



Peran AI dalam **Etika, Tatakelola dan Kebijakan**

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



YAYASAN PRIMA AGUS TEKNIK



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

Peran AI dalam Etika, Tatakelola dan Kebijakan



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 978-634-7227-43-0 (PDF)



9

786347

227430

Peran AI dalam Etika, Tatakelola dan Kebijakan

Penulis :

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

ISBN : 978-634-7227-43-0

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 Yuniarto, 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

Hak cipta dilindungi undang-undang

Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara
apapun tanpa ijin dari penulis

KATA PENGANTAR

Puji syukur kita panjatkan ke hadirat Allah SWT atas limpahan rahmat dan karunia-Nya sehingga buku dengan judul *“Peran AI dalam Etika, Tatakelola, dan Kebijakan”* dapat terselesaikan. Kehadiran kecerdasan buatan (Artificial Intelligence/AI) di era digital saat ini membawa perubahan besar dalam hampir semua aspek kehidupan manusia. AI tidak hanya berperan sebagai teknologi canggih yang mendorong kemajuan, tetapi juga sebagai kekuatan transformasi yang memengaruhi cara manusia berpikir, mengambil keputusan, serta mengatur kehidupan sosialnya.

Makna yang terkandung dalam judul buku ini menegaskan pentingnya menempatkan AI dalam kerangka yang lebih luas. Teknologi ini memang menawarkan peluang besar, mulai dari peningkatan efisiensi hingga percepatan inovasi, namun pada saat yang sama menghadirkan tantangan serius yang berkaitan dengan keadilan, akuntabilitas, serta arah kebijakan yang mengaturnya. Oleh sebab itu, membicarakan AI tidak bisa dilepaskan dari pertimbangan etika dan tata kelola yang matang agar penerapannya benar-benar membawa manfaat bagi masyarakat.

Buku ini terdiri atas 18 bab yang membahas berbagai dimensi etika AI mulai dari konsep dasar hingga aplikasinya dalam konteks global, sosial, hukum, dan keamanan. Masing-masing bab dibuat untuk memaparkan secara runtut dan sistematis kerangka kerja, peluang, risiko, hingga solusi yang relevan dalam membangun AI yang bertanggung jawab dan berkelanjutan.

Bab 1 membuka pembahasan dengan etika dalam pengembangan AI, menguraikan kerangka kerja lima prinsip etika sebagai dasar pengembangan AI yang harmonis dengan masyarakat. Pada Bab 2, pembaca diajak memahami peluang dan risiko penggunaan AI bagi masyarakat serta keuntungan ganda dari pendekatan etis dalam pemanfaatan AI.

Bab 3 membahas aturan dan perencanaan yang cermat dalam membangun AI yang dapat dipercaya, serta pentingnya etika sebagai fondasi legislasi. Bab 4 memaparkan tata kelola AI dan diskursus etika digital di Tiongkok sebagai studi kasus dengan fokus strategis yang spesifik.

Bab 5 mengeksplorasi bagaimana prinsip-prinsip etika dijalankan dalam praktik sehari-hari, termasuk dinamika etika bluewashing dan lobi etika yang dapat memengaruhi persepsi publik. Bab 6 menegaskan potensi AI sebagai kekuatan untuk kebaikan, menitikberatkan pada aspek delegasi tanggung jawab dan etika translational.

Bab 7 membedah masalah dan solusi terkait etika algoritma, dengan penekanan pada dampak diskriminasi dan keadilan pada otonomi serta privasi informasi. Selanjutnya, Bab 8 menyajikan tujuh faktor esensial dalam merancang AI untuk kebaikan sosial yang ideal.

Bab 9 memberikan alat dan metode praktis etika AI bagi publik demi memahami dan mengimplementasikan etika secara lebih nyata dalam berbagai aplikasi. Bab 10 menyoroti pentingnya kerangka pembelajaran mesin yang dapat ditafsirkan agar transparansi dan tanggung jawab tetap terjaga.

Bab 11 mengupas agen buatan dan sifat moralnya, membahas aspek moralitas dan etika informasi yang melandasi interaksi antara manusia dan AI. Bab 12 melakukan analisis interdisipliner terhadap kejahatan berbasis AI dan solusi yang memungkinkan untuk menghadapinya.

