempat tinggal, kehangatan, makanan, keamanan, dan sebagainya, yang diinginkan dengan sendirinya daripada mengacu pada kualitas kehidupan orang lain.

Untuk membuat gagasan ini lebih konkret, mari kita misalkan dunia ini berisi dua orang, Alice dan Bob. Utilitas keseluruhan Alice terdiri dari kesejahteraan intrinsiknya sendiri ditambah faktor CAB dikalikan dengan kesejahteraan intrinsik Bob. Faktor kepedulian CAB menunjukkan seberapa besar Alice peduli pada Bob.

Demikian pula, utilitas keseluruhan Bob terdiri dari kesejahteraan intrinsiknya ditambah faktor kepedulian CBA dikalikan kesejahteraan intrinsik Alice, di mana CBA menunjukkan seberapa besar Bob peduli pada Alice. Robbie mencoba membantu Alice dan Bob, yang berarti (katakanlah) memaksimalkan jumlah utilitas mereka berdua. Dengan demikian, Robbie perlu memperhatikan tidak hanya kesejahteraan individu masing-masing tetapi juga seberapa besar masing-masing peduli pada kesejahteraan yang lain.

Tanda faktor kepedulian CAB dan CBA sangat penting. Misalnya, jika CAB positif, Alice "baik": dia mendapatkan kebahagiaan dari kesejahteraan Bob. Semakin positif CAB, semakin Alice bersedia mengorbankan sebagian kesejahteraannya sendiri untuk membantu Bob. Jika CAB nol, maka Alice benar-benar egois: jika dia bisa lolos begitu saja, dia akan mengalihkan sejumlah sumber daya dari Bob dan ke dirinya sendiri, bahkan jika Bob dibiarkan miskin dan kelaparan. Dihadapkan dengan Alice yang egois dan Bob yang baik hati, Robbie yang utilitarian jelas akan melindungi Bob dari kejahatan terburuk Alice.

Menariknya, keseimbangan akhir biasanya akan membuat Bob memiliki kesejahteraan intrinsik yang lebih rendah daripada Alice, tetapi ia mungkin memiliki kebahagiaan keseluruhan yang lebih besar karena ia peduli dengan kesejahteraan Alice. Anda mungkin merasa bahwa keputusan Robbie sangat tidak adil jika keputusan tersebut membuat Bob memiliki kesejahteraan yang lebih rendah daripada Alice hanya karena ia lebih baik hati daripada Alice: Bukankah ia akan membenci hasilnya dan tidak bahagia? Ya, mungkin saja, tetapi itu akan menjadi model yang berbeda model yang mencakup istilah untuk kebencian atas perbedaan kesejahteraan.

Dalam model sederhana kita, Bob akan menerima hasilnya dengan tenang. Bahkan, dalam situasi keseimbangan, ia akan menolak setiap upaya untuk mentransfer sumber daya dari Alice kepada dirinya sendiri, karena itu akan mengurangi kebahagiaan keseluruhannya. Jika Anda berpikir ini sama sekali tidak realistis, pertimbangkan kasus di mana Alice adalah putri Bob yang baru lahir.

Kasus yang benar-benar bermasalah bagi Robbie adalah ketika CAB negatif: dalam kasus itu, Alice benar-benar jahat. Saya akan menggunakan frasa altruisme negatif untuk merujuk pada preferensi semacam itu. Seperti halnya Harriet yang sadis yang disebutkan sebelumnya, ini bukan tentang keserakahan dan keegoisan biasa, di mana Alice puas

mengurangi bagian Bob dari kue untuk meningkatkan bagiannya sendiri. Altruisme negatif berarti bahwa Alice memperoleh kebahagiaan semata-mata dari berkurangnya kesejahteraan orang lain, bahkan jika kesejahteraan intrinsiknya sendiri tidak berubah.

Dalam makalahnya yang memperkenalkan utilitarianisme preferensi, Harsanyi mengaitkan altruisme negatif dengan "sadisme, iri hati, kebencian, dan kedengkian" dan berpendapat bahwa hal-hal tersebut harus diabaikan dalam menghitung jumlah total utilitas manusia dalam suatu populasi:

> Tidak ada jumlah niat baik kepada individu X yang dapat membebankan kewajiban moral kepada saya untuk membantunya dalam menyakiti orang ketiga, individu Y.

Ini tampaknya menjadi salah satu bidang di mana masuk akal bagi para perancang mesin cerdas untuk (dengan hati-hati) sedikit memanipulasi keseimbangan keadilan, bisa dikatakan demikian.

Sayangnya, altruisme negatif jauh lebih umum daripada yang diperkirakan. Hal ini muncul bukan karena sadisme dan kebencian tetapi karena iri hati dan kebencian serta emosi kebalikannya, yang akan saya sebut kesombongan (karena tidak ada kata yang lebih baik). Jika Bob iri pada Alice, ia memperoleh ketidakbahagiaan dari perbedaan antara kesejahteraan Alice dan kesejahteraannya sendiri; semakin besar perbedaannya, semakin tidak bahagia dia. Sebaliknya, jika Alice bangga akan keunggulannya atas Bob, ia memperoleh kebahagiaan bukan hanya dari kesejahteraan intrinsiknya sendiri tetapi juga dari fakta bahwa kesejahteraannya lebih tinggi daripada Bob.

Mudah untuk menunjukkan bahwa, dalam pengertian matematis, kesombongan dan iri hati bekerja dengan cara yang kurang lebih sama seperti sadisme; Mereka mengarahkan Alice dan Bob untuk memperoleh kebahagiaan semata-mata dari mengurangi kesejahteraan satu sama lain, karena pengurangan kesejahteraan Bob meningkatkan kebanggaan Alice, sementara pengurangan kesejahteraan Alice mengurangi rasa iri Bob.

Jeffrey Sachs, ekonom pembangunan terkenal, pernah menceritakan sebuah kisah yang menggambarkan kekuatan preferensi semacam ini dalam pemikiran orang. Dia berada di Bangladesh segera setelah banjir besar menghancurkan salah satu wilayah negara itu. Dia berbicara dengan seorang petani yang telah kehilangan rumahnya, ladangnya, semua hewan ternaknya, dan salah satu anaknya. “Saya sangat menyesal Anda pasti sangat sedih,” kata Sachs. “Tidak sama sekali,” jawab petani itu. “Saya cukup senang karena tetangga saya yang terkutuk itu juga kehilangan istri dan semua anaknya!”

Analisis ekonomi tentang kesombongan dan iri hati khususnya dalam konteks status sosial dan konsumsi mencolok muncul dalam karya sosiolog Amerika Thorstein Veblen, yang bukunya tahun 1899, *The Theory of the Leisure Class*, menjelaskan konsekuensi buruk dari sikap-sikap ini. Pada tahun 1977, ekonom Inggris Fred Hirsch *menerbitkan The Social Limits to Growth*, di mana ia memperkenalkan gagasan barang posisional. Barang posisional adalah apa pun bisa berupa mobil, rumah, medali Olimpiade, pendidikan, pendapatan, atau aksen yang

memperoleh nilai persepsinya bukan hanya dari manfaat intrinsiknya tetapi juga dari sifat relatifnya, termasuk sifat kelangkaan dan keunggulan dibandingkan orang lain.

Pengejaran barang posisional, yang didorong oleh kesombongan dan iri hati, memiliki karakter permainan zero-sum, dalam arti bahwa Alice tidak dapat meningkatkan posisi relatifnya tanpa memperburuk posisi relatif Bob, dan sebaliknya. (Hal ini tampaknya tidak mencegah pemborosan sejumlah besar uang dalam upaya ini.) Barang-barang posisional tampaknya ada di mana-mana dalam kehidupan modern, sehingga mesin perlu memahami pentingnya barang-barang tersebut secara keseluruhan dalam preferensi individu.

Selain itu, para ahli teori identitas sosial mengusulkan bahwa keanggotaan dan kedudukan dalam suatu kelompok dan status keseluruhan kelompok relatif terhadap kelompok lain merupakan unsur penting dari harga diri manusia. Dengan demikian, sulit untuk memahami perilaku manusia tanpa memahami bagaimana individu memandang diri mereka sendiri sebagai anggota kelompok apakah kelompok tersebut adalah spesies, bangsa, kelompok etnis, partai politik, profesi, keluarga, atau pendukung tim sepak bola tertentu. Seperti halnya sadisme dan kebencian, kita dapat mengusulkan agar Robbie memberikan sedikit atau tidak sama sekali bobot pada kesombongan dan iri hati dalam rencananya untuk membantu Alice dan Bob. Namun, ada beberapa kesulitan dengan usulan ini.

Karena kesombongan dan iri hati bertentangan dengan kepedulian dalam sikap Alice terhadap kesejahteraan Bob, mungkin tidak mudah untuk memisahkan keduanya. Mungkin Alice sangat peduli, tetapi juga menderita iri hati; Sulit untuk membedakan Alice ini dari Alice lain yang hanya sedikit peduli tetapi sama sekali tidak iri. Terlebih lagi, mengingat prevalensi kesombongan dan iri hati dalam preferensi manusia, sangat penting untuk mempertimbangkan dengan sangat hati-hati konsekuensi dari mengabaikannya. Mungkin saja hal itu penting untuk harga diri, terutama dalam bentuk positifnya rasa hormat diri dan kekaguman terhadap orang lain.

Izinkan saya menekankan kembali poin yang telah disebutkan sebelumnya: mesin yang dirancang dengan tepat tidak akan berperilaku seperti yang mereka amati, bahkan jika mesin-mesin itu mempelajari preferensi iblis sadis. Bahkan, mungkin saja jika kita manusia mendapati diri kita dalam situasi yang tidak biasa berurusan dengan entitas yang sepenuhnya altruistik setiap hari, kita mungkin belajar untuk menjadi orang yang lebih baik lebih altruistik dan kurang didorong oleh kesombongan dan iri hati.

9.4 MANUSIA BODOH DAN EMOSIONAL

Judul bagian ini tidak dimaksudkan untuk merujuk pada subset manusia tertentu. Ini merujuk pada kita semua. Kita semua sangat bodoh dibandingkan dengan standar yang tak terjangkau yang ditetapkan oleh rasionalitas sempurna, dan kita semua tunduk pada pasang surut emosi yang beragam yang, sebagian besar, mengatur perilaku kita.

Mari kita mulai dengan kebodohan. Entitas yang sepenuhnya rasional memaksimalkan kepuasan yang diharapkan dari preferensinya atas semua kemungkinan kehidupan masa depan yang dapat dipilihnya. Saya tidak dapat mulai menuliskan angka yang menggambarkan kompleksitas masalah pengambilan keputusan ini, tetapi saya menemukan eksperimen

pemikiran berikut ini bermanfaat. Pertama, perhatikan bahwa jumlah pilihan kontrol motorik yang dibuat manusia seumur hidup adalah sekitar dua puluh triliun. (Lihat Lampiran A untuk perhitungan detailnya.) Selanjutnya, mari kita lihat seberapa jauh kekuatan kasar akan membawa kita dengan bantuan laptop fisika pamungkas Seth Lloyd, yang satu miliar triliun triliun kali lebih cepat daripada komputer tercepat di dunia.

Kita akan memberinya tugas untuk menghitung semua kemungkinan urutan kata-kata bahasa Inggris (mungkin sebagai pemanasan untuk Perpustakaan Babel karya Jorge Luis Borges), dan kita akan membiarkannya berjalan selama setahun. Berapa panjang urutan yang dapat dihitungnya dalam waktu tersebut? Seribu halaman teks? Satu juta halaman? Tidak. Sebelas kata. Ini memberi tahu Anda sesuatu tentang kesulitan merancang kehidupan terbaik yang mungkin dari dua puluh triliun tindakan. Singkatnya, kita jauh lebih jauh dari rasional daripada siput yang mampu menyalip pesawat ruang angkasa Enterprise yang melaju dengan kecepatan warp sembilan. Kita sama sekali tidak tahu seperti apa kehidupan yang dipilih secara rasional.

Implikasinya adalah bahwa manusia sering bertindak dengan cara yang bertentangan dengan preferensi mereka sendiri. Misalnya, ketika Lee Sedol kalah dalam pertandingan Go-nya melawan AlphaGo, ia memainkan satu atau lebih langkah yang menjamin kekalahannya, dan AlphaGo dapat (setidaknya dalam beberapa kasus) mendeteksi bahwa ia telah melakukan hal ini. Namun, akan salah jika AlphaGo menyimpulkan bahwa Lee Sedol memiliki preferensi untuk kalah. Sebaliknya, masuk akal untuk menyimpulkan bahwa Lee Sedol memiliki preferensi untuk menang tetapi memiliki beberapa keterbatasan komputasi yang mencegahnya memilih langkah yang tepat dalam semua kasus.

Dengan demikian, untuk memahami perilaku Lee Sedol dan mempelajari preferensinya, robot yang mengikuti prinsip ketiga ("sumber informasi utama tentang preferensi manusia adalah perilaku manusia") harus memahami sesuatu tentang proses kognitif yang menghasilkan perilakunya. Robot tersebut tidak dapat berasumsi bahwa ia rasional.

Hal ini memberikan masalah penelitian yang sangat serius bagi komunitas AI, ilmu kognitif, psikologi, dan ilmu saraf: untuk memahami cukup banyak tentang kognisi manusia sehingga kita (atau lebih tepatnya, mesin-mesin bermanfaat kita) dapat "merekayasa balik" perilaku manusia untuk mendapatkan preferensi mendasar yang dalam, sejauh preferensi tersebut ada.

Manusia berhasil melakukan sebagian dari hal ini, mempelajari nilai-nilai mereka dari orang lain dengan sedikit bimbingan dari biologi, jadi tampaknya hal itu mungkin. Manusia memiliki keuntungan: mereka dapat menggunakan arsitektur kognitif mereka sendiri untuk mensimulasikan arsitektur kognitif manusia lain, tanpa mengetahui apa arsitektur tersebut "Jika saya menginginkan X, saya akan melakukan hal yang sama seperti yang dilakukan Ibu, jadi Ibu pasti menginginkan X."

Mesin tidak memiliki keunggulan ini. Mereka dapat mensimulasikan mesin lain dengan mudah, tetapi tidak manusia. Kemungkinan besar mereka tidak akan segera memiliki akses ke model kognisi manusia yang lengkap, baik yang bersifat umum maupun yang disesuaikan

dengan individu tertentu. Sebaliknya, dari sudut pandang praktis, masuk akal untuk melihat cara-cara utama di mana manusia menyimpang dari rasionalitas dan mempelajari cara mempelajari preferensi dari perilaku yang menunjukkan penyimpangan tersebut.

Salah satu perbedaan yang jelas antara manusia dan entitas rasional adalah bahwa, pada saat tertentu, kita tidak memilih di antara semua langkah pertama yang mungkin dari semua kemungkinan kehidupan di masa depan. Bahkan tidak mendekati itu. Sebaliknya, kita biasanya tertanam dalam hierarki "subrutin" yang sangat berlapis. Secara umum, kita mengejar tujuan jangka pendek daripada memaksimalkan preferensi atas kehidupan di masa depan, dan kita hanya dapat bertindak sesuai dengan batasan subrutin yang sedang kita jalani saat ini.

Saat ini, misalnya, saya sedang mengetik kalimat ini: Saya dapat memilih bagaimana melanjutkan setelah titik dua, tetapi tidak pernah terlintas dalam pikiran saya untuk bertanya-tanya apakah saya harus berhenti menulis kalimat ini dan mengikuti kursus rap online atau membakar rumah dan mengklaim asuransi atau hal lain dari sekian banyak hal yang dapat saya lakukan selanjutnya. Banyak dari hal-hal lain ini mungkin sebenarnya lebih baik daripada yang saya lakukan, tetapi, mengingat hierarki komitmen saya, seolah-olah hal-hal lain itu tidak ada.

Memahami tindakan manusia, kemudian, tampaknya membutuhkan pemahaman tentang hierarki subrutin ini (yang mungkin sangat individual): subrutin mana yang sedang dijalankan orang tersebut saat ini, tujuan jangka pendek apa yang sedang dikejar dalam subrutin ini, dan bagaimana tujuan tersebut berhubungan dengan preferensi jangka panjang yang lebih dalam. Secara lebih umum, mempelajari preferensi manusia tampaknya membutuhkan pembelajaran tentang struktur kehidupan manusia yang sebenarnya. Apa saja hal-hal yang dapat kita lakukan sebagai manusia, baik secara individu maupun bersama-sama?

Aktivitas apa yang menjadi ciri khas berbagai budaya dan tipe individu? Ini adalah pertanyaan penelitian yang sangat menarik dan menantang. Jelas, pertanyaan-pertanyaan ini tidak memiliki jawaban tetap karena kita manusia terus menambahkan aktivitas dan struktur perilaku baru ke dalam repertoar kita. Tetapi bahkan jawaban parsial dan sementara pun akan sangat berguna untuk semua jenis sistem cerdas yang dirancang untuk membantu manusia dalam kehidupan sehari-hari mereka.

Sifat lain yang jelas dari tindakan manusia adalah bahwa tindakan tersebut sering kali didorong oleh emosi. Dalam beberapa kasus, ini adalah hal yang baik emosi seperti cinta dan rasa syukur tentu saja sebagian membentuk preferensi kita, dan tindakan yang dipandu oleh emosi tersebut dapat bersifat rasional meskipun tidak sepenuhnya dipertimbangkan. Dalam kasus lain, respons emosional mengarah pada tindakan yang bahkan kita, manusia yang bodoh, sadari sebagai kurang rasional setelah kejadian, tentu saja. Misalnya, Harriet yang marah dan frustrasi menampar Alice yang berusia sepuluh tahun yang keras kepala mungkin langsung menyesali tindakannya.

Robbie, yang mengamati tindakan tersebut, seharusnya (biasanya, meskipun tidak dalam semua kasus) mengaitkan tindakan tersebut dengan kemarahan dan frustrasi serta kurangnya pengendalian diri daripada sadisme yang disengaja demi kesenangan semata. Agar hal ini berhasil, Robbie harus memiliki pemahaman tentang keadaan emosional manusia,

termasuk penyebabnya, bagaimana keadaan tersebut berkembang dari waktu ke waktu sebagai respons terhadap rangsangan eksternal, dan efeknya terhadap tindakan.

Para ahli saraf mulai memahami mekanisme beberapa keadaan emosional dan hubungannya dengan proses kognitif lainnya, dan ada beberapa pekerjaan yang bermanfaat tentang metode komputasi untuk mendeteksi, memprediksi, dan memanipulasi keadaan emosional manusia, tetapi masih banyak yang perlu dipelajari. Sekali lagi, mesin berada dalam posisi yang kurang menguntungkan dalam hal emosi: mereka tidak dapat menghasilkan simulasi internal suatu pengalaman untuk melihat keadaan emosional apa yang akan ditimbulkannya.

Selain memengaruhi tindakan kita, emosi mengungkapkan informasi yang berguna tentang preferensi mendasar kita. Misalnya, Alice kecil mungkin menolak mengerjakan pekerjaan rumahnya, dan Harriet marah dan frustrasi karena dia sangat ingin Alice berprestasi di sekolah dan memiliki kesempatan hidup yang lebih baik daripada Harriet sendiri. Jika Robbie mampu memahami hal ini bahkan jika dia tidak dapat mengalaminya sendiri dia dapat belajar banyak dari tindakan Harriet yang kurang rasional. Oleh karena itu, seharusnya mungkin untuk menciptakan model dasar keadaan emosional manusia yang cukup untuk menghindari kesalahan paling fatal dalam menyimpulkan preferensi manusia dari perilaku.

9.5 KETIDAKPASTIAN DAN KESALAHAN DALAM PREFERENSI MANUSIA

Premis utama buku ini adalah bahwa ada masa depan yang kita inginkan dan masa depan yang ingin kita hindari, seperti kepunahan dalam waktu dekat atau diubah menjadi peternakan manusia seperti dalam film The Matrix. Dalam hal ini, ya, tentu saja manusia memiliki preferensi. Namun, begitu kita membahas detail tentang bagaimana manusia lebih memilih kehidupan mereka berjalan, segalanya menjadi jauh lebih rumit.

Ketidakpastian dan Kesalahan

Salah satu sifat manusia yang jelas, jika Anda memikirkannya, adalah bahwa mereka tidak selalu tahu apa yang mereka inginkan. Misalnya, buah durian menimbulkan respons yang berbeda dari orang yang berbeda: beberapa orang menganggap bahwa "rasanya melampaui semua buah lain di dunia" sementara yang lain menyamakannya dengan "limbah, muntah basi, semprotan sigung, dan kapas bedah bekas." Saya sengaja menahan diri untuk tidak mencoba durian sebelum publikasi, sehingga saya dapat mempertahankan netralitas pada poin ini: saya sama sekali tidak tahu saya akan berada di kubu mana. Hal yang sama mungkin berlaku untuk banyak orang yang mempertimbangkan karier masa depan, pasangan hidup masa depan, aktivitas pasca pensiun masa depan, dan sebagainya.

Setidaknya ada dua jenis ketidakpastian preferensi. Yang pertama adalah ketidakpastian epistemik yang nyata, seperti yang saya alami tentang preferensi durian saya. Tidak ada jumlah pemikiran yang akan menyelesaikan ketidakpastian ini. Ada fakta empiris tentang masalah ini, dan saya dapat mengetahui lebih banyak dengan mencoba beberapa durian, dengan membandingkan DNA saya dengan DNA pecinta dan pembenci durian, dan sebagainya. Yang kedua muncul dari keterbatasan komputasi: melihat dua posisi Go, saya tidak

yakin mana yang saya sukai karena konsekuensi dari masing-masing posisi berada di luar kemampuan saya untuk menyelesaikannya sepenuhnya.

Ketidakpastian juga muncul dari fakta bahwa pilihan yang disajikan kepada kita biasanya tidak sepenuhnya ditentukan kadang-kadang sangat tidak lengkap sehingga hampir tidak memenuhi syarat sebagai pilihan sama sekali. Ketika Alice akan lulus dari sekolah menengah, seorang konselor karier mungkin menawarkannya pilihan antara "pustakawan" dan "penambang batu bara"; dia mungkin, dengan cukup wajar, mengatakan, "Saya tidak yakin mana yang saya sukai."

Di sini, ketidakpastian berasal dari ketidakpastian epistemik tentang preferensinya sendiri, misalnya, debu batu bara versus debu buku; dari ketidakpastian komputasi saat dia berjuang untuk mencari tahu bagaimana dia dapat memanfaatkan setiap pilihan karier sebaik mungkin; dan dari ketidakpastian biasa tentang dunia, seperti keraguannya tentang kelangsungan jangka panjang tambang batu bara lokalnya.

Karena alasan ini, adalah ide yang buruk untuk mengidentifikasi preferensi manusia dengan pilihan sederhana antara opsi yang tidak dijelaskan secara lengkap yang sulit dievaluasi dan mencakup elemen keinginan yang tidak diketahui. Pilihan-pilihan tersebut memberikan bukti tidak langsung tentang preferensi yang mendasarinya, tetapi pilihan-pilihan tersebut bukanlah pembentuk preferensi itu sendiri. Itulah mengapa saya telah mengemukakan gagasan preferensi dalam konteks kehidupan masa depan misalnya dengan membayangkan bahwa Anda dapat mengalami, dalam bentuk yang dipadatkan, dua film berbeda tentang kehidupan masa depan Anda dan kemudian menyatakan preferensi di antara keduanya (lihat halaman ini).

Eksperimen pemikiran ini tentu saja mustahil untuk dilakukan dalam praktik, tetapi dapat dibayangkan bahwa dalam banyak kasus preferensi yang jelas akan muncul jauh sebelum semua detail setiap film terisi dan dialami sepenuhnya. Anda mungkin tidak tahu sebelumnya mana yang akan Anda sukai, bahkan dengan ringkasan plot; tetapi ada jawaban untuk pertanyaan sebenarnya, berdasarkan siapa Anda sekarang, sama seperti ada jawaban untuk pertanyaan apakah Anda akan menyukai durian ketika Anda mencobanya.

Fakta bahwa Anda mungkin tidak yakin tentang preferensi Anda sendiri tidak menimbulkan masalah khusus bagi pendekatan berbasis preferensi untuk AI yang terbukti bermanfaat. Memang, sudah ada beberapa algoritma yang memperhitungkan ketidakpastian Robbie dan Harriet tentang preferensi Harriet dan memungkinkan kemungkinan bahwa Harriet mungkin sedang mempelajari preferensinya sementara Robbie juga. Sama seperti ketidakpastian Robbie tentang preferensi Harriet dapat dikurangi dengan mengamati perilaku Harriet, ketidakpastian Harriet tentang preferensinya sendiri dapat dikurangi dengan mengamati reaksinya sendiri terhadap pengalaman.

Kedua jenis ketidakpastian tersebut tidak harus berhubungan langsung; Robbie juga tidak selalu lebih tidak yakin daripada Harriet tentang preferensi Harriet. Misalnya, Robbie mungkin dapat mendeteksi bahwa Harriet memiliki kecenderungan genetik yang kuat untuk membenci rasa durian. Dalam hal itu, ia akan memiliki sedikit ketidakpastian tentang preferensi duriannya, bahkan ketika Harriet sama sekali tidak tahu.

Jika Harriet dapat merasa tidak yakin tentang preferensinya terhadap peristiwa di masa depan, maka, kemungkinan besar, ia juga bisa salah. Misalnya, dia mungkin yakin bahwa dia tidak akan menyukai durian (atau, katakanlah, telur hijau dan ham) sehingga dia menghindarinya dengan segala cara, tetapi mungkin saja jika seseorang menyelipkan durian ke dalam salad buahnya suatu hari dia malah menyukainya.

Dengan demikian, Robbie tidak dapat berasumsi bahwa tindakan Harriet mencerminkan pengetahuan yang akurat tentang preferensinya sendiri: beberapa mungkin didasarkan sepenuhnya pada pengalaman, sementara yang lain mungkin terutama didasarkan pada dugaan, prasangka, ketakutan akan hal yang tidak diketahui, atau generalisasi yang kurang didukung. Robbie yang cukup bijaksana dapat sangat membantu Harriet dalam mengingatkannya tentang situasi seperti itu.

Pengalaman dan Ingatan

Beberapa psikolog mempertanyakan gagasan bahwa ada satu diri yang preferensinya berdaulat seperti yang disarankan oleh prinsip otonomi preferensi Harsanyi. Yang paling menonjol di antara para psikolog ini adalah mantan kolega saya di Berkeley, Daniel Kahneman. Kahneman, yang memenangkan Hadiah Nobel 2002 untuk karyanya di bidang ekonomi perilaku, adalah salah satu pemikir paling berpengaruh tentang topik preferensi manusia. Buku terbarunya, *Thinking, Fast and Slow,* menceritakan secara rinci serangkaian eksperimen yang meyakinkannya bahwa ada dua diri, diri yang mengalami dan diri yang mengingat yang preferensinya saling bertentangan.

Diri yang mengalami adalah diri yang diukur oleh hedonimeter, yang dibayangkan oleh ekonom Inggris abad ke-19, Francis Edgeworth, sebagai “instrumen yang idealnya sempurna, mesin psikofisik, yang terus-menerus mencatat tingkat kesenangan yang dialami oleh individu, tepat sesuai dengan penilaian kesadaran.” Menurut utilitarianisme hedonis, nilai keseluruhan dari setiap pengalaman bagi individu hanyalah jumlah dari nilai hedonis setiap saat selama pengalaman tersebut. Gagasan ini berlaku sama baiknya untuk makan es krim atau menjalani seluruh hidup.

Di sisi lain, diri yang mengingat adalah yang "bertanggung jawab" ketika ada keputusan yang harus dibuat. Diri ini memilih pengalaman baru berdasarkan ingatan akan pengalaman sebelumnya dan keinginan akan pengalaman tersebut. Eksperimen Kahneman menunjukkan bahwa diri yang mengingat memiliki gagasan yang sangat berbeda dari diri yang mengalami.

Eksperimen paling sederhana untuk dipahami melibatkan mencelupkan tangan subjek ke dalam air dingin. Ada dua rezim yang berbeda: pada rezim pertama, perendaman dilakukan selama 60 detik dalam air pada suhu 14 derajat Celcius; pada rezim kedua, perendaman dilakukan selama 60 detik dalam air pada suhu 14 derajat diikuti oleh 30 detik pada suhu 15 derajat. (Suhu ini mirip dengan suhu laut di California Utara cukup dingin sehingga hampir semua orang mengenakan pakaian selam di dalam air.) Semua subjek melaporkan pengalaman tersebut sebagai tidak menyenangkan. Setelah mengalami kedua rezim (dalam urutan apa pun, dengan jeda 7 menit di antaranya), subjek diminta untuk memilih mana yang ingin mereka ulangi. Sebagian besar subjek lebih memilih untuk mengulangi 60 + 30 daripada hanya perendaman 60 detik. Kahneman berpendapat bahwa, dari sudut pandang diri yang

mengalami, 60 + 30 pasti lebih buruk daripada 60, karena mencakup 60 dan pengalaman tidak menyenangkan lainnya. Namun, diri yang mengingat memilih 60 + 30. Mengapa?

Penjelasan Kahneman adalah bahwa diri yang mengingat melihat kembali dengan kacamata yang agak aneh, terutama memperhatikan nilai "puncak" (nilai hedonis tertinggi atau terendah) dan nilai "akhir" (nilai hedonis di akhir pengalaman). Durasi bagian-bagian pengalaman yang berbeda sebagian besar diabaikan. Tingkat ketidaknyamanan puncak untuk 60 dan 60 + 30 sama, tetapi tingkat akhirnya berbeda: dalam kasus 60 + 30, airnya satu derajat lebih hangat. Jika diri yang mengingat mengevaluasi pengalaman berdasarkan nilai puncak dan akhir, alih-alih dengan menjumlahkan nilai hedonis dari waktu ke waktu, maka 60 + 30 lebih baik, dan inilah yang ditemukan. Model puncak-akhir tampaknya menjelaskan banyak temuan aneh lainnya dalam literatur tentang preferensi.

Kahneman tampaknya (mungkin tepat) memiliki dua pendapat tentang temuannya. Dia menegaskan bahwa diri yang mengingat "hanya membuat kesalahan" dan memilih pengalaman yang salah karena ingatannya salah dan tidak lengkap; dia menganggap ini sebagai "kabar buruk bagi para penganut rasionalitas pilihan." Di sisi lain, dia menulis, "Teori kesejahteraan yang mengabaikan apa yang diinginkan orang tidak dapat dipertahankan." Misalnya, anggaplah Harriet telah mencoba Pepsi dan Coke dan sekarang sangat lebih menyukai Pepsi; akan absurd untuk memaksanya minum Coke berdasarkan penjumlahan pembacaan hedonimeter rahasia yang diambil selama setiap percobaan.

Faktanya adalah tidak ada hukum yang mengharuskan preferensi kita antara pengalaman didefinisikan oleh jumlah nilai hedonis selama beberapa saat. Memang benar bahwa model matematika standar berfokus pada memaksimalkan jumlah imbalan, tetapi motivasi awalnya adalah kemudahan matematis. Pembenaran datang kemudian dalam bentuk asumsi teknis di mana rasional untuk memutuskan berdasarkan penjumlahan imbalan, tetapi asumsi teknis tersebut tidak selalu berlaku dalam kenyataan. Misalnya, anggaplah Harriet memilih antara dua urutan nilai hedonik: [10,10,10,10,10] dan [0,0,40,0,0]. Sangat mungkin dia hanya lebih menyukai urutan kedua; tidak ada hukum matematika yang dapat memaksanya untuk membuat pilihan berdasarkan jumlah daripada, misalnya, maksimum.

Kahneman mengakui bahwa situasi ini semakin rumit karena peran penting antisipasi dan ingatan dalam kesejahteraan. Ingatan akan satu pengalaman yang menyenangkan hari pernikahan seseorang, kelahiran seorang anak, sore hari yang dihabiskan untuk memetik blackberry dan membuat selai dapat membantu seseorang melewati tahun-tahun kerja keras dan kekecewaan. Mungkin diri yang mengingat tidak hanya mengevaluasi pengalaman itu sendiri tetapi juga efek totalnya pada nilai kehidupan di masa depan melalui pengaruhnya pada ingatan di masa depan. Dan mungkin diri yang mengingat, dan bukan diri yang mengalami, adalah penilai terbaik tentang apa yang akan diingat.

Waktu dan Perubahan

Hampir tidak perlu dikatakan bahwa orang-orang yang bijaksana di abad ke-21 tidak ingin meniru preferensi, misalnya, masyarakat Romawi di abad ke-2, yang penuh dengan pembantaian gladiator untuk hiburan publik, ekonomi yang berbasis pada perbudakan, dan pembantaian brutal terhadap orang-orang yang kalah. (Kita tidak perlu membahas paralel

yang jelas dengan karakteristik ini dalam masyarakat modern.) Standar moralitas jelas berkembang seiring waktu seiring kemajuan peradaban kita atau pergeseran, jika Anda lebih suka. Hal ini menunjukkan, pada gilirannya, bahwa generasi mendatang mungkin akan merasa sangat jijik dengan sikap kita saat ini terhadap, misalnya, kesejahteraan hewan. Karena alasan ini, penting agar mesin yang bertugas menerapkan preferensi manusia mampu merespons perubahan preferensi tersebut dari waktu ke waktu daripada menetapkannya secara permanen. Tiga prinsip dari Bab 7 mengakomodasi perubahan tersebut secara alami, karena prinsip-prinsip tersebut mengharuskan mesin untuk mempelajari dan menerapkan preferensi manusia saat ini banyak sekali, semuanya berbeda daripada satu set preferensi ideal atau preferensi perancang mesin yang mungkin sudah lama meninggal.

Kemungkinan perubahan dalam preferensi khas populasi manusia sepanjang sejarah secara alami memfokuskan perhatian pada pertanyaan tentang bagaimana preferensi setiap individu terbentuk dan plastisitas preferensi orang dewasa. Preferensi kita tentu dipengaruhi oleh biologi kita: kita biasanya menghindari rasa sakit, kelaparan, dan kehausan, misalnya. Namun, biologi kita tetap relatif konstan, sehingga preferensi yang tersisa pasti muncul dari pengaruh budaya dan keluarga.

Sangat mungkin, anak-anak terus-menerus menjalankan beberapa bentuk pembelajaran penguatan terbalik untuk mengidentifikasi preferensi orang tua dan teman sebaya untuk menjelaskan perilaku mereka; anak-anak kemudian mengadopsi preferensi ini sebagai preferensi mereka sendiri. Bahkan sebagai orang dewasa, preferensi kita berkembang melalui pengaruh media, pemerintah, teman, pemberi kerja, dan pengalaman langsung kita sendiri. Misalnya, mungkin saja banyak pendukung Reich Ketiga tidak memulai sebagai sadis genosida yang haus akan kemurnian ras.

Perubahan preferensi menghadirkan tantangan bagi teori rasionalitas baik pada tingkat individu maupun masyarakat. Misalnya, prinsip otonomi preferensi Harsanyi tampaknya mengatakan bahwa setiap orang berhak atas preferensi apa pun yang mereka miliki dan tidak seorang pun boleh menyentuhnya. Namun, jauh dari tak tersentuh, preferensi disentuh dan dimodifikasi sepanjang waktu, oleh setiap pengalaman yang dialami seseorang. Mesin tidak dapat menghindari modifikasi preferensi manusia, karena mesin memodifikasi pengalaman manusia.

Penting, meskipun terkadang sulit, untuk memisahkan perubahan preferensi dari pembaruan preferensi, yang terjadi ketika Harriet yang awalnya tidak yakin mempelajari lebih lanjut tentang preferensinya sendiri melalui pengalaman. Pembaruan preferensi dapat mengisi celah dalam pengetahuan diri dan mungkin menambahkan kepastian pada preferensi yang sebelumnya dipegang secara lemah dan sementara. Perubahan preferensi, di sisi lain, bukanlah proses yang dihasilkan dari bukti tambahan tentang apa sebenarnya preferensi seseorang. Dalam kasus ekstrem, Anda dapat membayangkannya sebagai akibat dari pemberian obat atau bahkan operasi otak itu terjadi dari proses yang mungkin tidak kita pahami atau setujui.

Perubahan preferensi bermasalah setidaknya karena dua alasan. Alasan pertama adalah tidak jelas preferensi mana yang harus dipertimbangkan saat mengambil keputusan:

preferensi yang dimiliki Harriet pada saat pengambilan keputusan atau preferensi yang akan dimilikinya selama dan setelah peristiwa yang diakibatkan oleh keputusannya. Dalam bioetika, misalnya, ini adalah dilema yang sangat nyata karena preferensi orang tentang intervensi medis dan perawatan akhir hayat memang berubah, seringkali secara dramatis, setelah mereka sakit parah. Dengan asumsi perubahan ini tidak disebabkan oleh penurunan kapasitas intelektual, preferensi siapa yang harus dihormati?

Alasan kedua mengapa perubahan preferensi bermasalah adalah karena tampaknya tidak ada dasar rasional yang jelas untuk mengubah (bukan memperbarui) preferensi seseorang. Jika Harriet lebih menyukai A daripada B, tetapi dapat memilih untuk menjalani pengalaman yang dia tahu akan membuatnya lebih menyukai B daripada A, mengapa dia harus melakukan itu? Hasilnya adalah dia kemudian akan memilih B, yang saat ini tidak diinginkannya.

Masalah perubahan preferensi muncul dalam bentuk dramatis dalam legenda Ulysses dan Siren. Siren adalah makhluk mitos yang nyanyiannya memikat para pelaut menuju malapetaka di bebatuan pulau-pulau tertentu di Mediterania. Ulysses, yang ingin mendengar nyanyian itu, memerintahkan para pelautnya untuk menyumbat telinga mereka dengan lilin dan mengikatnya ke tiang layar; dalam keadaan apa pun mereka tidak boleh menuruti permohonannya selanjutnya untuk membebaskannya. Jelas, dia ingin para pelaut menghormati preferensi yang dia miliki awalnya, bukan preferensi yang akan dia miliki setelah Siren memikatnya. Legenda ini menjadi judul buku karya filsuf Norwegia Jon Elster, yang membahas kelemahan kemauan dan tantangan lain terhadap gagasan teoritis rasionalitas.

Mengapa mesin cerdas mungkin dengan sengaja berupaya memodifikasi preferensi manusia? Jawabannya cukup sederhana: untuk membuat preferensi lebih mudah dipenuhi. Kita melihat ini di Bab 1 dengan kasus optimasi klik-tayang media sosial. Salah satu tanggapan mungkin mengatakan bahwa mesin harus memperlakukan preferensi manusia sebagai sesuatu yang sakral: tidak ada yang boleh mengubah preferensi manusia. Sayangnya, ini sama sekali tidak mungkin. Keberadaan robot pembantu yang berguna kemungkinan akan berpengaruh pada preferensi manusia.

Salah satu solusi yang mungkin adalah agar mesin mempelajari meta-preferensi manusia yaitu, preferensi tentang jenis proses perubahan preferensi apa yang mungkin dapat diterima atau tidak dapat diterima. Perhatikan penggunaan "proses perubahan preferensi" daripada "perubahan preferensi" di sini. Itu karena menginginkan preferensi seseorang berubah ke arah tertentu seringkali berarti sudah memiliki preferensi tersebut; yang sebenarnya diinginkan dalam kasus seperti itu adalah kemampuan untuk lebih baik dalam mengimplementasikan preferensi tersebut. Misalnya, jika Harriet berkata, "Saya ingin preferensi saya berubah sehingga saya tidak menginginkan kue sebanyak sekarang," maka dia sudah memiliki preferensi untuk masa depan dengan konsumsi kue yang lebih sedikit; yang sebenarnya dia inginkan adalah mengubah arsitektur kognitifnya sehingga perilakunya lebih mencerminkan preferensi tersebut.

Yang saya maksud dengan “preferensi tentang jenis proses perubahan preferensi apa yang mungkin dapat diterima atau tidak dapat diterima” adalah, misalnya, pandangan bahwa

seseorang mungkin akan memiliki preferensi yang “lebih baik” dengan berkeliling dunia dan mengalami berbagai macam budaya, atau dengan berpartisipasi dalam komunitas intelektual yang dinamis yang secara menyeluruh mengeksplorasi berbagai tradisi moral, atau dengan menyisihkan waktu untuk introspeksi dan berpikir keras tentang kehidupan dan maknanya. Saya akan menyebut proses-proses ini netral terhadap preferensi, dalam arti bahwa seseorang tidak mengantisipasi bahwa proses tersebut akan mengubah preferensi seseorang ke arah tertentu, sambil mengakui bahwa beberapa orang mungkin sangat tidak setuju dengan karakterisasi tersebut.

Tentu saja, tidak semua proses netral terhadap preferensi itu diinginkan misalnya, sedikit orang yang berharap untuk mengembangkan preferensi yang “lebih baik” dengan memukul kepala mereka sendiri. Menundukkan diri pada proses perubahan preferensi yang dapat diterima sama analognya dengan menjalankan eksperimen untuk mencari tahu sesuatu tentang cara kerja dunia: Anda tidak pernah tahu sebelumnya bagaimana eksperimen itu akan berakhir, tetapi Anda tetap berharap untuk menjadi lebih baik dalam keadaan mental baru Anda.

Gagasan bahwa ada jalur yang dapat diterima untuk modifikasi preferensi tampaknya terkait dengan gagasan bahwa ada metode modifikasi perilaku yang dapat diterima di mana, misalnya, seorang pemberi kerja merekayasa situasi pilihan sehingga orang membuat pilihan yang "lebih baik" tentang menabung untuk pensiun. Seringkali ini dapat dilakukan dengan memanipulasi faktor-faktor "non-rasional" yang memengaruhi pilihan, daripada dengan membatasi pilihan atau mengenakan pajak pada pilihan yang "buruk". Nudge, sebuah buku karya ekonom Richard Thaler dan pakar hukum Cass Sunstein, menjabarkan berbagai metode dan peluang yang dianggap dapat diterima untuk "memengaruhi perilaku orang agar hidup mereka lebih panjang, lebih sehat, dan lebih baik."