Bab 13 mengelaborasi aturan kecerdasan buatan dalam mencegah perang siber, dengan fokus pada standar, prosedur, serta kerentanan AI dalam keamanan siber. Bab 14 mengangkat topik unik bot doa dan ibadah keagamaan di media sosial Twitter, melihat implikasi sosial dan budaya yang muncul.

Bab 15 membahas hubungan antara kecerdasan buatan, deepfake, dan masa depan etika pelayanan kesehatan, menyoroti perlunya proporsionalitas dan pertimbangan etika. Bab 16 memaparkan perkembangan kendaraan otonom, dengan fokus pada pendekatan dan implikasi etis dari perjalanan yang semakin otomatis.

Bab 17 fokus pada keadilan dan tata kelola AI dalam sektor keuangan, menyoroti bagaimana risiko AI dapat dikelola sepanjang siklus hidupnya. Bab 18 menutup dengan pembahasan tentang robot, pekerjaan, pajak, dan tanggung jawab, menggali masa depan AI dari sudut pandang integrasi sosial dan ekonomi.

Melalui buku ini, penulis berupaya mengajak pembaca untuk memahami bahwa perkembangan AI harus dilihat secara seimbang: sebagai sumber peluang yang dapat mendorong kemajuan, sekaligus sebagai fenomena yang menuntut regulasi dan pengawasan yang tepat. Dengan demikian, diskusi yang ditawarkan diharapkan dapat memberikan panduan bagi akademisi, praktisi, pembuat kebijakan, maupun masyarakat luas untuk memahami dinamika AI dari sisi teknologi, etika, dan tata kelola.

Penulis menyadari bahwa karya ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang. Semoga buku ini dapat memberi manfaat nyata, memperluas wawasan, serta menjadi inspirasi bagi siapa pun yang peduli terhadap arah perkembangan teknologi dan masa depan kemanusiaan.

Semarang, September 2025

Penulis

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

Di era digital yang terus berkembang pesat, AI bukan hanya soal teknologi semata, melainkan juga soal integrasi nilai-nilai etika dan tanggung jawab sosial agar dapat memberikan manfaat yang sebesar-besarnya sekaligus meminimalkan risiko yang mungkin ditimbulkannya. Semoga buku ini dapat menjadi rujukan yang bermanfaat bagi akademisi, pembuat kebijakan, praktisi teknologi, dan masyarakat luas yang ingin memahami dan menerapkan AI secara etis serta bertanggung jawab.

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

DAFTAR ISI

KATA PENGANTAR.....	ii
DAFTAR ISI.....	v
BAB 1 ETIKA DALAM PENGEMBANGAN AI.....	1
1.1 Pendahuluan.....	1
1.2 Kerangka Kerja Terpadu Lima Prinsip AI Dalam Masyarakat	4
1.3 Definisi Kecerdasan Buatan	5
1.4 Kerangka Kerja Terpadu Lima Prinsip Untuk AI Yang Etis	6
1.5 Etika AI: Dari Mana Dan Untuk Siapa?	11
1.6 Kesimpulan	12
BAB 2 PELUANG DAN RISIKO PENGGUNAAN AI	14
2.1 Pendahuluan.....	14
2.2 Peluang Dan Risiko Ai Bagi Masyarakat	15
2.3 Keuntungan Ganda Dari Pendekatan Etis Terhadap AI.....	19
2.4 Kerangka Prinsip Terpadu Untuk Ai Dalam Masyarakat	20
2.5 Rekomendasi Untuk Masyarakat Ai Yang Baik.....	26
2.6 Kesimpulan	31
BAB 3 ATURAN DALAM MEMBANGUN AI YANG TEPERCAYA	32
3.1 Perencanaan Yang Cermat, Alih-Alih Pengujian Beta	32
3.2 Etika Pertama Yang Menginformasikan Legislasi	33
3.3 Langkah Lanjutan Untuk Panggung Global.....	34
BAB 4 KEBIJAKAN, ETIKA, DAN REGULASI TERHADAP AI	36
4.1 Pendahuluan.....	36
4.2 Tata Kelola AI Di Tiongkok.....	37
4.3 Fokus Strategis AI Tiongkok	41
4.4 Perdebatan Tentang Etika Digital Dan AI Di Tiongkok	50
4.5 Kesimpulan	58
BAB 5 MENERJEMAHKAN PRINSIP KE DALAM PRAKTIK ETIKA DIGITAL.....	60
5.1 Pendahuluan.....	60
5.2 Belanja Etika	61
5.3 Etika Bluewashing.....	62
5.4 Lobi Etika	63
5.5 Pembuangan Etika	64
5.6 Pengabaian Etika	66
5.7 Kesimpulan	67
BAB 6 AI DAPAT MENJADI KEKUATAN UNTUK KEBAIKAN	68
6.1 Delegasi Dan Tanggung Jawab.....	69
6.2 Ketidaktampakan Dan Pengaruh	70

6.3	Etika Translasional	71
BAB 7	MASALAH DAN SOLUSI ETIKA ALGORITMA	73
7.1	Pendahuluan.....	73
7.2	Peta Etika Algoritma	74
7.3	Bukti Yang Tidak Konklusif	76
7.4	Bukti Yang Tak Terpenuhi Yang Menyebabkan Ketidakjelasan	77
7.5	Bukti Yang Salah Arah Menyebabkan Bias Yang Tidak Diinginkan	80
7.6	Hasil Yang Tidak Adil Yang Menyebabkan Diskriminasi	82
7.7	Dampak Transformatif Bagi Otonomi Dan Privasi Informasi	84
7.8	Ketertelusuran Yang Mengarah Pada Tanggung Jawab Moral	87
7.9	Kesimpulan	90
BAB 8	FAKTOR ESENSIAL MERANCANG AI UNTUK KEBAIKAN SOSIAL.....	91
8.1	Pendahuluan.....	91
8.2	Metodologi	93
8.3	Tujuh Faktor Esensial Untuk Keberhasilan Ai4sg	95
8.4	Kesimpulan	110
BAB 9	ALAT DAN METODE ETIKA AI UNTUK PUBLIK.....	113
9.1	Pendahuluan.....	113
9.2	Beralih Dari Prinsip Ke Praktik	116
9.3	Metodologi	119
9.4	Membangkit Hasil	123
9.5	Interpretasi Permodelan AI	124
9.6	Pandangan Masa Depan.....	129
9.7	Keterbatasan AI Publik.....	132
9.8	Kesimpulan	132
BAB 10	KERANGKA PEMBELAJARAN MESIN YANG DIINTERPRETASIKAN	135
10.1	Pendahuluan.....	135
10.2	Mengapa Menjelaskan Algoritma?	136
10.3	Latar Belakang Formal	139
10.4	Cakupan	144
10.5	Permainan Penjelasan	146
10.6	Pembelajaran Mesin Yang Diinterpretasikan	155
10.7	Keberatan	158
10.8	Kesimpulan	163
BAB 11	AGEN BUATAN DAN SIFAT MORALNYA.....	164
11.1	Pendahuluan.....	164
11.2	Apa Itu Agen?	167
11.3	Moralitas.....	177
11.4	Etika Informasi	184
11.5	Kesimpulan	187
BAB 12	ANALISIS INTERDISIPLINER KEJAHATAN AI DAN SOLUSINYA.....	189

12.1	Pendahuluan.....	189
12.2	Metodologi	192
12.3	Ancaman.....	193
12.4	Solusi Yang Mungkin Untuk Kejahatan Yang Didukung Kecerdasan Buatan	206
12.5	Kesimpulan	212
BAB 13	ATURAN KECERDASAN BUATAN DALAM MENCEGAH PERANG SIBER.....	215
13.1	Tanpa Aturan	216
13.2	Tiga Langkah	218
13.3	Ambiguitas Peran AI Dalam Keamanan Siber	218
13.4	Kerentanan AI	220
13.5	Standar Dan Prosedur Sertifikasi.....	221
13.6	Menjadikan AI Dalam Keamanan Siber Handal.....	222
BAB 14	BOT DOA DAN IBADAH KEAGAMAAN DI TWITTER.....	225
14.1	Pendahuluan.....	225
14.2	Aplikasi Salat Islam	226
14.3	Konteks Keagamaan	229
14.4	Implikasi Yang Lebih Luas	230
BAB 15	KECERDASAN BUATAN, DEEPFAKE, DAN MASA DEPAN EKTIPE	233
15.1	Etika AI Dalam Pelayanan Kesehatan	238
15.2	Metodologi	240
15.3	Temuan	244
15.4	Perlunya Etika Dan Proporsionalitas Dalam Pendekatan	257
15.5	Kesimpulan	262
BAB 16	KENDARAAN OTONOM.....	263
16.1	Pendahuluan.....	263
16.2	Perjalanan, Perjalanan, Dan Trip	264
16.3	Fokus Pada 'Bagaimana', Bukan 'Kapan' Atau 'Di Mana'	266
16.4	Kesimpulan	267
BAB 17	KEADILAN DAN TATA KELOLA AI DALAM SEKTOR KEUANGAN	268
17.1	Pendahuluan.....	268
17.2	Keadilan Dalam Industri Jasa Keuangan	270
17.3	Mengelola Risiko Ai Sepanjang Siklus Hidupnya	272
17.4	Desain Permodalan Ai Dalam Keuangan	273
17.5	Membangun	279
BAB 18	ROBOT, PEKERJAAN, PAJAK, DAN TANGGUNG JAWAB	283
18.1	Bagaimana Masa Depan Kecerdasan Buatan (Ai)	284
18.2	AI: Definisi Kerja	285
18.3	Masa Depan AI Dan Kebutuhan Akan Ludufikasi	287
18.4	Masa Depan AI Dari Masalah Sulit Kemasalah Kompleks	292
18.5	Kesimpulan	296
DAFTAR PUSTAKA		298