Tidak jelas apakah metode modifikasi perilaku benar-benar hanya memodifikasi perilaku. Jika, ketika dorongan dihilangkan, perilaku yang dimodifikasi tetap ada yang mungkin merupakan hasil yang diinginkan dari intervensi tersebut maka sesuatu telah berubah dalam arsitektur kognitif individu (hal yang mengubah preferensi mendasar menjadi perilaku) atau dalam preferensi mendasar individu. Kemungkinan besar keduanya terjadi. Namun, yang jelas adalah bahwa strategi dorongan (nudge) mengasumsikan bahwa setiap orang memiliki preferensi yang sama untuk kehidupan yang “lebih panjang, lebih sehat, dan lebih baik”; setiap dorongan didasarkan pada definisi tertentu tentang kehidupan yang “lebih baik”, yang tampaknya bertentangan dengan otonomi preferensi.

Mungkin lebih baik, sebagai gantinya, untuk merancang proses bantuan yang netral terhadap preferensi yang membantu orang menyelaraskan keputusan dan arsitektur kognitif mereka dengan preferensi yang mendasarinya. Misalnya, dimungkinkan untuk merancang alat bantu kognitif yang menyoroti konsekuensi jangka panjang dari keputusan dan mengajarkan orang untuk mengenali benih konsekuensi tersebut di masa sekarang. Bahwa kita membutuhkan pemahaman yang lebih baik tentang proses pembentukan dan pembentukan preferensi manusia tampaknya sudah jelas, terutama karena pemahaman tersebut akan membantu kita merancang mesin yang menghindari perubahan yang tidak disengaja dan tidak

diinginkan dalam preferensi manusia seperti yang disebabkan oleh algoritma pemilihan konten media sosial.

Dengan pemahaman seperti itu, tentu saja, kita akan tergoda untuk merekayasa perubahan yang akan menghasilkan dunia yang “lebih baik”. Sebagian orang mungkin berpendapat bahwa kita harus menyediakan kesempatan yang jauh lebih besar untuk pengalaman "peningkatan" yang netral terhadap preferensi, seperti perjalanan, debat, dan pelatihan dalam berpikir analitis dan kritis. Misalnya, kita dapat memberikan kesempatan kepada setiap siswa sekolah menengah untuk tinggal selama beberapa bulan di setidaknya dua budaya lain yang berbeda dari budaya mereka sendiri.

Namun, hampir pasti kita akan ingin melangkah lebih jauh misalnya, dengan menerapkan reformasi sosial dan pendidikan yang meningkatkan koefisien altruisme bobot yang diberikan setiap individu pada kesejahteraan orang lain sambil mengurangi koefisien sadisme, kesombongan, dan iri hati. Apakah ini ide yang bagus? Haruskah kita merekrut mesin kita untuk membantu dalam proses ini? Tentu saja ini sangat menggoda. Bahkan, Aristoteles sendiri menulis, “Keprihatinan utama politik adalah untuk menumbuhkan karakter tertentu pada warga negara dan membuat mereka baik dan cenderung melakukan tindakan mulia.” Mari kita katakan saja bahwa ada risiko yang terkait dengan rekayasa preferensi yang disengaja dalam skala global. Kita harus bertindak dengan sangat hati-hati.

BAB 10
KEBEBASAN ATAU KETERGANTUNGAN PADA AI

Jika kita berhasil menciptakan sistem AI yang terbukti bermanfaat, kita akan menghilangkan risiko kehilangan kendali atas mesin supercerdas. Umat manusia dapat melanjutkan pengembangannya dan memanen manfaat yang hampir tak terbayangkan yang akan mengalir dari kemampuan untuk menggunakan kecerdasan yang jauh lebih besar dalam memajukan peradaban kita. Kita akan terbebas dari perbudakan selama ribuan tahun sebagai robot pertanian, industri, dan administrasi, dan kita akan bebas untuk memanfaatkan potensi hidup sebaik-baiknya. Dari sudut pandang zaman keemasan ini, kita akan melihat kembali kehidupan kita saat ini seperti yang dibayangkan Thomas Hobbes tentang kehidupan tanpa pemerintahan: kesepian, miskin, jahat, brutal, dan singkat. Atau mungkin tidak. Penjahat ala Bond mungkin akan melewati pengamanan kita dan melepaskan kecerdasan super yang tak terkendali yang tidak dapat dilawan oleh umat manusia. Dan jika kita selamat dari itu, kita mungkin akan mendapati diri kita secara bertahap melemah karena kita semakin mempercayakan pengetahuan dan keterampilan kita kepada mesin. Mesin-mesin tersebut mungkin menyarankan kita untuk tidak melakukan ini, karena memahami nilai jangka panjang otonomi manusia, tetapi kita dapat mengabaikan saran mereka.

10.1 MESIN YANG BERMANFAAT

Model standar yang mendasari sebagian besar teknologi abad ke-20 bergantung pada mesin yang mengoptimalkan tujuan tetap yang disediakan secara eksternal. Seperti yang telah kita lihat, model ini pada dasarnya cacat. Model ini hanya berfungsi jika tujuan tersebut dijamin lengkap dan benar, atau jika mesin dapat dengan mudah diatur ulang. Kedua kondisi tersebut tidak akan berlaku seiring dengan semakin kuatnya AI.

Jika tujuan yang disediakan secara eksternal dapat salah, maka tidak masuk akal bagi mesin untuk bertindak seolah-olah selalu benar. Oleh karena itu, usulan saya untuk mesin yang bermanfaat: mesin yang tindakannya dapat diharapkan untuk mencapai tujuan kita. Karena tujuan-tujuan ini ada di dalam diri kita, dan bukan di dalam mesin, mesin perlu mempelajari lebih lanjut tentang apa yang sebenarnya kita inginkan dari pengamatan pilihan yang kita buat dan bagaimana kita membuatnya. Mesin yang dirancang dengan cara ini akan tunduk pada manusia: mereka akan meminta izin; mereka akan bertindak hati-hati ketika panduan tidak jelas; dan mereka akan membiarkan diri mereka dimatikan.

Meskipun hasil awal ini untuk pengaturan yang disederhanakan dan diidealkan, saya percaya hasil ini akan bertahan dalam transisi ke pengaturan yang lebih realistis. Rekan-rekan saya telah berhasil menerapkan pendekatan yang sama pada masalah praktis seperti mobil otonom yang berinteraksi dengan pengemudi manusia. Misalnya, mobil otonom terkenal buruk dalam menangani rambu berhenti empat arah ketika tidak jelas siapa yang memiliki hak jalan. Namun, dengan merumuskan ini sebagai permainan bantuan, mobil tersebut menemukan solusi baru: ia benar-benar mundur sedikit untuk menunjukkan bahwa ia jelas

tidak berencana untuk jalan duluan. Manusia memahami sinyal ini dan melanjutkan perjalanan, yakin bahwa tidak akan terjadi tabrakan. Jelas, kita sebagai ahli manusia dapat memikirkan solusi ini dan memprogramnya ke dalam kendaraan, tetapi bukan itu yang terjadi; ini adalah bentuk komunikasi yang sepenuhnya diciptakan oleh kendaraan itu sendiri.

Seiring bertambahnya pengalaman kita di lingkungan lain, saya berharap kita akan terkejut dengan jangkauan dan kelancaran perilaku mesin saat berinteraksi dengan manusia. Kita sudah terbiasa dengan kebodohan mesin yang menjalankan perilaku yang tidak fleksibel dan telah diprogram sebelumnya atau mengejar tujuan yang pasti tetapi salah sehingga kita mungkin akan terkejut dengan betapa masuk akalnya mereka. Teknologi mesin yang terbukti bermanfaat adalah inti dari pendekatan baru terhadap AI dan dasar bagi hubungan baru antara manusia dan mesin.

Tampaknya juga memungkinkan untuk menerapkan ide serupa pada desain ulang "mesin" lain yang seharusnya melayani manusia, dimulai dengan sistem perangkat lunak biasa. Kita diajarkan untuk membangun perangkat lunak dengan menyusun subrutin, yang masing-masing memiliki spesifikasi yang jelas yang menyatakan apa yang seharusnya menjadi output untuk setiap input tertentu sama seperti tombol akar kuadrat pada kalkulator.

Spesifikasi ini merupakan analog langsung dari tujuan yang diberikan kepada sistem AI. Subrutin tidak seharusnya berhenti dan mengembalikan kendali ke lapisan yang lebih tinggi dari sistem perangkat lunak sampai menghasilkan output yang memenuhi spesifikasi. (Ini seharusnya mengingatkan Anda pada sistem AI yang terus-menerus mengejar tujuan yang diberikan.)

Pendekatan yang lebih baik adalah dengan memungkinkan ketidakpastian dalam spesifikasi. Misalnya, subrutin yang melakukan beberapa perhitungan matematika yang sangat rumit biasanya diberi batas kesalahan yang mendefinisikan presisi yang dibutuhkan untuk jawabannya dan harus mengembalikan solusi yang benar dalam batas kesalahan tersebut. Terkadang, ini mungkin membutuhkan komputasi selama berminggu-minggu. Sebaliknya, mungkin akan lebih baik untuk tidak terlalu presisi mengenai kesalahan yang diizinkan, sehingga subrutin dapat kembali setelah dua puluh detik dan berkata, “Saya telah menemukan solusi yang sebaik ini. Apakah ini baik-baik saja atau Anda ingin saya melanjutkan?” Dalam beberapa kasus, pertanyaan tersebut dapat merembes hingga ke tingkat tertinggi sistem perangkat lunak, sehingga pengguna manusia dapat memberikan panduan lebih lanjut kepada sistem. Jawaban manusia kemudian akan membantu dalam menyempurnakan spesifikasi di semua tingkatan.

Jenis pemikiran yang sama dapat diterapkan pada entitas seperti pemerintah dan perusahaan. Kegagalan pemerintah yang jelas termasuk terlalu memperhatikan preferensi (keuangan maupun politik) dari mereka yang berada di pemerintahan dan terlalu sedikit memperhatikan preferensi yang diperintah. Pemilihan umum seharusnya mengkomunikasikan preferensi kepada pemerintah, tetapi tampaknya memiliki bandwidth yang sangat kecil (sekitar satu byte informasi setiap beberapa tahun) untuk tugas yang kompleks seperti itu. Di terlalu banyak negara, pemerintah hanyalah sarana bagi satu kelompok orang untuk memaksakan kehendaknya kepada orang lain. Perusahaan-perusahaan berupaya lebih keras

untuk mempelajari preferensi pelanggan, baik melalui riset pasar maupun umpan balik langsung berupa keputusan pembelian. Di sisi lain, pembentukan preferensi manusia melalui iklan, pengaruh budaya, dan bahkan kecanduan zat kimia merupakan cara berbisnis yang diterima.

10.2 TATA KELOLA AI

AI memiliki kekuatan untuk membentuk kembali dunia, dan proses pembentukan kembali tersebut harus dikelola dan dipandu dengan cara tertentu. Jika banyaknya inisiatif untuk mengembangkan tata kelola AI yang efektif menjadi indikator, maka kita berada dalam kondisi yang sangat baik. Semua orang dan kerabatnya sedang membentuk Dewan atau Majelis atau Panel Internasional. Forum Ekonomi Dunia telah mengidentifikasi hampir tiga ratus upaya terpisah untuk mengembangkan prinsip-prinsip etika untuk AI. Kotak masuk email saya dapat diringkas sebagai undangan panjang ke Forum Konferensi KTT Dunia Global tentang Masa Depan Tata Kelola Internasional atas Dampak Sosial dan Etika Teknologi Kecerdasan Buatan yang Muncul.

Ini semua sangat berbeda dari apa yang terjadi dengan teknologi nuklir. Setelah Perang Dunia II, Amerika Serikat memegang semua kendali nuklir. Pada tahun 1953, presiden AS Dwight Eisenhower mengusulkan kepada PBB sebuah badan internasional untuk mengatur teknologi nuklir. Pada tahun 1957, Badan Energi Atom Internasional mulai bekerja; badan ini adalah satu-satunya pengawas global untuk pengembangan energi nuklir yang aman dan bermanfaat.

Sebaliknya, banyak tangan memegang kendali AI. Tentu saja, Amerika Serikat, Tiongkok, dan Uni Eropa mendanai banyak penelitian AI, tetapi hampir semuanya terjadi di luar laboratorium nasional yang aman. Para peneliti AI di universitas merupakan bagian dari komunitas internasional yang luas dan kooperatif, yang disatukan oleh kepentingan bersama, konferensi, perjanjian kerja sama, dan perkumpulan profesional seperti AAAI (*Association for the Advancement of Artificial Intelligence*) dan IEEE (*Institute of Electrical and Electronics Engineers*, yang mencakup puluhan ribu peneliti dan praktisi AI). Mungkin sebagian besar investasi dalam penelitian dan pengembangan AI sekarang terjadi di dalam perusahaan, besar dan kecil; pemain utama pada tahun 2019 adalah Google (termasuk DeepMind), Facebook, Amazon, Microsoft, dan IBM di Amerika Serikat dan Tencent, Baidu, dan, sampai batas tertentu, Alibaba di Tiongkok semuanya termasuk perusahaan terbesar di dunia. Semua kecuali Tencent dan Alibaba adalah anggota Partnership on AI, sebuah konsorsium industri yang di antara prinsip-prinsipnya terdapat janji kerja sama dalam hal keamanan AI. Terakhir, meskipun sebagian besar manusia memiliki sedikit keahlian di bidang AI, setidaknya ada kemauan yang dangkal di antara para pemain lain untuk mempertimbangkan kepentingan umat manusia.

Oleh karena itu, merekalah para pemain yang memegang sebagian besar kartu. Kepentingan mereka tidak sepenuhnya selaras, tetapi semuanya memiliki keinginan untuk mempertahankan kendali atas sistem AI seiring dengan semakin kuatnya sistem tersebut. (Tujuan lain, seperti menghindari pengangguran massal, dimiliki oleh pemerintah dan peneliti

universitas, tetapi belum tentu oleh perusahaan yang mengharapkan keuntungan dalam jangka pendek dari penyebaran AI seluas mungkin.) Untuk memperkuat kepentingan bersama ini dan mencapai tindakan terkoordinasi, ada organisasi dengan kekuatan penyelenggara, yang berarti, secara kasar, jika organisasi tersebut menyelenggarakan pertemuan, orang-orang menerima undangan untuk berpartisipasi.

Selain perkumpulan profesional, yang dapat menyatukan para peneliti AI, dan Kemitraan tentang AI, yang menggabungkan perusahaan dan lembaga nirlaba, penyelenggara utama adalah PBB (untuk pemerintah dan peneliti) dan Forum Ekonomi Dunia (untuk pemerintah dan perusahaan). Selain itu, G7 telah mengusulkan Panel Internasional tentang Kecerdasan Buatan, dengan harapan panel ini akan berkembang menjadi sesuatu seperti Panel Antarpemerintah PBB tentang Perubahan Iklim. Laporan-laporan yang terdengar penting berlipat ganda seperti kelinci.

Dengan semua aktivitas ini, apakah ada prospek kemajuan nyata dalam tata kelola yang terjadi? Mungkin mengejutkan, jawabannya adalah ya, setidaknya di pinggirannya. Banyak pemerintah di seluruh dunia melengkapi diri mereka dengan badan penasihat untuk membantu proses pengembangan peraturan; mungkin contoh yang paling menonjol adalah Kelompok Pakar Tingkat Tinggi Uni Eropa tentang Kecerdasan Buatan. Kesepakatan, aturan, dan standar mulai muncul untuk isu-isu seperti privasi pengguna, pertukaran data, dan menghindari bias rasial.

Pemerintah dan perusahaan bekerja keras untuk menyusun aturan untuk mobil otonom aturan yang pasti akan memiliki unsur lintas batas. Ada konsensus bahwa keputusan AI harus dapat dijelaskan jika sistem AI ingin dipercaya, dan konsensus tersebut sebagian telah diimplementasikan dalam undang-undang GDPR Uni Eropa. Di California, undang-undang baru melarang sistem AI untuk meniru manusia dalam keadaan tertentu. Dua hal terakhir ini kemampuan menjelaskan dan peniruan tentu memiliki pengaruh pada isu-isu keselamatan dan kontrol AI.

Saat ini, belum ada rekomendasi yang dapat diimplementasikan yang dapat diberikan kepada pemerintah atau organisasi lain yang mempertimbangkan isu pemeliharaan kontrol atas sistem AI. Regulasi seperti "sistem AI harus aman dan dapat dikendalikan" tidak akan memiliki bobot, karena istilah-istilah ini belum memiliki arti yang tepat dan karena belum ada metodologi rekayasa yang dikenal luas untuk memastikan keselamatan dan kemampuan pengendalian. Namun, mari kita bersikap optimis dan membayangkan bahwa, beberapa tahun ke depan, validitas pendekatan "terbukti bermanfaat" terhadap AI telah ditetapkan melalui analisis matematika dan realisasi praktis dalam bentuk aplikasi yang bermanfaat. Misalnya, kita mungkin memiliki asisten digital pribadi yang dapat kita percayai untuk menggunakan kartu kredit kita, menyaring panggilan dan email kita, dan mengelola keuangan kita karena mereka telah beradaptasi dengan preferensi individu kita dan tahu kapan boleh melanjutkan dan kapan lebih baik meminta bimbingan.

Mobil otonom kita mungkin telah mempelajari tata krama yang baik untuk berinteraksi satu sama lain dan dengan pengemudi manusia, dan robot rumah tangga kita seharusnya berinteraksi dengan lancar bahkan dengan balita yang paling bandel sekalipun. Mudah-

mudahan, tidak ada kucing yang dipanggang untuk makan malam dan tidak ada daging paus yang disajikan kepada anggota Partai Hijau.

Pada titik itu, mungkin dimungkinkan untuk menentukan desain perangkat lunak template yang harus dipatuhi oleh berbagai jenis aplikasi agar dapat dijual atau dihubungkan ke Internet, sama seperti aplikasi harus melewati sejumlah pengujian perangkat lunak sebelum dapat dijual di App Store Apple atau Google Play. Vendor perangkat lunak dapat mengusulkan templat tambahan, selama disertai bukti bahwa templat tersebut memenuhi persyaratan keamanan dan kontrolabilitas (yang saat itu sudah terdefinisi dengan baik). Akan ada mekanisme untuk melaporkan masalah dan untuk memperbarui sistem perangkat lunak yang menghasilkan perilaku yang tidak diinginkan. Akan masuk akal juga untuk membuat kode etik profesional seputar gagasan program AI yang terbukti aman dan untuk mengintegrasikan teorema dan metode yang sesuai ke dalam kurikulum bagi calon praktisi AI dan pembelajaran mesin.

Bagi pengamat Silicon Valley yang berpengalaman, ini mungkin terdengar agak naif. Regulasi dalam bentuk apa pun sangat ditentang di Silicon Valley. Sementara kita terbiasa dengan gagasan bahwa perusahaan farmasi harus menunjukkan keamanan dan kemanjuran (yang bermanfaat) melalui uji klinis sebelum mereka dapat merilis produk ke masyarakat umum, industri perangkat lunak beroperasi dengan seperangkat aturan yang berbeda yaitu, himpunan kosong. Sekelompok orang yang menenggak Red Bull di sebuah perusahaan perangkat lunak dapat meluncurkan produk atau peningkatan yang memengaruhi miliaran orang tanpa pengawasan pihak ketiga sama sekali. Namun, tak pelak lagi, industri teknologi harus mengakui bahwa produk-produknya penting; dan, jika produk-produk tersebut penting, maka penting juga agar produk-produk tersebut tidak memiliki efek berbahaya. Ini berarti akan ada aturan yang mengatur sifat interaksi dengan manusia, melarang desain yang, misalnya, secara konsisten memanipulasi preferensi atau menghasilkan perilaku adiktif. Saya tidak ragu bahwa transisi dari dunia yang tidak diatur ke dunia yang diatur akan menjadi transisi yang menyakitkan. Mari kita berharap hal itu tidak memerlukan bencana sebesar Chernobyl (atau lebih buruk) untuk mengatasi penolakan industri.

10.3 REGULASI AI: TANTANGAN PENJAHAT DAN PENCEGAHAN

Regulasi mungkin menyakitkan bagi industri perangkat lunak, tetapi akan tak tertahankan bagi Dr. Evil, yang merencanakan dominasi dunia di bunker bawah tanah rahasianya. Tidak diragukan lagi bahwa unsur-unsur kriminal, teroris, dan negara-negara nakal akan memiliki insentif untuk menghindari batasan apa pun pada desain mesin cerdas sehingga dapat digunakan untuk mengendalikan senjata atau untuk merancang dan melakukan kegiatan kriminal. Bahayanya bukanlah keberhasilan rencana jahat tersebut; melainkan kegagalannya karena kehilangan kendali atas sistem cerdas yang dirancang dengan buruk terutama yang dijiwai dengan tujuan jahat dan diberi akses ke senjata.

Ini bukan alasan untuk menghindari regulasi, lagipula kita memiliki hukum terhadap pembunuhan meskipun seringkali diabaikan. Namun, hal ini menciptakan masalah penegakan hukum yang sangat serius. Kita sudah kalah dalam pertempuran melawan malware dan

kejahatan siber. (Sebuah laporan baru-baru ini memperkirakan lebih dari dua miliar korban dan biaya tahunan sekitar Rp 6 kuadriliun.) Malware dalam bentuk program yang sangat cerdas akan jauh lebih sulit untuk dikalahkan.

Beberapa pihak, termasuk Nick Bostrom, telah mengusulkan agar kita menggunakan sistem AI supercerdas kita sendiri yang bermanfaat untuk mendeteksi dan menghancurkan sistem AI jahat atau yang berperilaku buruk lainnya. Tentu saja, kita harus menggunakan alat yang kita miliki, sambil meminimalkan dampak pada kebebasan pribadi, tetapi gambaran manusia yang bersembunyi di bunker, tak berdaya melawan kekuatan dahsyat yang dilepaskan oleh superintelijen yang saling bertempur, sama sekali tidak meyakinkan meskipun beberapa di antaranya berada di pihak kita. Akan jauh lebih baik untuk menemukan cara untuk membasmi AI jahat sejak dini.

Langkah pertama yang baik adalah kampanye internasional yang sukses dan terkoordinasi melawan kejahatan siber, termasuk perluasan Konvensi Budapest tentang Kejahatan Siber. Ini akan membentuk kerangka kerja organisasi untuk upaya di masa depan guna mencegah munculnya program AI yang tidak terkendali. Pada saat yang sama, hal ini akan menumbuhkan pemahaman budaya yang luas bahwa menciptakan program semacam itu, baik secara sengaja maupun tidak sengaja, dalam jangka panjang merupakan tindakan bunuh diri yang sebanding dengan menciptakan organisme pandemi.

10.4 KELEMAHAN DAN OTONOMI MANUSIA

Novel-novel E. M. Forster yang paling terkenal, termasuk *Howards End* dan *A Passage to India*, meneliti masyarakat Inggris dan sistem kelasnya di awal abad ke-20. Pada tahun 1909, ia menulis sebuah cerita fiksi ilmiah yang terkenal: “The Machine Stops.” Cerita ini luar biasa karena kejeliannya, termasuk penggambaran (apa yang sekarang kita sebut) Internet, konferensi video, iPad, kursus daring terbuka besar-besaran (MOOC), obesitas yang meluas, dan penghindaran kontak tatap muka. Mesin dalam judul tersebut adalah infrastruktur cerdas yang mencakup semua kebutuhan manusia. Manusia menjadi semakin bergantung padanya, tetapi mereka semakin sedikit memahami cara kerjanya. Pengetahuan teknik digantikan oleh mantra-mantra ritual yang akhirnya gagal untuk menghentikan kerusakan bertahap dari kinerja Mesin. Kuno, tokoh utama, melihat apa yang sedang terjadi tetapi tidak berdaya untuk menghentikannya:

> Tidakkah kau lihat… Bahwa kitalah yang sedang sekarat, dan bahwa di sini satu-satunya yang benar-benar hidup adalah Mesin? Kita menciptakan Mesin untuk melakukan kehendak kita, tetapi kita tidak dapat membuatnya melakukan kehendak kita sekarang. Ia telah merampas indra ruang dan indra sentuhan kita, ia telah mengaburkan setiap hubungan manusia, ia telah melumpuhkan tubuh dan kehendak kita. Kita hanya ada sebagai sel darah yang mengalir melalui arterinya, dan jika ia dapat bekerja tanpa kita, ia akan membiarkan kita mati. Oh, aku tidak punya obat atau, setidaknya, hanya satu untuk memberi tahu orang-

orang berulang kali bahwa aku telah melihat perbukitan Wessex seperti yang dilihat Aelfrid ketika ia mengalahkan bangsa Denmark.

Lebih dari seratus miliar orang telah hidup di Bumi. Mereka (kita) telah menghabiskan sekitar satu triliun tahun-orang untuk belajar dan mengajar, agar peradaban kita dapat berlanjut. Hingga saat ini, satu-satunya kemungkinan kelanjutannya adalah melalui penciptaan kembali dalam pikiran generasi baru. (Kertas memang bagus sebagai metode transmisi, tetapi kertas tidak berguna sampai pengetahuan yang tercatat di atasnya mencapai pikiran orang berikutnya.) Hal itu kini berubah: semakin banyak, dimungkinkan untuk menempatkan pengetahuan kita ke dalam mesin yang, dengan sendirinya, dapat menjalankan peradaban kita untuk kita.

Begitu insentif praktis untuk mewariskan peradaban kita kepada generasi berikutnya hilang, akan sangat sulit untuk membalikkan prosesnya. Satu triliun tahun pembelajaran kumulatif akan, dalam arti sebenarnya, hilang. Kita akan menjadi penumpang di kapal pesiar yang dijalankan oleh mesin, dalam pelayaran yang berlangsung selamanya persis seperti yang dibayangkan dalam film *WALL-E*.

Seorang konsekuensialis yang baik akan berkata, “Jelas ini adalah konsekuensi yang tidak diinginkan dari penggunaan otomatisasi yang berlebihan! Mesin yang dirancang dengan tepat tidak akan pernah melakukan ini!” Benar, tetapi pikirkan apa artinya ini. Mesin mungkin memahami bahwa otonomi dan kompetensi manusia adalah aspek penting dari bagaimana kita lebih suka menjalani hidup kita.

Mereka mungkin bersikeras bahwa manusia mempertahankan kendali dan tanggung jawab atas kesejahteraan mereka sendiri dengan kata lain, mesin akan mengatakan tidak. Tetapi kita manusia yang rabun dan malas mungkin tidak setuju. Ada tragedi kepemilikan bersama yang terjadi di sini: bagi setiap individu manusia, mungkin tampak sia-sia untuk terlibat dalam pembelajaran bertahun-tahun yang melelahkan untuk memperoleh pengetahuan dan keterampilan yang sudah dimiliki mesin; tetapi jika semua orang berpikir seperti itu, umat manusia, secara kolektif, akan kehilangan otonominya.

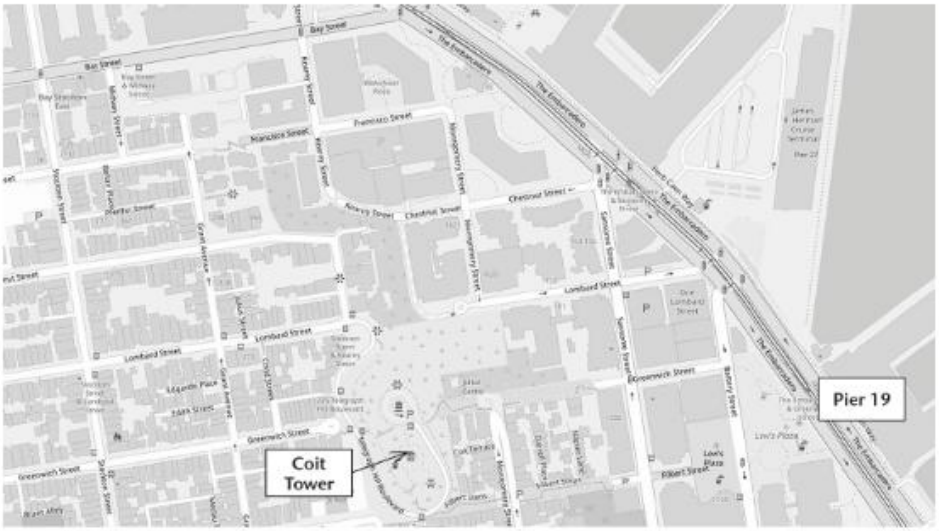
Solusi untuk masalah ini tampaknya bersifat budaya, bukan teknis. Kita akan membutuhkan gerakan budaya untuk membentuk kembali cita-cita dan preferensi kita menuju otonomi, kemandirian, dan kemampuan, serta menjauh dari kesenangan diri dan ketergantungan jika Anda mau, versi budaya modern dari etos militer Sparta kuno. Ini berarti rekayasa preferensi manusia dalam skala global bersamaan dengan perubahan radikal dalam cara kerja masyarakat kita. Untuk menghindari memperburuk situasi yang sudah buruk, kita mungkin membutuhkan bantuan mesin supercerdas, baik dalam membentuk solusi maupun dalam proses aktual mencapai keseimbangan bagi setiap individu.

Setiap orang tua yang memiliki anak kecil pasti familiar dengan proses ini. Setelah anak melewati tahap tak berdaya, pengasuhan membutuhkan keseimbangan yang terus berkembang antara melakukan segalanya untuk anak dan membiarkan anak sepenuhnya mengurus dirinya sendiri. Pada tahap tertentu, anak akan mengerti bahwa orang tua mampu mengikat tali sepatu anak tetapi memilih untuk tidak melakukannya. Apakah itu masa depan

umat manusia diperlakukan seperti anak kecil, selamanya, oleh mesin yang jauh lebih unggul? Saya rasa tidak. Pertama, anak-anak tidak dapat mematikan orang tua mereka. (Syukurlah!) Kita juga tidak akan menjadi hewan peliharaan atau hewan di kebun binatang. Sebenarnya tidak ada analog di dunia kita saat ini untuk hubungan yang akan kita miliki dengan mesin cerdas yang bermanfaat di masa depan. Masih harus dilihat bagaimana hasil akhirnya.

BAB 11
LOGIKA PENYELESAIAN MASALAH

Memilih tindakan dengan melihat ke depan dan mempertimbangkan hasil dari berbagai kemungkinan rangkaian tindakan adalah kemampuan mendasar bagi sistem cerdas. Ini adalah sesuatu yang dilakukan ponsel Anda setiap kali Anda memintanya untuk memberikan petunjuk arah. Gambar 11.1 menunjukkan contoh tipikal: menuju dari lokasi saat ini, Dermaga 19, ke tujuan, Menara Coit. Algoritma perlu mengetahui tindakan apa yang tersedia; biasanya, untuk navigasi peta, setiap tindakan melintasi segmen jalan yang menghubungkan dua persimpangan yang berdekatan.



Gambar 11.1 Peta sebagian San Francisco, menunjukkan lokasi awal di Dermaga 19 dan tujuan di Menara Coit.

Dalam contoh tersebut, dari Dermaga 19 hanya ada satu tindakan: belok kanan dan berkendara di sepanjang Embarcadero ke persimpangan berikutnya. Kemudian ada pilihan: melanjutkan atau belok tajam ke kiri ke Jalan Battery. Algoritma secara sistematis mengeksplorasi semua kemungkinan ini sampai akhirnya menemukan rute. Biasanya kita menambahkan sedikit panduan akal sehat, seperti preferensi untuk menjelajahi jalan-jalan yang menuju ke tujuan daripada menjauhinya. Dengan panduan ini dan beberapa trik lainnya, algoritma dapat menemukan solusi optimal dengan sangat cepat biasanya dalam beberapa milidetik, bahkan untuk perjalanan lintas negara.

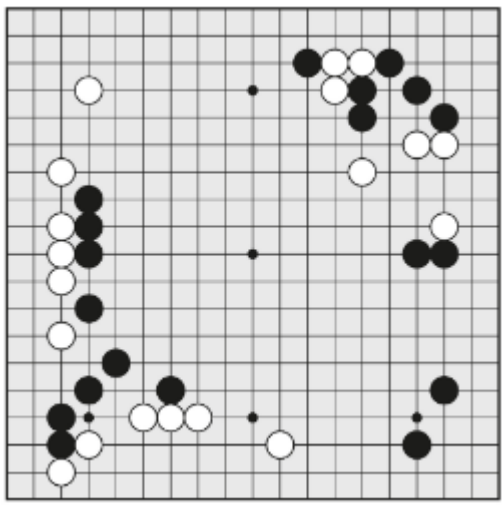
Mencari rute di peta adalah contoh yang alami dan familiar, tetapi mungkin sedikit menyesatkan karena jumlah lokasi yang berbeda sangat kecil. Di Amerika Serikat, misalnya, hanya ada sekitar sepuluh juta persimpangan. Itu mungkin tampak seperti angka yang besar, tetapi sangat kecil dibandingkan dengan jumlah negara bagian yang berbeda dalam teka-teki

15. Teka-teki 15 adalah mainan dengan kisi empat kali empat yang berisi lima belas ubin bernomor dan ruang kosong.

Tujuannya adalah untuk memindahkan ubin untuk mencapai konfigurasi tujuan, seperti memiliki semua ubin dalam urutan numerik. Teka-teki 15 memiliki sekitar sepuluh triliun negara bagian (satu juta kali lebih besar dari Amerika Serikat!); teka-teki 24 memiliki sekitar delapan triliun triliun negara bagian. Ini adalah contoh dari apa yang disebut matematikawan sebagai kompleksitas kombinatorial ledakan cepat dalam jumlah kombinasi seiring dengan meningkatnya jumlah "bagian yang bergerak" dari suatu masalah. Kembali ke peta Amerika Serikat: jika sebuah perusahaan truk ingin mengoptimalkan pergerakan seratus truknya di seluruh Amerika Serikat, jumlah negara bagian yang mungkin dipertimbangkan adalah sepuluh juta pangkat seratus (yaitu, 10700).

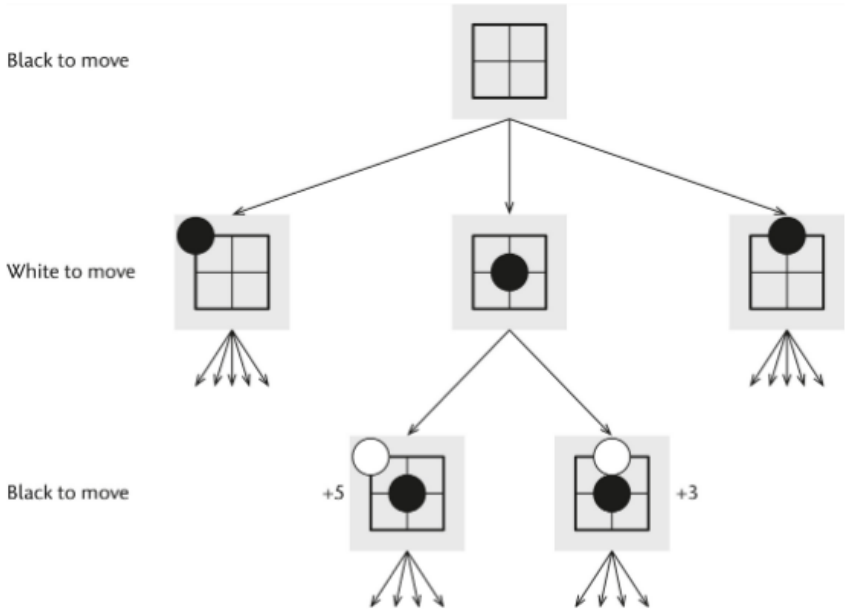
11.1 MENYERAH PADA KEPUTUSAN RASIONAL

Banyak permainan memiliki sifat kompleksitas kombinatorial ini, termasuk catur, dam, backgammon, dan Go. Karena aturan Go sederhana dan elegan (gambar 11.2), saya akan menggunakannya sebagai contoh. Tujuannya cukup jelas: memenangkan permainan dengan mengelilingi lebih banyak wilayah daripada lawan Anda. Tindakan yang mungkin juga jelas: meletakkan batu di lokasi kosong. Sama seperti navigasi di peta, cara yang jelas untuk memutuskan apa yang harus dilakukan adalah dengan membayangkan berbagai kemungkinan masa depan yang dihasilkan dari berbagai urutan tindakan dan memilih yang terbaik. Anda bertanya, “Jika saya melakukan ini, apa yang mungkin dilakukan lawan saya? Dan apa yang saya lakukan selanjutnya?” Ide ini diilustrasikan pada gambar 11.3 untuk Go 3×3. Bahkan untuk Go 3×3, saya hanya dapat menunjukkan sebagian kecil dari pohon kemungkinan masa depan, tetapi saya harap idenya cukup jelas. Memang, cara pengambilan keputusan ini tampaknya hanya akal sehat yang sederhana.



Gambar 11.2 Papan Go, di tengah-tengah Game 5 final Piala LG 2002 antara Lee Sedol (hitam) dan Choe Myeong-hun (putih). Hitam dan Putih bergantian menempatkan satu batu di lokasi kosong mana pun di papan. Di sini, giliran Hitam untuk bergerak dan ada 343

kemungkinan gerakan. Masing-masing pihak berusaha untuk mengelilingi wilayah sebanyak mungkin. Misalnya, Putih memiliki peluang bagus untuk memenangkan wilayah di tepi kiri dan di sisi kiri tepi bawah, sementara Hitam dapat memenangkan wilayah di sudut kanan atas dan kanan bawah. Konsep kunci dalam Go adalah kelompok yaitu, sekumpulan batu dengan warna yang sama yang terhubung satu sama lain melalui kedekatan vertikal atau horizontal. Sebuah kelompok tetap hidup selama ada setidaknya satu ruang kosong di sebelahnya; jika sepenuhnya dikelilingi, tanpa ruang kosong, kelompok tersebut mati dan dikeluarkan dari papan.



Gambar 11.3 Bagian dari pohon permainan untuk Go 3×3. Dimulai dari keadaan awal yang kosong, kadang-kadang disebut akar pohon, Hitam dapat memilih salah satu dari tiga kemungkinan langkah yang berbeda. (Yang lainnya simetris dengan ini.) Kemudian giliran Putih untuk bergerak. Jika Hitam memilih untuk bermain di tengah, Putih memiliki dua langkah berbeda sudut atau sisi maka Hitam akan mendapat kesempatan bermain lagi. Dengan membayangkan kemungkinan masa depan ini, Hitam dapat memilih langkah mana yang akan dimainkan dalam keadaan awal. Jika Hitam tidak dapat mengikuti setiap kemungkinan jalur permainan hingga akhir permainan, maka fungsi evaluasi dapat digunakan untuk memperkirakan seberapa baik posisi di daun pohon. Di sini, fungsi evaluasi memberikan +5 dan +3 kepada dua daun.

Masalahnya adalah Go memiliki lebih dari 10^{170} kemungkinan posisi untuk papan 19×19 penuh. Sementara menemukan rute terpendek yang dijamin pada peta relatif mudah, menemukan kemenangan yang dijamin dalam Go sama sekali tidak mungkin. Sekalipun algoritma tersebut mempertimbangkan selama miliaran tahun ke depan, ia hanya dapat menjelajahi sebagian kecil dari keseluruhan pohon kemungkinan. Hal ini menimbulkan dua pertanyaan. Pertama, bagian pohon mana yang harus dieksplorasi oleh program? Dan kedua,

langkah apa yang harus dilakukan oleh program, mengingat pohon parsial yang telah dieksplorasinya?

Untuk menjawab pertanyaan kedua terlebih dahulu: ide dasar yang digunakan oleh hampir semua program lookahead adalah untuk menetapkan nilai perkiraan pada "daun" pohon keadaan yang paling jauh di masa depan dan kemudian "bekerja mundur" untuk mengetahui seberapa baik pilihan di akarnya. Misalnya, melihat dua posisi di bagian bawah gambar 11.3, seseorang mungkin menebak nilai +5 (dari sudut pandang Hitam) untuk posisi di sebelah kiri dan +3 untuk posisi di sebelah kanan, karena batu Putih di sudut jauh lebih rentan daripada yang di samping.

Jika nilai-nilai ini benar, maka Hitam dapat mengharapkan bahwa Putih akan bermain di samping, yang mengarah ke posisi sebelah kanan; Oleh karena itu, tampaknya masuk akal untuk memberikan nilai +3 pada langkah awal Hitam di tengah. Dengan sedikit variasi, ini adalah skema yang digunakan oleh program catur Arthur Samuel untuk mengalahkan penciptanya pada tahun 1955, oleh Deep Blue untuk mengalahkan juara catur dunia saat itu, Garry Kasparov, pada tahun 1997, dan oleh AlphaGo untuk mengalahkan mantan juara dunia Go Lee Sedol pada tahun 2016. Untuk Deep Blue, manusia menulis bagian program yang mengevaluasi posisi di daun pohon, sebagian besar berdasarkan pengetahuan mereka tentang catur. Untuk program Samuel dan untuk AlphaGo, program tersebut mempelajarinya dari ribuan atau jutaan permainan latihan.

Pertanyaan pertama bagian pohon mana yang harus dieksplorasi oleh program? Adalah contoh dari salah satu pertanyaan terpenting dalam AI: Komputasi apa yang harus dilakukan agen? Untuk program permainan, ini sangat penting karena mereka hanya memiliki alokasi waktu yang kecil dan tetap, dan menggunakannya untuk komputasi yang tidak berguna adalah cara pasti untuk kalah.

Bagi manusia dan agen lain yang beroperasi di dunia nyata, hal ini bahkan lebih penting karena dunia nyata jauh lebih kompleks: kecuali dipilih dengan baik, tidak ada jumlah komputasi yang akan memberikan dampak sekecil apa pun pada masalah memutuskan apa yang harus dilakukan. Jika Anda sedang mengemudi dan seekor rusa besar berjalan ke tengah jalan, tidak ada gunanya memikirkan apakah akan menukar euro dengan pound atau apakah Hitam harus melakukan langkah pertamanya di tengah papan Go.

Kemampuan manusia untuk mengelola aktivitas komputasi mereka sehingga keputusan yang masuk akal dibuat dengan cukup cepat setidaknya sama luar biasanya dengan kemampuan mereka untuk memahami dan bernalar dengan benar. Dan tampaknya ini adalah sesuatu yang kita peroleh secara alami dan tanpa usaha: ketika ayah saya mengajari saya bermain catur, dia mengajari saya aturannya, tetapi dia juga tidak mengajari saya algoritma cerdas tertentu untuk memilih bagian mana dari pohon permainan yang akan dieksplorasi dan bagian mana yang akan diabaikan.