BAB 1

ETIKA DALAM PENGEMBANGAN AI

1.1 PENDAHULUAN

Perkembangan kecerdasan buatan (AI) telah melalui berbagai fase atau "musim" yang mencerminkan dinamika antara harapan, pencapaian, serta kekecewaan dalam penerapannya. Setiap periode tersebut tidak hanya membawa kemajuan teknologi, tetapi juga menimbulkan berbagai persoalan etis yang semakin kompleks. Keberhasilan AI dalam menyelesaikan tugas-tugas tertentu telah membuka peluang baru, namun di sisi lain, kegagalannya mengungkap risiko dan tantangan moral yang perlu mendapat perhatian serius. Oleh karena itu, pendekatan etika dalam pengembangan dan penggunaan AI menjadi semakin penting untuk memastikan bahwa teknologi ini dapat memberikan manfaat tanpa mengabaikan nilai-nilai kemanusiaan.

Masalah dengan metafora musiman adalah sifatnya yang siklus. Jika Anda mengatakan bahwa kecerdasan buatan (AI) melewati musim dingin yang buruk, Anda juga harus ingat bahwa musim dingin akan kembali, dan Anda sebaiknya bersiap. Musim dingin AI adalah tahap ketika teknologi, bisnis, dan media keluar dari gelembung hangat dan nyaman mereka, mendinginkan diri, meredam spekulasi fiksi ilmiah dan sensasi yang tidak masuk akal, dan memahami apa yang sebenarnya dapat atau tidak dapat dilakukan AI sebagai sebuah teknologi, tanpa berlebihan. Investasi menjadi lebih selektif, dan jurnalis berhenti menulis tentang AI, untuk mengejar topik-topik populer lainnya dan memicu tren berikutnya.

AI telah mengalami beberapa musim dingin. Di antara yang paling signifikan, ada satu di akhir tahun tujuh puluhan, dan satu lagi di pergantian tahun delapan puluhan dan sembilan puluhan. Hari ini, kita berbicara tentang musim dingin lain yang dapat diprediksi. AI tunduk pada siklus sensasi ini karena ia merupakan harapan atau ketakutan yang telah kita miliki sejak kita diusir dari surga: suatu bentuk agensi yang melakukan segalanya untuk kita, alih-alih kita, lebih baik dari kita, dengan semua keuntungan yang indah (kita akan berlibur selamanya) dan risiko yang mengerikan (kita akan diperbudak) yang menyertainya.

Bagi sebagian orang, berspekulasi tentang semua ini tak tertahankan. Ini adalah dunia liar dari skenario "bagaimana jika". Namun saya harap pembaca akan memaafkan saya untuk momen "Sudah kubilang". Untuk beberapa waktu, saya telah memperingatkan terhadap komentator dan "pakar", yang bersaing untuk melihat siapa yang bisa menceritakan kisah tertinggi. Sebuah jaring mitos pun terjadi. Mereka berbicara tentang AI seolah-olah itu adalah obat mujarab utama, yang akan menyelesaikan segalanya dan mengatasi segalanya; atau sebagai bencana terakhir, kecerdasan super yang akan menghancurkan jutaan pekerjaan, menggantikan pengacara dan dokter, jurnalis dan peneliti, pengemudi truk dan pengemudi taksi, dan berakhir dengan mendominasi manusia seolah-olah mereka adalah hewan peliharaan yang terbaik. Banyak yang mengikuti Elon Musk dalam menyatakan pengembangan AI sebagai risiko eksistensial terbesar yang dijalankan oleh umat manusia. Seolah-olah sebagian besar umat manusia tidak hidup dalam kesengsaraan dan penderitaan. Seolah-olah

perang, kelaparan, polusi, pemanasan global, ketidakadilan sosial, dan fundamentalisme adalah fiksi ilmiah, atau hanya gangguan yang dapat diabaikan, tidak layak untuk dipertimbangkan.