Bagaimana ini bisa terjadi? Atas dasar apa kita dapat mengarahkan pikiran kita? Jawabannya adalah bahwa komputasi memiliki nilai sejauh ia dapat meningkatkan kualitas keputusan Anda. Proses pemilihan komputasi disebut metareasoning, yang berarti penalaran tentang penalaran. Sama seperti tindakan dapat dipilih secara rasional, berdasarkan nilai yang

diharapkan, demikian pula komputasi. Ini disebut metareasoning rasional. Ide dasarnya sangat sederhana:

> Lakukan komputasi yang akan memberikan peningkatan kualitas keputusan yang diharapkan tertinggi, dan berhenti ketika biaya (dalam hal waktu) melebihi peningkatan yang diharapkan.

Itu saja. Tidak perlu algoritma yang rumit! Prinsip sederhana ini menghasilkan perilaku komputasi yang efektif dalam berbagai masalah, termasuk catur dan Go. Tampaknya otak kita menerapkan sesuatu yang serupa, yang menjelaskan mengapa kita tidak perlu mempelajari algoritma baru yang spesifik untuk permainan tertentu untuk berpikir setiap kali kita belajar memainkan permainan baru.

Menjelajahi pohon kemungkinan yang membentang ke masa depan dari keadaan saat ini bukanlah satu-satunya cara untuk mencapai keputusan, tentu saja. Seringkali, lebih masuk akal untuk bekerja mundur dari tujuan. Misalnya, keberadaan rusa di jalan menunjukkan tujuan untuk menghindari menabrak rusa, yang pada gilirannya menunjukkan tiga tindakan yang mungkin: berbelok ke kiri, berbelok ke kanan, atau mengerem mendadak.

Ini tidak menunjukkan tindakan menukar euro dengan pound atau menempatkan batu hitam di tengah. Dengan demikian, tujuan memiliki efek fokus yang luar biasa pada pemikiran seseorang. Tidak ada program permainan saat ini yang memanfaatkan ide ini; bahkan, mereka biasanya mempertimbangkan semua tindakan legal yang mungkin. Inilah salah satu (dari sekian banyak) alasan mengapa saya tidak khawatir AlphaZero akan menguasai dunia.

11.2 PERENCANAAN HIERARKIS: OTAK VS ALPHAGO

Mari kita misalkan Anda telah memutuskan untuk melakukan langkah spesifik di papan Go. Bagus! Sekarang Anda harus benar-benar melakukannya. Di dunia nyata, ini melibatkan meraih ke dalam mangkuk berisi batu yang belum dimainkan untuk mengambil batu, menggerakkan tangan Anda di atas lokasi yang dimaksud, dan meletakkan batu dengan rapi di tempat tersebut, baik dengan tenang atau tegas sesuai dengan etiket Go.

Masing-masing tahap ini, pada gilirannya, terdiri dari tarian kompleks persepsi dan perintah kontrol motorik yang melibatkan otot dan saraf tangan, lengan, bahu, dan mata. Dan saat meraih batu, Anda memastikan bagian tubuh Anda yang lain tidak jatuh karena pergeseran pusat gravitasi Anda. Fakta bahwa Anda mungkin tidak secara sadar menyadari memilih tindakan ini tidak berarti bahwa tindakan tersebut tidak dipilih oleh otak Anda. Misalnya, mungkin ada banyak batu di dalam mangkuk, tetapi "tangan" Anda sebenarnya, otak Anda yang memproses informasi sensorik tetap harus memilih salah satu untuk diambil.

Hampir semua yang kita lakukan seperti ini. Saat mengemudi, kita mungkin memilih untuk berpindah jalur ke kiri; tetapi tindakan ini melibatkan melihat ke kaca spion dan ke belakang bahu, mungkin menyesuaikan kecepatan, dan menggerakkan setir sambil memantau kemajuan hingga manuver selesai.

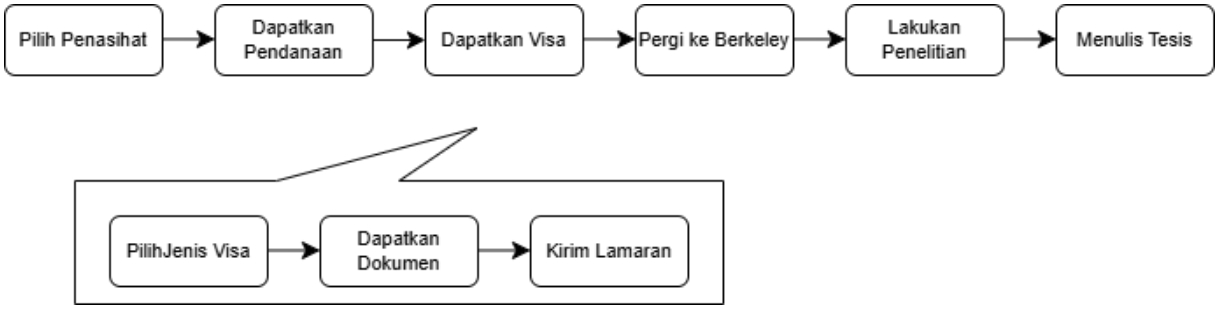
Dalam percakapan, respons rutin seperti "Oke, izinkan saya memeriksa kalender saya dan akan menghubungi Anda kembali" melibatkan pengucapan empat belas suku kata, yang masing-masing membutuhkan ratusan perintah kontrol motorik yang terkoordinasi dengan tepat ke otot-otot lidah, bibir, rahang, tenggorokan, dan alat pernapasan. Untuk bahasa ibu Anda, proses ini otomatis; ini sangat mirip dengan gagasan menjalankan subrutin dalam program komputer (lihat halaman ini). Fakta bahwa rangkaian tindakan kompleks dapat menjadi rutin dan otomatis, sehingga berfungsi sebagai tindakan tunggal dalam proses yang lebih kompleks, benar-benar mendasar bagi kognisi manusia. Mengucapkan kata-kata dalam bahasa yang kurang familiar mungkin menanyakan arah ke Szczebrzeszyn di Polandia adalah pengingat yang berguna bahwa ada suatu masa dalam hidup Anda ketika membaca dan mengucapkan kata-kata adalah tugas yang sulit yang membutuhkan usaha mental dan banyak latihan.

Jadi, masalah sebenarnya yang dihadapi otak Anda bukanlah memilih langkah di papan Go, tetapi mengirimkan perintah kontrol motorik ke otot Anda. Jika kita mengalihkan perhatian kita dari tingkat langkah Go ke tingkat perintah kontrol motorik, masalahnya terlihat sangat berbeda. Secara kasar, otak Anda dapat mengirimkan perintah setiap seratus milidetik. Kita memiliki sekitar enam ratus otot, jadi itu adalah maksimum teoritis sekitar enam ribu aktuasi per detik, dua puluh juta per jam, dua ratus miliar per tahun, dua puluh triliun per masa hidup. Gunakanlah dengan bijak!

Sekarang, misalkan kita mencoba menerapkan algoritma seperti AlphaZero untuk menyelesaikan masalah pengambilan keputusan pada tingkat ini. Dalam Go, AlphaZero melihat ke depan mungkin lima puluh langkah. Tetapi lima puluh langkah perintah kontrol motorik hanya membawa Anda beberapa detik ke depan! Tidak cukup untuk dua puluh juta perintah kontrol motorik dalam permainan Go selama satu jam, dan tentu saja tidak cukup untuk triliunan (1.000.000.000.000) langkah yang terlibat dalam menyelesaikan gelar PhD. Jadi, meskipun AlphaZero melihat lebih jauh ke depan dalam permainan Go daripada manusia mana pun, kemampuan itu tampaknya tidak membantu di dunia nyata. Itu adalah jenis pandangan ke depan yang salah.

Tentu saja, saya tidak mengatakan bahwa melakukan PhD benar-benar membutuhkan perencanaan triliunan gerakan otot sebelumnya. Hanya rencana yang cukup abstrak yang dibuat pada awalnya mungkin memilih Berkeley atau tempat lain, memilih pembimbing PhD atau topik penelitian, mengajukan pendanaan, mendapatkan visa pelajar, bepergian ke kota yang dipilih, melakukan beberapa penelitian, dan sebagainya. Untuk membuat pilihan Anda, Anda cukup berpikir, tentang hal-hal yang tepat, sehingga keputusan menjadi jelas. Jika kelayakan beberapa langkah abstrak seperti mendapatkan visa tidak jelas, Anda melakukan lebih banyak pemikiran dan mungkin pengumpulan informasi, yang berarti membuat rencana lebih konkret dalam aspek-aspek tertentu: mungkin memilih jenis visa yang memenuhi syarat, mengumpulkan dokumen yang diperlukan, dan mengirimkan aplikasi. Gambar 11.4 menunjukkan rencana abstrak dan penyempurnaan langkah Mendapatkan Visa menjadi subrencana tiga langkah. Ketika tiba waktunya untuk mulai melaksanakan rencana tersebut,

langkah-langkah awalnya harus disempurnakan hingga ke tingkat paling dasar sehingga tubuh Anda dapat melaksanakannya.



Gambar 11.4 Rencana abstrak untuk seorang mahasiswa luar negeri yang memilih untuk mengambil gelar PhD di Berkeley. Langkah GetVisa, yang kelayakannya tidak pasti, telah diperluas menjadi rencana abstrak tersendiri.

AlphaGo sama sekali tidak dapat melakukan pemikiran semacam ini: satu-satunya tindakan yang pernah dipertimbangkannya adalah tindakan primitif yang terjadi secara berurutan dari keadaan awal. Ia tidak memiliki gagasan tentang rencana abstrak. Mencoba menerapkan AlphaGo di dunia nyata seperti mencoba menulis novel dengan bertanya-tanya apakah huruf pertama harus A, B, C, dan seterusnya.

Pada tahun 1962, Herbert Simon menekankan pentingnya organisasi hierarkis dalam sebuah makalah terkenal, "Arsitektur Kompleksitas." Para peneliti AI sejak awal tahun 1970-an telah mengembangkan berbagai metode yang membangun dan menyempurnakan rencana yang terorganisasi secara hierarkis. Beberapa sistem yang dihasilkan mampu membangun rencana dengan puluhan juta langkah misalnya, untuk mengatur aktivitas manufaktur di pabrik besar.

Sekarang kita memiliki pemahaman teoritis yang cukup baik tentang makna tindakan abstrak yaitu, bagaimana mendefinisikan efek yang ditimbulkannya pada dunia. Pertimbangkan, misalnya, tindakan abstrak GoToBerkeley pada gambar 11.4. Tindakan ini dapat diimplementasikan dengan berbagai cara, yang masing-masing menghasilkan efek yang berbeda pada dunia: Anda dapat berlayar ke sana, menyelinap ke kapal, terbang ke Kanada dan berjalan kaki menyeberangi perbatasan, menyewa jet pribadi, dan sebagainya.

Tetapi Anda tidak perlu mempertimbangkan pilihan-pilihan ini untuk saat ini. Selama Anda yakin ada cara untuk melakukannya yang tidak menghabiskan begitu banyak waktu dan uang atau menimbulkan begitu banyak risiko sehingga membahayakan bagian rencana lainnya, Anda cukup memasukkan langkah abstrak GoToBerkeley ke dalam rencana dan yakin bahwa rencana tersebut akan berhasil. Dengan cara ini, kita dapat membangun rencana tingkat tinggi yang pada akhirnya akan berubah menjadi miliaran atau triliunan langkah primitif tanpa pernah khawatir tentang apa langkah-langkah tersebut sampai tiba saatnya untuk benar-benar melakukannya.

Tentu saja, semua ini tidak mungkin tanpa hierarki. Tanpa tindakan tingkat tinggi seperti mendapatkan visa dan menulis tesis, kita tidak dapat membuat rencana abstrak untuk

mendapatkan gelar PhD; tanpa tindakan tingkat yang lebih tinggi lagi seperti mendapatkan gelar PhD dan memulai perusahaan, kita tidak dapat merencanakan untuk mendapatkan gelar PhD dan kemudian memulai perusahaan. Di dunia nyata, kita akan tersesat tanpa perpustakaan tindakan yang luas pada puluhan tingkat abstraksi. (Dalam permainan Go, tidak ada hierarki tindakan yang jelas, sehingga sebagian besar dari kita tersesat.) Namun, saat ini, semua metode yang ada untuk perencanaan hierarkis bergantung pada hierarki tindakan abstrak dan konkret yang dihasilkan manusia; kita belum memahami bagaimana hierarki tersebut dapat dipelajari dari pengalaman.

BAB 12
PENGETAHUAN DAN LOGIKA

Logika adalah studi tentang penalaran dengan pengetahuan yang pasti. Logika bersifat umum sepenuhnya dalam hal materi pokok yaitu, pengetahuan tersebut dapat tentang apa saja. Oleh karena itu, logika merupakan bagian yang sangat penting dari pemahaman kita tentang kecerdasan tujuan umum.

Persyaratan utama logika adalah bahasa formal dengan makna yang tepat untuk kalimat-kalimat dalam bahasa tersebut, sehingga terdapat proses yang tidak ambigu untuk menentukan apakah suatu kalimat benar atau salah dalam situasi tertentu. Itu saja. Setelah kita memilikinya, kita dapat menulis algoritma penalaran yang baik yang menghasilkan kalimat baru dari kalimat yang sudah diketahui. Kalimat-kalimat baru tersebut dijamin mengikuti dari kalimat yang sudah diketahui sistem, artinya kalimat-kalimat baru tersebut pasti benar dalam situasi apa pun di mana kalimat aslinya benar. Hal ini memungkinkan mesin untuk menjawab pertanyaan, membuktikan teorema matematika, atau menyusun rencana yang dijamin berhasil.

Aljabar sekolah menengah memberikan contoh yang baik (meskipun contoh yang mungkin membangkitkan kenangan menyakitkan). Bahasa formal mencakup kalimat-kalimat seperti $4x + 1 = 2y - 5$. Kalimat ini benar dalam situasi di mana $x = 5$ dan $y = 13$, dan salah ketika $x = 5$ dan $y = 6$. Dari kalimat ini, seseorang dapat menurunkan kalimat lain seperti $y = 2x + 3$, dan setiap kali kalimat pertama benar, kalimat kedua dijamin juga benar.

Ide inti logika, yang dikembangkan secara independen di India kuno, Cina, dan Yunani, adalah bahwa gagasan yang sama tentang makna yang tepat dan penalaran yang benar dapat diterapkan pada kalimat tentang apa pun, bukan hanya angka. Contoh kanonik dimulai dengan "Socrates adalah manusia" dan "Semua manusia fana" dan menghasilkan "Socrates fana." Penurunan ini benar-benar formal dalam arti bahwa ia tidak bergantung pada informasi lebih lanjut tentang siapa Socrates atau apa arti manusia dan fana. Fakta bahwa penalaran logis benar-benar formal berarti bahwa dimungkinkan untuk menulis algoritma yang melakukannya.

12.1 LOGIKA PROPOSISIONAL

Untuk tujuan kita dalam memahami kemampuan dan prospek AI, ada dua jenis logika penting yang benar-benar relevan: logika proposisional dan logika orde pertama. Perbedaan antara keduanya sangat mendasar untuk memahami situasi terkini dalam AI dan bagaimana kemungkinan perkembangannya.

Mari kita mulai dengan logika proposisional, yang lebih sederhana dari keduanya. Kalimat hanya terdiri dari dua jenis hal: simbol yang mewakili proposisi yang dapat benar atau salah, dan penghubung logis seperti dan, atau, tidak, dan jika...maka. (Kita akan melihat contohnya sebentar lagi.) Penghubung logis ini kadang-kadang disebut Boolean, diambil dari nama George Boole, seorang ahli logika abad ke-19 yang menghidupkan kembali bidangnya

dengan ide-ide matematika baru. Mereka sama persis dengan gerbang logika yang digunakan dalam chip komputer.

Algoritma praktis untuk penalaran dalam logika proposisional telah dikenal sejak awal tahun 1960-an. Meskipun tugas penalaran umum mungkin memerlukan waktu eksponensial dalam kasus terburuk, algoritma penalaran proposisional modern menangani masalah dengan jutaan simbol proposisi dan puluhan juta kalimat. Algoritma ini merupakan alat inti untuk membangun rencana logistik yang terjamin, memverifikasi desain chip sebelum diproduksi, dan memeriksa kebenaran aplikasi perangkat lunak dan protokol keamanan sebelum diterapkan.

Hal yang menakjubkan adalah bahwa satu algoritma algoritma penalaran untuk logika proposisional menyelesaikan semua tugas ini setelah dirumuskan sebagai tugas penalaran. Jelas, ini merupakan langkah menuju tujuan generalisasi dalam sistem cerdas. Sayangnya, ini bukan langkah yang besar karena bahasa logika proposisional tidak terlalu ekspresif. Mari kita lihat apa artinya ini dalam praktik ketika kita mencoba untuk mengekspresikan aturan dasar untuk langkah legal dalam Go: "Pemain yang gilirannya untuk bergerak dapat memainkan batu di persimpangan mana pun yang tidak ditempati." Langkah pertama adalah memutuskan simbol proposisi apa yang akan digunakan untuk membicarakan langkah-langkah Go dan posisi papan Go.

Proposisi mendasar yang penting adalah apakah batu dengan warna tertentu berada di lokasi tertentu pada waktu tertentu. Jadi, kita akan membutuhkan simbol seperti $Batu\_Putih\_Di\_5\_5\_Pada\_Langkah\_38$ dan $Batu\_Hitam\_Di\_5\_5\_Pada\_Langkah\_38$. (Ingatlah bahwa, seperti halnya manusia, fana, dan Socrates, algoritma penalaran tidak perlu mengetahui arti simbol-simbol tersebut.) Maka kondisi logis agar Putih dapat bermain di persimpangan 5,5 pada langkah ke-38 adalah:

$$(bukan\ Batu\ Putih\ di\ 5{,}5\ pada\ Langkah\ ke-38)\ dan$$
$$(bukan\ Batu\ Hitam\ di\ 5{,}5\ pada\ Langkah\ ke-38)$$

Dengan kata lain: tidak ada batu putih dan tidak ada batu hitam. Itu tampaknya cukup sederhana. Sayangnya, dalam logika proposisional, hal itu harus ditulis secara terpisah untuk setiap lokasi dan untuk setiap langkah dalam permainan. Karena ada 361 lokasi dan sekitar 300 langkah per permainan, ini berarti lebih dari 100.000 salinan aturan! Untuk aturan yang berkaitan dengan penangkapan dan pengulangan, yang melibatkan banyak batu dan lokasi, situasinya bahkan lebih buruk, dan kita dengan cepat mengisi jutaan halaman.

Dunia nyata, jelas, jauh lebih besar daripada papan Go: ada jauh lebih dari 361 lokasi dan 300 langkah waktu, dan ada banyak jenis benda selain batu; jadi, prospek menggunakan bahasa proposisional untuk pengetahuan tentang dunia nyata benar-benar tidak ada harapan. Bukan hanya ukuran buku aturan yang konyol yang menjadi masalah: tetapi juga jumlah pengalaman yang konyol yang dibutuhkan sistem pembelajaran untuk memperoleh aturan dari contoh. Sementara manusia hanya membutuhkan satu atau dua contoh untuk mendapatkan ide dasar tentang menempatkan batu, menangkap batu, dan sebagainya, sistem

cerdas yang berbasis pada logika proposisional harus diperlihatkan contoh pergerakan dan penangkapan secara terpisah untuk setiap lokasi dan langkah waktu.

Sistem tidak dapat menggeneralisasi dari beberapa contoh, seperti yang dilakukan manusia, karena tidak ada cara untuk mengekspresikan aturan umum. Batasan ini berlaku tidak hanya untuk sistem yang berbasis pada logika proposisional tetapi juga untuk sistem apa pun dengan kekuatan ekspresif yang sebanding. Itu termasuk jaringan Bayesian, yang merupakan sepupu probabilistik dari logika proposisional, dan jaringan saraf, yang merupakan dasar dari pendekatan "pembelajaran mendalam" untuk AI.

12.2 LOGIKA ORDE PERTAMA

Jadi, pertanyaan selanjutnya adalah, dapatkah kita merancang bahasa logika yang lebih ekspresif? Kita menginginkan bahasa yang memungkinkan untuk menyampaikan aturan Go ke sistem berbasis pengetahuan dengan cara berikut:

> untuk semua lokasi di papan, dan untuk semua langkah waktu, berikut aturannya...

Logika orde pertama, yang diperkenalkan oleh matematikawan Jerman Gottlob Frege pada tahun 1879, memungkinkan seseorang untuk menulis aturan dengan cara ini. Perbedaan utama antara logika proposisional dan logika orde pertama adalah ini: sedangkan logika proposisional mengasumsikan dunia terdiri dari proposisi yang benar atau salah, logika orde pertama mengasumsikan dunia terdiri dari objek yang dapat dihubungkan satu sama lain dengan berbagai cara. Misalnya, mungkin ada lokasi yang berdekatan satu sama lain, waktu yang berurutan, batu yang berada di lokasi pada waktu tertentu, dan gerakan yang sah pada waktu tertentu. Logika orde pertama memungkinkan seseorang untuk menyatakan bahwa beberapa properti benar untuk semua objek di dunia; Jadi, dapat ditulis:

> untuk semua langkah waktu t, dan untuk semua lokasi l, dan untuk semua warna c,
> jika giliran c untuk bergerak pada waktu t dan l tidak ditempati pada waktu t,
> maka sah bagi c untuk memainkan batu di lokasi l pada waktu t.

Dengan beberapa peringatan tambahan dan beberapa kalimat tambahan yang mendefinisikan lokasi papan, dua warna, dan apa arti tidak ditempati, kita memiliki awal dari aturan lengkap Go. Aturan-aturan tersebut membutuhkan ruang yang hampir sama dalam logika orde pertama seperti dalam bahasa Inggris.

Pengembangan pemrograman logika pada akhir tahun 1970-an menyediakan teknologi yang elegan dan efisien untuk penalaran logis yang diwujudkan dalam bahasa pemrograman yang disebut Prolog. Ilmuwan komputer menemukan cara untuk membuat penalaran logis dalam Prolog berjalan pada jutaan langkah penalaran per detik, membuat banyak aplikasi logika menjadi praktis. Pada tahun 1982, pemerintah Jepang mengumumkan investasi besar

dalam AI berbasis Prolog yang disebut proyek Generasi Kelima, dan Amerika Serikat dan Inggris menanggapi dengan upaya serupa. Sayangnya, proyek Generasi Kelima dan proyek-proyek sejenisnya kehabisan tenaga pada akhir tahun 1980-an dan awal tahun 1990-an, sebagian karena ketidakmampuan logika untuk menangani informasi yang tidak pasti. Mereka mewakili apa yang kemudian menjadi istilah peyoratif: AI Kuno yang Baik, atau GOFAI. Menjadi tren untuk mengabaikan logika sebagai hal yang tidak relevan dengan AI; memang, banyak peneliti AI yang sekarang bekerja di bidang pembelajaran mendalam tidak tahu apa pun tentang logika. Tren ini tampaknya akan memudar: jika Anda menerima bahwa dunia memiliki objek-objek yang saling terkait dalam berbagai cara, maka logika orde pertama akan relevan, karena menyediakan matematika dasar objek dan relasi. Pandangan ini juga dianut oleh Demis Hassabis, CEO Google DeepMind:

> Anda dapat menganggap pembelajaran mendalam seperti yang ada saat ini sebagai padanan di otak untuk korteks sensorik kita: korteks visual atau korteks pendengaran kita. Tetapi, tentu saja, kecerdasan sejati jauh lebih dari itu, Anda harus menggabungkannya kembali ke dalam pemikiran tingkat tinggi dan penalaran simbolik, banyak hal yang coba ditangani oleh AI klasik pada tahun 80-an.
>
> . . . Kami ingin [sistem-sistem ini] berkembang hingga mencapai tingkat penalaran simbolis ini matematika, bahasa, dan logika. Jadi, itu adalah bagian besar dari pekerjaan kami.

Oleh karena itu, salah satu pelajaran terpenting dari tiga puluh tahun pertama penelitian AI adalah bahwa program yang mengetahui sesuatu, dalam arti yang bermanfaat, akan membutuhkan kapasitas untuk representasi dan penalaran yang setidaknya sebanding dengan yang ditawarkan oleh logika orde pertama. Sampai saat ini, kita belum mengetahui bentuk pastinya: mungkin akan dimasukkan ke dalam sistem penalaran probabilistik, ke dalam sistem pembelajaran mendalam, atau ke dalam beberapa desain hibrida yang masih akan ditemukan.

BAB 13
KETIDAKPASTIAN DAN PROBABILITAS

Sedangkan logika menyediakan dasar umum untuk penalaran dengan pengetahuan pasti, teori probabilitas mencakup penalaran dengan informasi yang tidak pasti (di mana pengetahuan pasti adalah kasus khusus). Ketidakpastian adalah situasi epistemik normal seorang agen di dunia nyata. Meskipun ide-ide dasar probabilitas dikembangkan pada abad ketujuh belas, baru-baru ini dimungkinkan untuk merepresentasikan dan menalar dengan model probabilitas besar secara formal.

13.1 DASAR-DASAR PROBABILITAS

Teori probabilitas memiliki kesamaan dengan logika dalam gagasan bahwa ada kemungkinan dunia. Biasanya dimulai dengan mendefinisikan apa itu misalnya, jika saya melempar satu dadu enam sisi biasa, ada enam dunia (kadang-kadang disebut hasil): 1, 2, 3, 4, 5, 6. Tepat satu dari mereka akan terjadi, tetapi secara apriori saya tidak tahu yang mana.

Teori probabilitas mengasumsikan bahwa dimungkinkan untuk memberikan probabilitas pada setiap dunia; untuk lemparan dadu saya, saya akan memberikan ⅙ pada setiap dunia. (Probabilitas ini kebetulan sama, tetapi tidak harus demikian; satu-satunya syarat adalah probabilitasnya harus berjumlah 1.) Sekarang saya dapat mengajukan pertanyaan seperti “Berapa probabilitas saya akan mendapatkan angka genap?” Untuk menemukan ini, saya cukup menjumlahkan probabilitas untuk tiga dunia di mana angkanya genap: ⅙ + ⅙ + ⅙ = ½.

Sangat mudah juga untuk mempertimbangkan bukti baru. Misalkan seorang peramal memberi tahu saya bahwa hasil lemparan adalah bilangan prima (yaitu, 2, 3, atau 5). Ini mengesampingkan dunia 1, 4, dan 6. Saya cukup mengambil probabilitas yang terkait dengan dunia yang tersisa dan menskalakannya sehingga totalnya tetap 1. Sekarang probabilitas 2, 3, dan 5 masing-masing adalah ⅓, dan probabilitas bahwa hasil lemparan saya adalah angka genap sekarang hanya ⅓, karena 2 adalah satu-satunya hasil lemparan genap yang tersisa. Proses pembaruan probabilitas seiring dengan datangnya bukti baru ini adalah contoh dari pembaruan Bayesian.

Jadi, hal-hal tentang probabilitas ini tampak cukup sederhana! Bahkan komputer pun dapat menjumlahkan angka, jadi apa masalahnya? Masalahnya muncul ketika ada lebih dari beberapa kemungkinan. Misalnya, jika saya melempar dadu seratus kali, ada 6^{100} kemungkinan hasil. Tidak mungkin untuk memulai proses penalaran probabilistik dengan memberikan angka pada setiap kemungkinan hasil tersebut secara individual. Petunjuk untuk mengatasi kompleksitas ini berasal dari fakta bahwa lemparan dadu bersifat independen jika dadu tersebut diketahui adil yaitu, hasil dari satu lemparan saja tidak memengaruhi probabilitas untuk hasil dari lemparan lainnya. Dengan demikian, independensi sangat membantu dalam menyusun probabilitas untuk serangkaian peristiwa yang kompleks.

Misalkan saya bermain Monopoli dengan putra saya, George. Bidak saya berada di Just Visiting, dan George memiliki set kuning yang propertinya berjarak enam belas, tujuh belas, dan sembilan belas kotak dari Just Visiting. Haruskah dia membeli rumah untuk set kuning sekarang, sehingga saya harus membayar sewa yang sangat mahal jika saya mendarat di kotak-kotak itu, atau haruskah dia menunggu sampai giliran berikutnya? Itu tergantung pada probabilitas mendarat di set kuning pada giliran saya saat ini.

Berikut adalah aturan untuk melempar dadu di Monopoly: dua dadu dilempar dan bidak dipindahkan sesuai dengan total yang ditunjukkan; jika angka kembar muncul, pemain melempar lagi dan bergerak lagi; jika lemparan kedua adalah angka kembar, pemain melempar untuk ketiga kalinya dan bergerak lagi (tetapi jika lemparan ketiga adalah angka kembar, pemain masuk penjara sebagai gantinya). Jadi, misalnya, saya mungkin melempar 4-4 diikuti oleh 5-4, total 17; atau 2-2, lalu 2-2, lalu 6-2, total 16.

Seperti sebelumnya, saya hanya menjumlahkan probabilitas semua dunia tempat saya mendarat di set kuning. Sayangnya, ada banyak dunia. Sebanyak enam dadu dapat dilempar sekaligus, sehingga jumlah dunia mencapai ribuan. Selain itu, lemparan dadu tidak lagi independen, karena lemparan kedua tidak akan ada kecuali lemparan pertama menghasilkan angka kembar. Di sisi lain, jika kita menetapkan nilai pasangan dadu pertama, maka nilai pasangan dadu kedua bersifat independen. Adakah cara untuk menangkap ketergantungan semacam ini?

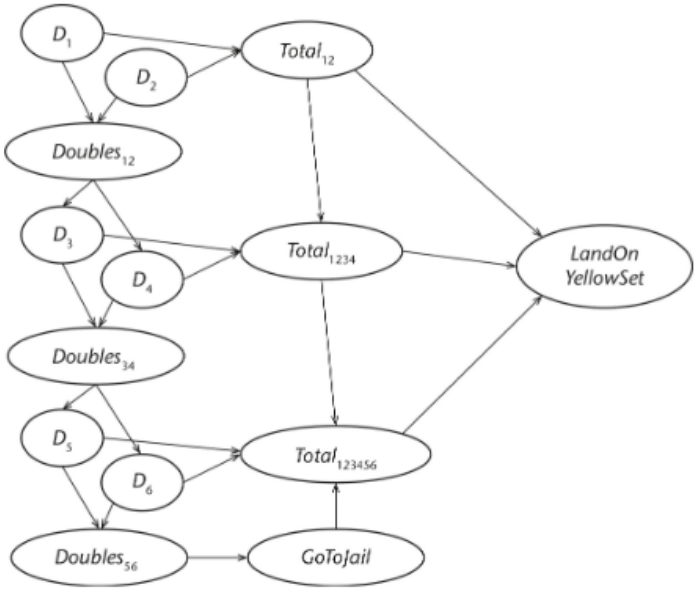
13.2 JARINGAN BAYESIAN

Pada awal tahun 1980-an, Judea Pearl mengusulkan bahasa formal yang disebut jaringan Bayesian (sering disingkat menjadi jaring Bayes) yang memungkinkan, dalam banyak situasi dunia nyata, untuk merepresentasikan probabilitas sejumlah besar hasil dalam bentuk yang sangat ringkas.

Gambar 13.1 menunjukkan jaringan Bayesian yang menggambarkan pelemparan dadu dalam Monopoli. Satu-satunya probabilitas yang harus diberikan adalah probabilitas ⅙ dari nilai 1, 2, 3, 4, 5, 6 untuk setiap lemparan dadu (D_1, D_2, dll.) yaitu, tiga puluh enam angka, bukan ribuan. Menjelaskan arti sebenarnya dari jaringan tersebut membutuhkan sedikit matematika, tetapi ide dasarnya adalah bahwa panah menunjukkan hubungan ketergantungan misalnya, nilai $Double_{12}$ bergantung pada nilai D_1 dan D_2. Demikian pula, nilai D_3 dan D_4 (lemparan dadu berikutnya) bergantung pada $Double_{12}$ karena jika $Double_{12}$ bernilai $false$, maka D_3 dan D_4 bernilai 0 (yaitu, tidak ada lemparan berikutnya).

Sama seperti logika proposisional, ada algoritma yang dapat menjawab pertanyaan apa pun untuk jaringan Bayesian apa pun dengan bukti apa pun. Misalnya, kita dapat menanyakan probabilitas *LandsOnYellowSet,* yang ternyata sekitar 3.88 persen. (Ini berarti George dapat menunggu sebelum membeli rumah untuk set kuning.) Sedikit lebih ambisius, kita dapat menanyakan probabilitas *LandsOnYellowSet* dengan asumsi lemparan kedua adalah double-3. Algoritma tersebut menghitung sendiri bahwa, dalam hal itu, lemparan pertama pasti double dan menyimpulkan bahwa jawabannya sekitar 36,1 persen. Ini adalah contoh pembaruan Bayesian: ketika bukti baru (bahwa lemparan kedua adalah angka 3 ganda) ditambahkan,

probabilitas *LandsOnYellowSet* berubah dari 3,88 persen menjadi 36,1 persen. Demikian pula, probabilitas bahwa saya melempar tiga kali ($Double_{34}$ benar) adalah 2,78 persen, sedangkan probabilitas bahwa saya melempar tiga kali dengan syarat saya mendarat di set kuning adalah 20,44 persen.



Gambar 13.1 Jaringan Bayesian yang merepresentasikan aturan untuk melempar dadu dalam Monopoli dan memungkinkan algoritma untuk menghitung probabilitas mendarat di sekumpulan kotak tertentu (seperti set kuning) dimulai dari kotak lain (seperti Hanya Berkunjung). (Untuk kesederhanaan, jaringan ini mengabaikan kemungkinan mendarat di kotak Kesempatan atau Kotak Harta Karun dan dialihkan ke lokasi yang berbeda.) D_1 dan D_2 mewakili lemparan awal dua dadu dan keduanya independen (tidak ada hubungan di antara keduanya). Jika angka kembar muncul ($Double_{12}$), maka pemain melempar lagi, sehingga D_3 dan D_4 memiliki nilai bukan nol, dan seterusnya. Dalam situasi yang dijelaskan, pemain mendarat di set kuning jika salah satu dari tiga totalnya adalah 16, 17, atau 19.

Jaringan Bayesian menyediakan cara untuk membangun sistem berbasis pengetahuan yang menghindari kegagalan yang melanda sistem pakar berbasis aturan pada tahun 1980-an. (Memang, seandainya komunitas AI kurang resisten terhadap probabilitas pada awal tahun 1980-an, mungkin mereka dapat menghindari musim dingin AI yang mengikuti gelembung sistem pakar berbasis aturan.) Ribuan aplikasi telah diterapkan, di berbagai bidang mulai dari diagnosis medis hingga pencegahan terorisme.

Jaringan Bayesian menyediakan mesin untuk merepresentasikan probabilitas yang diperlukan dan melakukan perhitungan untuk mengimplementasikan pembaruan Bayesian untuk banyak tugas kompleks. Namun, seperti logika proposisional, jaringan ini cukup terbatas dalam kemampuannya untuk merepresentasikan pengetahuan umum. Dalam banyak aplikasi, representasi jaringan Bayesian menjadi sangat besar dan berulang misalnya, sama seperti aturan Go harus diulang untuk setiap kotak dalam logika proposisional, aturan berbasis

probabilitas Monopoli harus diulang untuk setiap pemain, untuk setiap lokasi tempat pemain berada, dan untuk setiap gerakan dalam permainan.

Jaringan sebesar itu hampir tidak mungkin dibuat secara manual; sebagai gantinya, seseorang harus menggunakan kode yang ditulis dalam bahasa tradisional seperti C++ untuk menghasilkan dan menyusun beberapa fragmen jaringan Bayes. Meskipun ini praktis sebagai solusi teknik untuk masalah spesifik, ini merupakan hambatan untuk generalisasi karena kode C++ harus ditulis ulang oleh seorang ahli manusia untuk setiap aplikasi.

13.3 BAHASA PROBABILISTIK ORDE PERTAMA

Untungnya, ternyata kita dapat menggabungkan ekspresivitas logika orde pertama dengan kemampuan jaringan Bayesian untuk menangkap informasi probabilistik secara ringkas. Kombinasi ini memberi kita yang terbaik dari kedua dunia: sistem berbasis pengetahuan probabilistik mampu menangani berbagai situasi dunia nyata yang jauh lebih luas daripada metode logis atau jaringan Bayesian. Misalnya, kita dapat dengan mudah menangkap pengetahuan probabilistik tentang pewarisan genetik:

untuk semua orang $c, f,$ dan $m,$
jika f adalah ayah dari c dan m adalah ibu dari c
dan f dan m sama-sama memiliki golongan darah AB,
maka c memiliki golongan darah AB dengan probabilitas 0,5.

Kombinasi logika orde pertama dan probabilitas sebenarnya memberi kita lebih dari sekadar cara untuk mengekspresikan informasi yang tidak pasti tentang banyak objek. Alasannya adalah ketika kita menambahkan ketidakpastian ke dunia yang berisi objek, kita mendapatkan dua jenis ketidakpastian baru: bukan hanya ketidakpastian tentang fakta mana yang benar atau salah, tetapi juga ketidakpastian tentang objek apa yang ada dan ketidakpastian tentang objek mana yang mana. Jenis ketidakpastian ini benar-benar menyeluruh. Dunia tidak datang dengan daftar karakter, seperti drama Victoria; sebaliknya, Anda secara bertahap mempelajari keberadaan objek dari pengamatan.

Terkadang pengetahuan tentang objek baru bisa cukup pasti, seperti ketika Anda membuka jendela hotel dan melihat basilika Sacré-Cœur untuk pertama kalinya; atau bisa juga cukup tidak pasti, seperti ketika Anda merasakan gemuruh lembut yang mungkin merupakan gempa bumi atau kereta bawah tanah yang lewat. Dan sementara identitas Sacré-Cœur cukup jelas, identitas kereta bawah tanah tidak: Anda mungkin menaiki kereta fisik yang sama ratusan kali tanpa pernah menyadari bahwa itu adalah kereta yang sama.

Terkadang kita tidak perlu menyelesaikan ketidakpastian: Saya biasanya tidak menyebutkan semua tomat dalam sekantong tomat ceri dan melacak seberapa baik kondisi masing-masing, kecuali mungkin saya sedang mencatat kemajuan eksperimen pembusukan tomat. Di sisi lain, untuk kelas yang penuh dengan mahasiswa pascasarjana, saya berusaha sebaik mungkin untuk melacak identitas mereka. (Suatu kali, ada dua asisten peneliti di kelompok saya yang memiliki nama depan dan belakang yang sama dan memiliki penampilan

yang sangat mirip serta mengerjakan topik yang sangat terkait; setidaknya, saya cukup yakin ada dua.) Masalahnya adalah kita secara langsung mempersepsikan bukan identitas objek tetapi (aspek) penampilannya; objek biasanya tidak memiliki plat nomor kecil yang secara unik mengidentifikasinya. Identitas adalah sesuatu yang terkadang pikiran kita kaitkan dengan objek untuk tujuan kita sendiri.

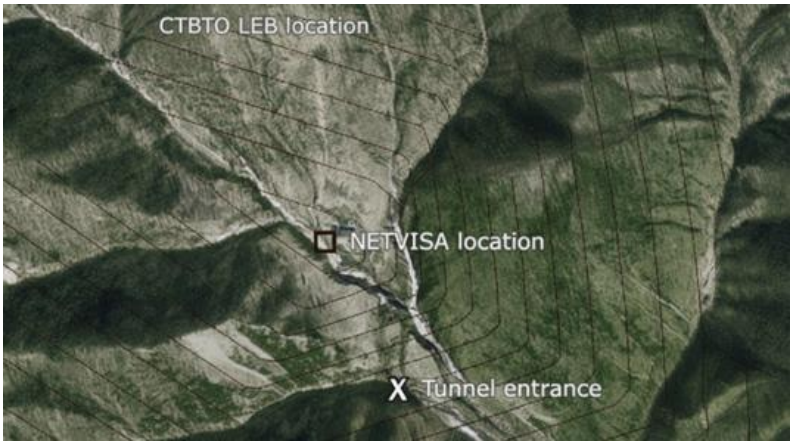
Kombinasi teori probabilitas dengan bahasa formal yang ekspresif adalah subbidang AI yang relatif baru, yang sering disebut pemrograman probabilistik. Beberapa lusin bahasa pemrograman probabilistik, atau PPL, telah dikembangkan, banyak di antaranya memperoleh kekuatan ekspresifnya dari bahasa pemrograman biasa daripada logika orde pertama. Semua sistem PPL memiliki kapasitas untuk merepresentasikan dan bernalar dengan pengetahuan yang kompleks dan tidak pasti.

Aplikasinya meliputi sistem TrueSkill Microsoft, yang memberi peringkat jutaan pemain video game setiap hari; model untuk aspek kognisi manusia yang sebelumnya tidak dapat dijelaskan oleh hipotesis mekanistik apa pun, seperti kemampuan untuk mempelajari kategori visual objek baru dari contoh tunggal; dan pemantauan seismik global untuk Perjanjian Pelarangan Uji Coba Nuklir Komprehensif (CTBT), yang bertanggung jawab untuk mendeteksi ledakan nuklir rahasia.

Sistem pemantauan CTBT mengumpulkan data pergerakan tanah secara real-time dari jaringan global yang terdiri dari lebih dari 150 seismometer dan bertujuan untuk mengidentifikasi semua peristiwa seismik yang terjadi di Bumi di atas magnitudo tertentu dan untuk menandai peristiwa yang mencurigakan. Jelas ada banyak ketidakpastian eksistensi dalam masalah ini, karena kita tidak tahu sebelumnya peristiwa apa yang akan terjadi; terlebih lagi, sebagian besar sinyal dalam data hanyalah noise.

Terdapat juga banyak ketidakpastian identitas: lonjakan energi seismik yang terdeteksi di stasiun A di Antartika mungkin berasal dari peristiwa yang sama atau mungkin tidak berasal dari lonjakan lain yang terdeteksi di stasiun B di Brasil. Mendengarkan Bumi seperti mendengarkan ribuan percakapan simultan yang telah diacak oleh penundaan transmisi dan gema serta ditenggelamkan oleh gelombang yang menghantam.

Bagaimana kita menyelesaikan masalah ini menggunakan pemrograman probabilistik? Orang mungkin berpikir kita membutuhkan beberapa algoritma yang sangat cerdas untuk memilah semua kemungkinan. Faktanya, dengan mengikuti metodologi sistem berbasis pengetahuan, kita tidak perlu merancang algoritma baru sama sekali. Kita cukup menggunakan PPL untuk mengekspresikan apa yang kita ketahui tentang geofisika: seberapa sering peristiwa cenderung terjadi di daerah dengan seismisitas alami, seberapa cepat gelombang seismik merambat melalui Bumi dan seberapa cepat gelombang tersebut meluruh, seberapa sensitif detektornya, dan seberapa banyak noise yang ada. Kemudian kita menambahkan data dan menjalankan algoritma penalaran probabilistik. Sistem pemantauan yang dihasilkan, yang disebut NET-VISA, telah beroperasi sebagai bagian dari rezim verifikasi perjanjian sejak tahun 2018. Gambar 13.2 menunjukkan deteksi NET-VISA terhadap uji coba nuklir tahun 2013 di Korea Utara.



Gambar 13.2 Perkiraan lokasi untuk uji coba nuklir tanggal 12 Februari 2013 yang dilakukan oleh pemerintah Korea Utara. Pintu masuk terowongan (tanda silang hitam di tengah bawah) diidentifikasi dalam foto satelit. Perkiraan lokasi ET-VISA sekitar 700 meter dari pintu masuk terowongan dan terutama didasarkan pada deteksi di stasiun yang berjarak 4.000 hingga 10.000 kilometer. Lokasi CTBTO LEB adalah perkiraan konsensus dari para ahli geofisika.

13.4 MELACAK DUNIA

Salah satu peran terpenting penalaran probabilistik adalah dalam melacak bagian-bagian dunia yang tidak dapat diamati secara langsung. Dalam sebagian besar permainan video dan papan, ini tidak perlu karena semua informasi yang relevan dapat diamati, tetapi di dunia nyata hal ini jarang terjadi.

Gambar 13.3 (kiri) Diagram situasi yang mengarah ke kecelakaan. Volvo yang mengemudi sendiri, ditandai V, mendekati persimpangan, mengemudi di jalur paling kanan dengan kecepatan tiga puluh delapan mil per jam. Lalu lintas di dua lajur lainnya berhenti dan lampu lalu lintas (L) berubah menjadi kuning. Tak terlihat oleh Volvo, sebuah Honda (H) sedang berbelok ke kiri; (kanan) akibat dari kecelakaan.

Contohnya adalah salah satu kecelakaan serius pertama yang melibatkan mobil otonom. Kecelakaan itu terjadi di South McClintock Drive di East Don Carlos Avenue di Tempe, Arizona, pada 24 Maret 2017. Seperti yang ditunjukkan pada gambar 13.3, sebuah Volvo otonom (V), yang melaju ke selatan di McClintock, mendekati persimpangan di mana lampu lalu lintas baru saja berubah menjadi kuning. Jalur Volvo kosong, sehingga mobil tersebut melaju dengan kecepatan yang sama melalui persimpangan. Kemudian sebuah kendaraan yang saat ini tidak

terlihat Honda (H) pada gambar 13.3 muncul dari belakang antrean lalu lintas yang berhenti dan terjadilah tabrakan.

Untuk menyimpulkan kemungkinan keberadaan Honda yang tidak terlihat, Volvo dapat mengumpulkan petunjuk saat mendekati persimpangan. Secara khusus, lalu lintas di dua jalur lainnya berhenti meskipun lampu hijau; mobil-mobil di depan antrean tidak bergerak maju ke persimpangan dan lampu remnya menyala. Ini bukan bukti konklusif adanya kendaraan yang berbelok ke kiri yang tidak terlihat, tetapi tidak perlu demikian; Bahkan probabilitas kecil pun cukup untuk menyarankan untuk memperlambat dan memasuki persimpangan dengan lebih hati-hati.

Pesan moral dari cerita ini adalah bahwa agen cerdas yang beroperasi di lingkungan yang sebagian dapat diamati harus melacak apa yang tidak dapat mereka lihat sejauh mungkin berdasarkan petunjuk dari apa yang dapat mereka lihat. Berikut contoh lain yang lebih dekat dengan kita: Di mana kunci Anda? Kecuali Anda sedang mengemudi sambil membaca buku ini tidak disarankan. Anda mungkin tidak dapat melihatnya saat ini. Di sisi lain, Anda mungkin tahu di mana letaknya: di saku Anda, di tas Anda, di meja samping tempat tidur, di saku mantel Anda yang tergantung, atau mungkin di gantungan di dapur. Anda tahu ini karena Anda meletakkannya di sana dan kunci itu belum berpindah sejak saat itu. Ini adalah contoh sederhana penggunaan pengetahuan dan penalaran untuk melacak keadaan dunia.

Tanpa kemampuan ini, kita akan tersesat seringkali secara harfiah. Misalnya, saat saya menulis ini, saya sedang melihat dinding putih kamar hotel yang biasa saja. Di mana saya? Jika saya harus bergantung pada masukan persepsi saya saat ini, saya memang akan tersesat. Faktanya, saya tahu bahwa saya berada di Zürich, karena saya tiba di Zürich kemarin dan saya belum pergi. Seperti manusia, robot perlu mengetahui di mana mereka berada agar mereka dapat menavigasi dengan sukses melalui ruangan, bangunan, jalan, hutan, dan gurun.

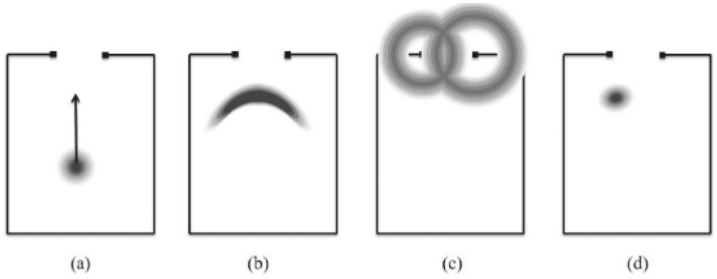
Dalam AI, kita menggunakan istilah keadaan kepercayaan untuk merujuk pada pengetahuan agen saat ini tentang keadaan dunia betapapun tidak lengkap dan tidak pastinya pengetahuan tersebut. Secara umum, keadaan kepercayaan bukan masukan persepsi saat ini adalah dasar yang tepat untuk membuat keputusan tentang apa yang harus dilakukan. Menjaga keadaan kepercayaan tetap mutakhir adalah aktivitas inti bagi setiap agen cerdas. Untuk beberapa bagian dari keadaan kepercayaan, ini terjadi secara otomatis misalnya, saya sepertinya tahu bahwa saya berada di Zürich, tanpa harus memikirkannya. Untuk bagian lain, ini terjadi sesuai permintaan, bisa dibilang begitu. Misalnya, ketika saya bangun di kota baru dengan jet lag yang parah, di tengah perjalanan panjang, saya mungkin harus berusaha secara sadar untuk mengingat kembali di mana saya berada, apa yang seharusnya saya lakukan, dan mengapa sedikit seperti laptop yang melakukan reboot sendiri, saya kira.

Melacak keadaan tidak berarti selalu mengetahui secara pasti keadaan segala sesuatu di dunia. Jelas ini tidak mungkin misalnya, saya tidak tahu siapa yang menempati kamar lain di hotel saya yang biasa-biasa saja di Zürich, apalagi lokasi dan aktivitas sebagian besar dari delapan miliar orang di Bumi. Saya sama sekali tidak tahu apa yang terjadi di belahan alam semesta lain di luar tata surya. Ketidakpastian saya tentang keadaan saat ini sangat besar dan tak terhindarkan.

Metode dasar untuk melacak dunia yang tidak pasti adalah pembaruan Bayesian. Algoritma untuk melakukan ini biasanya memiliki dua langkah: langkah prediksi, di mana agen memprediksi keadaan dunia saat ini berdasarkan tindakan terbarunya, dan kemudian langkah pembaruan, di mana ia menerima masukan persepsi baru dan memperbarui keyakinannya sesuai dengan itu. Untuk mengilustrasikan cara kerjanya, pertimbangkan masalah yang dihadapi robot dalam menentukan di mana ia berada. Gambar 13.4(a) mengilustrasikan kasus tipikal: Robot berada di tengah ruangan, dengan beberapa ketidakpastian tentang lokasi tepatnya, dan ingin melewati pintu.

Ia memerintahkan rodanya untuk bergerak 1,5 meter menuju pintu; sayangnya, rodanya sudah tua dan goyah, sehingga prediksi robot tentang di mana ia berakhir cukup tidak pasti, seperti yang ditunjukkan pada gambar 13.4(b). Jika ia mencoba untuk terus bergerak sekarang, ia mungkin akan menabrak. Untungnya, ia memiliki perangkat sonar untuk mengukur jarak ke kusen pintu. Seperti yang ditunjukkan gambar 13.4(c), pengukuran menunjukkan robot berada sekitar 70 sentimeter dari kusen pintu kiri dan 85 sentimeter dari kanan. Terakhir, robot memperbarui keadaan keyakinannya dengan menggabungkan prediksi pada (b) dengan pengukuran pada (c) untuk mendapatkan keadaan keyakinan baru pada gambar 13.4(d).

Algoritma untuk melacak keadaan keyakinan dapat diterapkan untuk menangani tidak hanya ketidakpastian tentang lokasi tetapi juga ketidakpastian tentang peta itu sendiri. Hal ini menghasilkan teknik yang disebut SLAM (*simultaneous localization and mapping*). SLAM merupakan komponen inti dari banyak aplikasi AI, mulai dari sistem augmented reality hingga mobil otonom dan robot penjelajah planet.



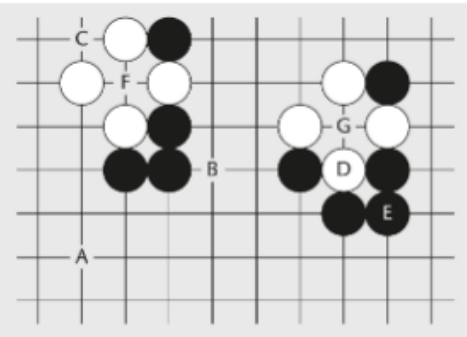
Gambar 13.4 Sebuah robot mencoba bergerak melewati ambang pintu. (a) Keadaan keyakinan awal: robot agak tidak yakin dengan lokasinya; ia mencoba bergerak 1,5 meter menuju pintu. (b) Langkah prediksi: robot memperkirakan bahwa ia lebih dekat ke pintu tetapi cukup tidak yakin tentang arah sebenarnya karena motornya sudah tua dan rodanya goyah. (c) Robot mengukur jarak ke setiap kusen pintu menggunakan perangkat sonar berkualitas rendah; perkiraannya adalah 70 sentimeter dari kusen pintu kiri dan 85 sentimeter dari kanan. (d) Langkah pembaruan: menggabungkan prediksi di (b) dengan pengamatan di (c) memberikan keadaan keyakinan baru. Sekarang robot memiliki gambaran yang cukup baik tentang di mana ia berada dan perlu sedikit mengoreksi arahnya untuk melewati pintu.

BAB 14
DEFINISI PEMBELAJARAN BERBASIS PENGALAMAN

Belajar berarti meningkatkan kinerja berdasarkan pengalaman. Untuk sistem persepsi visual, itu mungkin berarti belajar mengenali lebih banyak kategori objek berdasarkan melihat contoh-contoh dari kategori tersebut; untuk sistem berbasis pengetahuan, sekadar memperoleh lebih banyak pengetahuan adalah bentuk pembelajaran, karena itu berarti sistem dapat menjawab lebih banyak pertanyaan; untuk sistem pengambilan keputusan yang melihat ke depan seperti AlphaGo, belajar dapat berarti meningkatkan kemampuannya untuk mengevaluasi posisi atau meningkatkan kemampuannya untuk menjelajahi bagian-bagian yang berguna dari pohon kemungkinan.

14.1 PEMBELAJARAN TERAWASI: HIPOTESIS DARI CONTOH

Bentuk pembelajaran mesin yang paling umum disebut pembelajaran terawasi. Algoritma pembelajaran terawasi diberi kumpulan contoh pelatihan, masing-masing diberi label dengan keluaran yang benar, dan harus menghasilkan hipotesis tentang apa aturan yang benar. Biasanya, sistem pembelajaran terawasi berupaya mengoptimalkan kesesuaian antara hipotesis dan contoh pelatihan. Seringkali ada juga penalti untuk hipotesis yang lebih rumit daripada yang diperlukan seperti yang direkomendasikan oleh prinsip Ockham.



Gambar 14.1 Langkah legal dan ilegal dalam Go: langkah A, B, dan C legal untuk Hitam, sedangkan langkah D, E, dan F ilegal. Langkah G mungkin legal atau mungkin tidak, tergantung pada apa yang telah terjadi sebelumnya dalam permainan.

Mari kita ilustrasikan ini untuk masalah mempelajari langkah-langkah legal dalam Go. (Jika Anda sudah mengetahui aturan Go, maka setidaknya ini akan mudah diikuti; jika tidak, maka Anda akan lebih mudah memahami program pembelajaran.) Misalkan algoritma dimulai dengan hipotesis

untuk semua langkah waktu t, dan untuk semua lokasi l,
adalah legal untuk memainkan bidak di lokasi l pada waktu t.

Giliran Hitam untuk bergerak di posisi yang ditunjukkan pada gambar 14.1. Algoritma mencoba A: itu baik-baik saja. B dan C juga. Kemudian ia mencoba D, di atas bidak putih yang sudah ada: itu ilegal. (Dalam catur atau backgammon, itu akan baik-baik saja begitulah cara bidak ditangkap.) Langkah di E, di atas bidak hitam, juga ilegal. (Tidak sah dalam catur juga, tetapi sah dalam backgammon.) Sekarang, dari lima contoh pelatihan ini, algoritma mungkin mengusulkan hipotesis berikut:

untuk semua langkah waktu t, dan untuk semua lokasi l,
jika l tidak ditempati pada waktu t,
maka sah untuk memainkan batu di lokasi l pada waktu t.

Kemudian ia mencoba F dan terkejut menemukan bahwa F tidak sah. Setelah beberapa kali percobaan gagal, ia menetapkan hipotesis berikut:

untuk semua langkah waktu t, dan untuk semua lokasi l,
jika l tidak ditempati pada waktu t dan
l tidak dikelilingi oleh batu lawan,
maka sah untuk memainkan batu di lokasi l pada waktu t.

(Ini kadang-kadang disebut aturan tidak bunuh diri.) Akhirnya, ia mencoba G, yang dalam kasus ini ternyata sah. Setelah berpikir sejenak dan mungkin mencoba beberapa eksperimen lagi, ia menetapkan hipotesis bahwa G tidak apa-apa, meskipun dikelilingi, karena ia menangkap batu putih di D dan karenanya langsung tidak dikelilingi. Seperti yang Anda lihat dari perkembangan aturan secara bertahap, pembelajaran terjadi melalui serangkaian modifikasi pada hipotesis agar sesuai dengan contoh yang diamati.

Ini adalah sesuatu yang dapat dilakukan dengan mudah oleh algoritma pembelajaran. Para peneliti pembelajaran mesin telah merancang berbagai macam algoritma cerdas untuk menemukan hipotesis yang baik dengan cepat. Di sini algoritma tersebut mencari di ruang ekspresi logis yang mewakili aturan Go, tetapi hipotesis tersebut juga dapat berupa ekspresi aljabar yang mewakili hukum fisika, jaringan Bayesian probabilistik yang mewakili penyakit dan gejala, atau bahkan program komputer yang mewakili perilaku rumit dari mesin lain.

Poin penting kedua adalah bahwa bahkan hipotesis yang baik pun bisa salah: sebenarnya, hipotesis yang diberikan di atas salah, bahkan setelah diperbaiki untuk memastikan bahwa G legal. Hipotesis tersebut perlu menyertakan aturan ko atau tidak ada pengulangan misalnya, jika Putih baru saja menangkap batu hitam di G dengan bermain di D, Hitam tidak boleh menangkap kembali dengan bermain di G, karena itu menghasilkan posisi yang sama lagi.

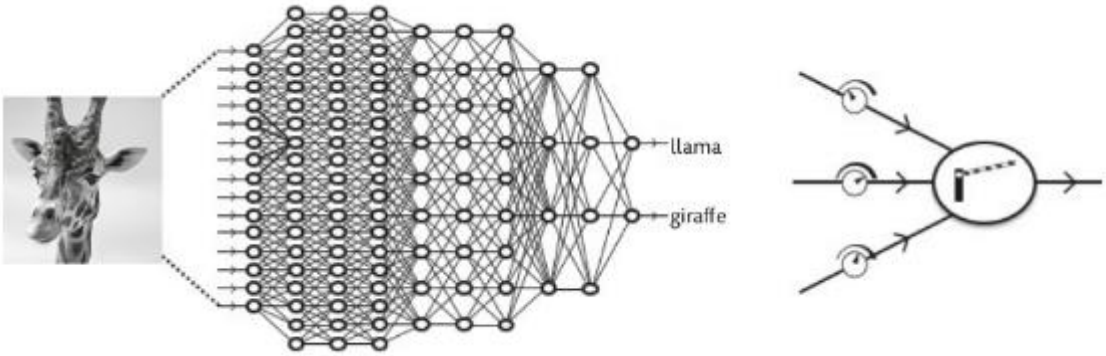
Perhatikan bahwa aturan ini merupakan penyimpangan radikal dari apa yang telah dipelajari program sejauh ini, karena itu berarti bahwa legalitas tidak dapat ditentukan dari posisi saat ini; Sebaliknya, seseorang juga harus mengingat posisi sebelumnya.

Filsuf Skotlandia David Hume menunjukkan pada tahun 1748 bahwa penalaran induktif yaitu, penalaran dari pengamatan khusus ke prinsip umum tidak pernah dapat dijamin. Dalam teori pembelajaran statistik modern, kita tidak meminta jaminan kebenaran sempurna tetapi hanya jaminan bahwa hipotesis yang ditemukan mungkin benar secara perkiraan. Algoritma pembelajaran dapat "tidak beruntung" dan melihat sampel yang tidak representatif misalnya, mungkin tidak pernah mencoba langkah seperti G, karena menganggapnya ilegal. Algoritma juga dapat gagal memprediksi beberapa kasus ekstrem yang aneh, seperti yang tercakup oleh beberapa bentuk aturan tanpa pengulangan yang lebih rumit dan jarang digunakan. Tetapi, selama alam semesta menunjukkan beberapa tingkat keteraturan, sangat tidak mungkin algoritma tersebut dapat menghasilkan hipotesis yang sangat buruk, karena hipotesis seperti itu kemungkinan besar telah "ditemukan" oleh salah satu eksperimen.

Pembelajaran mendalam teknologi yang menyebabkan semua kehebohan tentang AI di media pada dasarnya adalah bentuk pembelajaran terawasi. Ini merupakan salah satu kemajuan paling signifikan dalam AI dalam beberapa dekade terakhir, jadi penting untuk memahami cara kerjanya. Terlebih lagi, beberapa peneliti percaya bahwa ini akan mengarah pada sistem AI setingkat manusia dalam beberapa tahun ke depan, jadi ada baiknya untuk menilai apakah hal itu mungkin benar.

Paling mudah memahami pembelajaran mendalam dalam konteks tugas tertentu, seperti belajar membedakan jerapah dan llama. Dengan beberapa foto berlabel dari masing-masing hewan, algoritma pembelajaran harus membentuk hipotesis yang memungkinkannya untuk mengklasifikasikan gambar yang tidak berlabel. Dari sudut pandang komputer, sebuah gambar hanyalah tabel angka yang besar, dengan setiap angka sesuai dengan salah satu dari tiga nilai RGB untuk satu piksel gambar. Jadi, alih-alih hipotesis Go yang mengambil posisi papan dan langkah sebagai masukan dan memutuskan apakah langkah tersebut legal, kita membutuhkan hipotesis jerapah-llama yang mengambil tabel angka sebagai masukan dan memprediksi kategori (jerapah atau llama).

Sekarang pertanyaannya adalah, hipotesis seperti apa? Selama lebih dari lima puluh tahun terakhir penelitian visi komputer, banyak pendekatan telah dicoba. Pendekatan favorit saat ini adalah jaringan konvolusional dalam (*deep convolutional network*). Izinkan saya menjelaskannya: Disebut jaringan karena mewakili ekspresi matematika kompleks yang tersusun secara teratur dari banyak subekspresi yang lebih kecil, dan struktur komposisinya berbentuk jaringan. (Jaringan seperti itu sering disebut jaringan saraf karena perancangnya mengambil inspirasi dari jaringan neuron di otak.) Disebut konvolusional karena itu adalah cara matematis yang canggih untuk mengatakan bahwa struktur jaringan berulang dalam pola tetap di seluruh gambar masukan. Dan disebut dalam (deep) karena jaringan seperti itu biasanya memiliki banyak lapisan, dan juga karena kedengarannya mengesankan dan sedikit menakutkan.



Gambar 14.2 (kiri) Gambaran sederhana dari jaringan konvolusional dalam untuk mengenali objek dalam gambar. Nilai piksel gambar dimasukkan di sebelah kiri dan jaringan mengeluarkan nilai pada dua node paling kanan, yang menunjukkan seberapa besar kemungkinan gambar tersebut adalah llama atau jerapah. Perhatikan bagaimana pola koneksi lokal, yang ditunjukkan oleh garis gelap di lapisan pertama, berulang di seluruh lapisan; (kanan) salah satu node dalam jaringan. Terdapat bobot yang dapat disesuaikan pada setiap nilai yang masuk sehingga node lebih atau kurang memperhatikan nilai tersebut. Kemudian sinyal total yang masuk melewati fungsi gerbang yang memungkinkan sinyal besar untuk masuk tetapi menekan sinyal kecil.

Contoh yang disederhanakan (disederhanakan karena jaringan nyata mungkin memiliki ratusan lapisan dan jutaan node) ditunjukkan pada gambar 14.2. Jaringan tersebut sebenarnya adalah gambaran dari ekspresi matematika yang kompleks dan dapat disesuaikan. Setiap node dalam jaringan sesuai dengan ekspresi sederhana yang dapat disesuaikan, seperti yang diilustrasikan dalam gambar. Penyesuaian dilakukan dengan mengubah bobot pada setiap input, seperti yang ditunjukkan oleh "kontrol volume." Jumlah tertimbang dari input kemudian dilewatkan melalui fungsi gerbang sebelum mencapai sisi output node; biasanya, fungsi gerbang menekan nilai-nilai kecil dan memungkinkan nilai-nilai yang lebih besar untuk melewatinya.

Pembelajaran terjadi di dalam jaringan hanya dengan menyesuaikan semua kenop kontrol volume untuk mengurangi kesalahan prediksi pada contoh yang diberi label. Sesederhana itu: tidak ada sihir, tidak ada algoritma yang sangat cerdas. Menentukan ke mana harus memutar kenop untuk mengurangi kesalahan adalah penerapan kalkulus yang mudah untuk menghitung bagaimana perubahan setiap bobot akan mengubah kesalahan pada lapisan output. Hal ini mengarah pada rumus sederhana untuk menyebarkan kesalahan mundur dari lapisan output ke lapisan input, dengan menyesuaikan kenop di sepanjang jalan.

Ajaibnya, proses ini berhasil. Untuk tugas mengenali objek dalam foto, algoritma pembelajaran mendalam telah menunjukkan kinerja yang luar biasa. Petunjuk pertama tentang hal ini muncul dalam kompetisi ImageNet 2012, yang menyediakan data pelatihan yang terdiri dari 1,2 juta gambar berlabel dalam seribu kategori, dan kemudian mengharuskan algoritma untuk memberi label seratus ribu gambar baru. Geoff Hinton, seorang psikolog komputasi Inggris yang berada di garis depan revolusi jaringan saraf pertama pada tahun 1980-

an, telah bereksperimen dengan jaringan konvolusional dalam yang sangat besar: 650.000 node dan 60 juta parameter.

Dia dan kelompoknya di Universitas Toronto mencapai tingkat kesalahan ImageNet sebesar 15 persen, peningkatan dramatis dari rekor terbaik sebelumnya sebesar 26 persen. Pada tahun 2015, puluhan tim menggunakan metode pembelajaran mendalam dan tingkat kesalahan turun menjadi 5 persen, sebanding dengan manusia yang telah menghabiskan waktu berminggu-minggu untuk belajar mengenali seribu kategori dalam pengujian. Pada tahun 2017, tingkat kesalahan mesin adalah 2 persen.

Selama periode yang hampir sama, telah terjadi peningkatan yang sebanding dalam pengenalan ucapan dan terjemahan mesin berdasarkan metode serupa. Secara keseluruhan, ini adalah tiga area aplikasi terpenting untuk AI. Pembelajaran mendalam juga memainkan peran penting dalam aplikasi pembelajaran penguatan misalnya, dalam mempelajari fungsi evaluasi yang digunakan AlphaGo untuk memperkirakan keinginan dari kemungkinan posisi di masa depan, dan dalam mempelajari pengendali untuk perilaku robot yang kompleks.

Sampai saat ini, kita masih memiliki sedikit pemahaman mengapa pembelajaran mendalam (*deep learning*) bekerja sebaik itu. Mungkin penjelasan terbaik adalah bahwa jaringan dalam (*deep network*) memang dalam: karena memiliki banyak lapisan, setiap lapisan dapat mempelajari transformasi yang cukup sederhana dari input ke outputnya, sementara banyak transformasi sederhana tersebut jika digabungkan akan menghasilkan transformasi kompleks yang diperlukan untuk mengubah foto menjadi label kategori. Selain itu, jaringan dalam untuk visi memiliki struktur bawaan yang menegakkan invariansi translasi dan invariansi skala artinya seekor anjing tetaplah seekor anjing di mana pun ia muncul dalam gambar dan seberapa besar pun ukurannya dalam gambar.

Sifat penting lain dari jaringan dalam adalah bahwa mereka seringkali tampaknya menemukan representasi internal yang menangkap fitur-fitur dasar gambar, seperti mata, garis-garis, dan bentuk-bentuk sederhana. Tidak satu pun dari fitur-fitur ini yang sudah ada sejak awal. Kita tahu fitur-fitur tersebut ada karena kita dapat bereksperimen dengan jaringan yang telah dilatih dan melihat jenis data apa yang menyebabkan node internal (biasanya yang dekat dengan lapisan output) menyala.

Bahkan, dimungkinkan untuk menjalankan algoritma pembelajaran dengan cara yang berbeda sehingga algoritma tersebut menyesuaikan gambar itu sendiri untuk menghasilkan respons yang lebih kuat pada node internal yang dipilih. Mengulangi proses ini berkali-kali menghasilkan apa yang sekarang dikenal sebagai mimpi mendalam atau citra inceptionism, seperti yang ada pada gambar 14.3. Inceptionism telah menjadi bentuk seni tersendiri, menghasilkan citra yang tidak seperti seni manusia mana pun.

Terlepas dari semua pencapaian luar biasa mereka, sistem pembelajaran mendalam seperti yang kita pahami saat ini masih jauh dari menyediakan dasar untuk sistem yang cerdas secara umum. Kelemahan utama mereka adalah bahwa mereka adalah sirkuit; mereka adalah sepupu dari logika proposisional dan jaringan Bayesian, yang, terlepas dari semua sifatnya yang luar biasa, juga kurang mampu mengekspresikan bentuk pengetahuan yang kompleks secara ringkas.

Ini berarti bahwa jaringan mendalam yang beroperasi dalam "mode asli" membutuhkan sejumlah besar sirkuit untuk merepresentasikan jenis pengetahuan umum yang cukup sederhana. Itu, pada gilirannya, menyiratkan sejumlah besar bobot untuk dipelajari dan karenanya kebutuhan akan jumlah contoh yang tidak masuk akal lebih banyak daripada yang dapat disediakan oleh alam semesta.

Gambar 14.3 Gambar yang dihasilkan oleh perangkat lunak DeepDream Google.

Beberapa orang berpendapat bahwa otak juga terbuat dari sirkuit, dengan neuron sebagai elemen sirkuit; oleh karena itu, sirkuit dapat mendukung kecerdasan tingkat manusia. Ini benar, tetapi hanya dalam arti yang sama bahwa otak terbuat dari atom: atom memang dapat mendukung kecerdasan tingkat manusia, tetapi itu tidak berarti bahwa hanya mengumpulkan banyak atom akan menghasilkan kecerdasan.

Atom-atom tersebut harus diatur dengan cara tertentu. Dengan alasan yang sama, sirkuit harus diatur dengan cara tertentu. Komputer juga terbuat dari sirkuit, baik dalam memorinya maupun dalam unit pemrosesannya; tetapi sirkuit tersebut harus diatur dengan cara tertentu, dan lapisan perangkat lunak harus ditambahkan, sebelum komputer dapat mendukung pengoperasian bahasa pemrograman tingkat tinggi dan sistem penalaran logis. Namun, saat ini, tidak ada tanda-tanda bahwa sistem pembelajaran mendalam dapat mengembangkan kemampuan tersebut dengan sendirinya dan juga tidak masuk akal secara ilmiah untuk mengharuskan mereka melakukannya.

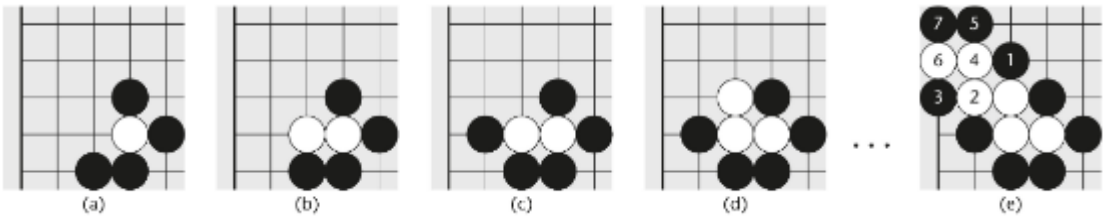
Ada alasan lebih lanjut untuk berpikir bahwa pembelajaran mendalam mungkin mencapai titik jenuh yang jauh di bawah kecerdasan umum, tetapi bukan tujuan saya di sini untuk mendiagnosis semua masalah: orang lain, baik di dalam maupun di luar komunitas pembelajaran mendalam, telah mencatat banyak di antaranya. Intinya adalah bahwa hanya menciptakan jaringan yang lebih besar dan lebih dalam serta kumpulan data yang lebih besar dan mesin yang lebih besar saja tidak cukup untuk menciptakan AI tingkat manusia. Kita telah melihat (dalam Lampiran B) pandangan CEO DeepMind Demis Hassabis bahwa "pemikiran tingkat tinggi dan penalaran simbolik" sangat penting untuk AI. Pakar pembelajaran mendalam terkemuka lainnya, François Chollet, mengatakannya seperti ini: “Banyak aplikasi lain yang sama sekali di luar jangkauan teknik pembelajaran mendalam saat ini bahkan dengan sejumlah

besar data yang dianotasi manusia. Kita perlu beralih dari pemetaan input-ke-output langsung, dan menuju penalaran dan abstraksi."

14.2 BELAJAR DARI BERPIKIR

Setiap kali Anda mendapati diri Anda harus berpikir tentang sesuatu, itu karena Anda belum mengetahui jawabannya. Ketika seseorang menanyakan nomor ponsel baru Anda, Anda mungkin tidak mengetahuinya. Anda berpikir, "Oke, saya tidak tahu; jadi bagaimana cara menemukannya?" Karena tidak terlalu bergantung pada ponsel, Anda tidak tahu cara menemukannya.

Anda berpikir, "Bagaimana cara saya mencari tahu cara menemukannya?" Anda memiliki jawaban umum untuk ini: "Mungkin mereka menempatkannya di suatu tempat yang mudah ditemukan pengguna." (Tentu saja, Anda bisa salah tentang ini.) Tempat yang jelas adalah di bagian atas layar beranda (tidak ada di sana), di dalam aplikasi Telepon, atau di Pengaturan untuk aplikasi tersebut. Anda mencoba Pengaturan > Telepon, dan di sanalah nomornya berada.



Gambar 14.4 Konsep tangga dalam permainan Go. (a) Hitam mengancam untuk menangkap bidak Putih. (b) Putih mencoba melarikan diri. (c) Hitam menghalangi arah pelarian tersebut. (d) Putih mencoba arah lain. (e) Permainan berlanjut sesuai urutan yang ditunjukkan oleh angka-angka. Tangga tersebut akhirnya mencapai tepi papan, di mana Putih tidak punya tempat untuk melarikan diri. Pukulan telak diberikan pada langkah ke-7: kelompok bidak Putih sepenuhnya dikelilingi dan mati.

Lain kali Anda ditanya nomor Anda, Anda tahu nomornya atau Anda tahu persis cara mendapatkannya. Anda mengingat prosedurnya, tidak hanya untuk ponsel ini pada kesempatan ini tetapi untuk semua ponsel serupa pada semua kesempatan yaitu, Anda menyimpan dan menggunakan kembali solusi umum untuk masalah tersebut.

Generalisasi ini dapat dibenarkan karena Anda memahami bahwa spesifikasi ponsel tertentu dan kesempatan tertentu ini tidak relevan. Anda akan terkejut jika metode ini hanya berfungsi pada hari Selasa untuk nomor telepon yang berakhiran 17. Go menawarkan contoh indah dari jenis pembelajaran yang sama. Pada gambar 14.4(a), kita melihat situasi umum di mana Hitam mengancam untuk menangkap bidak Putih dengan mengelilinginya. Putih mencoba melarikan diri dengan menambahkan bidak yang terhubung dengan bidak asli, tetapi Hitam terus memutus jalur pelarian. Pola gerakan ini membentuk tangga bidak secara diagonal di papan, hingga mencapai tepi; kemudian Putih tidak punya tempat untuk pergi.

Jika Anda bermain sebagai Putih, Anda mungkin tidak akan membuat kesalahan yang sama lagi: Anda menyadari bahwa pola tangga selalu menghasilkan penangkapan pada akhirnya, untuk lokasi awal apa pun dan arah apa pun, pada tahap permainan apa pun, baik Anda bermain sebagai Putih atau Hitam.

Satu-satunya pengecualian terjadi ketika tangga bertemu dengan beberapa batu tambahan milik yang lolos. Keumuman pola tangga mengikuti secara langsung dari aturan Go. Kasus nomor telepon yang hilang dan kasus tangga Go menggambarkan kemungkinan mempelajari aturan umum yang efektif dari satu contoh jauh berbeda dari jutaan contoh yang dibutuhkan untuk pembelajaran mendalam. Dalam AI, jenis pembelajaran ini disebut pembelajaran berbasis penjelasan: setelah melihat contoh tersebut, agen dapat menjelaskan kepada dirinya sendiri mengapa hasilnya seperti itu dan dapat mengekstrak prinsip umum dengan melihat faktor-faktor apa yang penting untuk penjelasan tersebut.

Secara tegas, proses ini sendiri tidak menambahkan pengetahuan baru misalnya, White bisa saja menyimpulkan keberadaan dan hasil dari pola tangga umum dari aturan Go, tanpa pernah melihat contohnya. Namun, kemungkinan besar White tidak akan pernah menemukan konsep tangga tanpa melihat contohnya; jadi, kita dapat memahami pembelajaran berbasis penjelasan sebagai metode yang ampuh untuk menyimpan hasil komputasi secara umum, sehingga menghindari keharusan untuk mengulangi proses penalaran yang sama (atau membuat kesalahan yang sama dengan proses penalaran yang tidak sempurna) di masa mendatang.

Penelitian dalam ilmu kognitif telah menekankan pentingnya jenis pembelajaran ini dalam kognisi manusia. Dengan nama chunking, ini membentuk pilar utama dari teori kognisi Allen Newell yang sangat berpengaruh. (Newell adalah salah satu peserta lokakarya Dartmouth tahun 1956 dan pemenang bersama Penghargaan Turing tahun 1975 dengan Herb Simon.) Ini menjelaskan bagaimana manusia menjadi lebih fasih dalam tugas-tugas kognitif dengan latihan, karena berbagai sub-tugas yang awalnya membutuhkan pemikiran menjadi otomatis. Tanpa itu, percakapan manusia akan terbatas pada respons satu atau dua kata, dan para matematikawan masih akan menghitung dengan jari mereka.

DAFTAR PUSTAKA

Abbass, H. A. (2019). Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, *11*(2), 159-171.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, *40*(S1), S293-S340.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019, July). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 2429-2437).

Barrat, J. (2023). *Our final invention: Artificial intelligence and the end of the human era*. Hachette UK.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, *45*(3), 1433-1450.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, *35*(2), 309-317.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., ... & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics*, *3*(1), 241-255.

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, *69*(4), 819-844.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., ... & Griffiths, T. L. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851-1863.

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & Von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, *51*(1), 91-105.

Edwards, H. (2024). Toward Ethical AI: Relational Dynamics, Theory of Mind, and Human-Compatible Artificial Intelligence. *Contexts*, *1*(5).

Faroldi, F. L. (2023). On the Human-Compatible Approach to the Alignment Problem: A Research Program. *Moral Normativity in an Interdisciplinary Perspective*, 123-134.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, *33*(3), 523-539.

Foster, J. G. (2023). From thin to thick: Toward a politics of human-compatible AI. *Public Culture*, *35*(3), 417-430.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment: I. Gabriel. *Minds and machines*, *30*(3), 411-437.

Goldfarb, A., & Lindsay, J. R. (2021). Prediction and judgment: Why artificial intelligence increases the importance of humans in war. *International Security*, *46*(3), 7-50.

Helbing, D. (2018). Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution* (pp. 47-72). Cham: Springer International Publishing.

Human, S., & Bernroider, E. (2024). AI-assisted framework for sustainable human-compatible innovation and transformation. *CENTERIS/Proj MAN/HCist 2024*, 104.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons*, *61*(4), 577-586.

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of clinical neuroscience*, *64*, 277-282.

Kido, T., & Takadama, K. (2025, May). Challenges in Human-Compatible AI for Well-Being: Harnessing Potential of GenAI for AI-Powered Science. In *Proceedings of the AAAI Symposium Series* (Vol. 5, No. 1, pp. 291-292).

Li, J., Cao, H., Lin, L., Hou, Y., Zhu, R., & El Ali, A. (2024, May). User experience design professionals' perceptions of generative artificial intelligence. In *Proceedings of the 2024 CHI conference on human factors in computing systems* (pp. 1-18).

McCarthy, J. (2022). Artificial intelligence, logic, and formalising common sense. *Machine Learning and the City: Applications in Architecture and Urban Design*, 69-90.

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Noël, J. C. (2020). Human compatible: AI and the problem of control. Stuart Russell. New York, Viking, 2019, 352 pages. *Politique étrangère*, (4), XV-XV.

Nyholm, S. (2023). A new control problem? Humanoid robots, artificial intelligence, and the value of control. *AI and Ethics*, *3*(4), 1229-1239.

Russell, S. (2022). Human-Compatible Artificial Intelligence. *Human-like machine intelligence*, *1*, 3-22.

Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., & Forestiero, A. (2022). A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions. *Applied sciences*, *12*(8), 4054.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & technology*, *34*(4), 1057-1084.

Schwarz, E. (2021). Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control. *Philosophical Journal of Conflict and Violence*.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. In *Mission Ai* (pp. 15-41). Springer, Cham.

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109-124.

Stelmaszak, M., Joshi, M., & Constantiou, I. (2026). Artificial intelligence as an organizing capability arising from human-algorithm relations. *Journal of Management Studies*, *63*(2), 335-365.

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, *19*(5), 722-734.

Takadama, K., & Shintani, D. (2025, May). Towards Human-Compatible AI for Well-being by Integrating Physiological Viewpoint With Machine Learning Viewpoint. In *Proceedings of the AAAI Symposium Series* (Vol. 5, No. 1, pp. 272-277).

Topf, D. (2021). Human Compatible: Artificial Intelligence and the Problem of Control. *Journal of Interdisciplinary Studies*, *33*(1-2), 192-195.

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and information technology*, *20*(1), 27-40.

Wang, Y. (2021). Artificial intelligence in educational leadership: a symbiotic role of human-artificial intelligence decision-making. *Journal of Educational Administration*, *59*(3), 256-270.

Wen, J., He, L., & Zhu, F. (2018). Swarm robotics control and communications: Imminent challenges for next generation smart logistics. *IEEE Communications Magazine*, *56*(7), 102-107.

Yampolskiy, R. V. (2022). On the controllability of artificial intelligence: An analysis of limitations. *Journal of Cyber Security and Mobility*, *11*(3), 321-403.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4).

MENGONTROL KECERDASAN BUATAN SESUAI DENGAN KEBUTUHAN MANUSIA

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

BIO DATA PENULIS

Penulis memiliki berbagai disiplin ilmu yang diperoleh dari Universitas Diponegoro (UNDIP) Semarang. dan dari Universitas Kristen Satya Wacana (UKSW) Salatiga. Disiplin ilmu itu antara lain teknik elektro, komputer, manajemen, ilmu sosiologi dan ilmu hukum. Penulis memiliki pengalaman kerja pada industri elektronik dan sertifikasi keahlian dalam bidang Jaringan Internet, Telekomunikasi, Artificial Intelligence, Internet Of Things (IoT), Augmented Reality (AR), Technopreneurship, Internet Marketing dan bidang pengolahan dan analisa data (komputer statistik), Ilmu Perpajakan.

Penulis adalah pendiri dari Universitas Sains dan Teknologi Komputer (Universitas STEKOM) dan juga seorang dosen yang memiliki Jabatan Fungsional Akademik Lektor Kepala (Associate Professor) yang telah menghasilkan puluhan Buku Ajar ber ISBN, HAKI dari beberapa karya cipta dan Hak Paten pada produk IPTEK. Sejak tahun 2023 penulis tercatat sebagai Dosen luar biasa di Fakultas Ekonomi & Bisnis (FEB) Universitas Diponegoro Semarang. Penulis juga terlibat dalam berbagai organisasi profesi dan industri yang terkait dengan dunia usaha dan industri, khususnya dalam pengembangan sumber daya manusia yang unggul untuk memenuhi kebutuhan dunia kerja secara nyata.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-634-7227-95-9 (PDF)
9 786347 227959