Mereka bersikeras bahwa hukum dan peraturan akan selalu terlambat dan tidak akan pernah bisa mengejar AI, padahal sebenarnya hukum dan norma bukanlah tentang kecepatan tetapi tentang arah inovasi, karena mereka harus mengarahkan perkembangan masyarakat yang tepat (jika kita suka ke mana kita menuju, kita tidak bisa pergi ke sana dengan cukup cepat). Saat ini, kita tahu bahwa undang-undang akan datang, setidaknya di UE. Mereka mengklaim AI adalah kotak hitam ajaib, yang tidak akan pernah bisa kita jelaskan, padahal sebenarnya itu adalah masalah tingkat abstraksi yang benar untuk menafsirkan interaksi kompleks yang direayasa bahkan lalu lintas mobil di pusat kota menjadi kotak hitam jika Anda ingin tahu mengapa setiap individu ada di sana pada saat itu. Saat ini ada perkembangan yang berkembang dari alat yang memadai untuk memantau dan memahami bagaimana sistem pembelajaran mesin mencapai hasilnya.

Mereka menyebarkan skeptisisme tentang kemungkinan kerangka etika yang akan mensintesis apa yang kita maksud dengan AI yang baik secara sosial, padahal kenyataannya Uni Eropa, OECD, dan Tiongkok telah menyatu pada prinsip-prinsip yang sangat mirip yang menawarkan platform bersama untuk perjanjian lebih lanjut. Sofis mencari berita utama. Mereka seharusnya malu dan meminta maaf. Tidak hanya karena komentar mereka yang tidak dapat dipertahankan, tetapi juga karena ketidakbertanggungjawaban dan alarmisme yang besar, yang telah menyesatkan opini publik baik tentang teknologi yang berpotensi berguna – yang dapat memberikan solusi yang membantu, dari obat-obatan hingga sistem keamanan dan pemantauan dan tentang risiko nyata yang kita tahu konkret tetapi jauh lebih sederhana, dari manipulasi pilihan sehari-hari hingga meningkatnya tekanan pada privasi individu dan kelompok, dari konflik siber hingga penggunaan AI oleh kejahatan terorganisir untuk pencucian uang dan pencurian identitas.

Risiko setiap musim panas AI adalah ekspektasi yang terlalu tinggi berubah menjadi gangguan massal. Risiko setiap musim dingin AI adalah reaksi baliknya berlebihan, kekecewaannya terlalu negatif, dan solusi yang berpotensi berharga terbuang bersama air ilusi. Mengelola dunia adalah tugas yang semakin kompleks: kota-kota besar dan "smartifikasi"-nya menawarkan contoh yang baik. Dan kita memiliki masalah planet seperti pemanasan global, ketidakadilan sosial, dan migrasi yang membutuhkan tingkat koordinasi yang semakin tinggi untuk dipecahkan.

Tampaknya jelas bahwa kita membutuhkan semua teknologi canggih yang dapat kita rancang, kembangkan, dan terapkan untuk mengatasi tantangan-tantangan ini, dan semua kecerdasan manusia yang dapat kita gunakan untuk menempatkan teknologi ini demi masa depan yang lebih baik. AI dapat memainkan peran penting dalam semua ini karena kita membutuhkan cara yang semakin cerdas untuk memproses data dalam jumlah besar, secara berkelanjutan dan efisien. Namun AI harus diperlakukan sebagai teknologi normal, bukan sebagai keajaiban atau wabah, melainkan sebagai salah satu dari sekian banyak solusi yang telah berhasil diciptakan oleh kecerdasan manusia. Inilah sebabnya mengapa perdebatan etika

merupakan dan akan selalu menjadi pertanyaan yang sepenuhnya manusiawi, dan sangat krusial, seperti yang ditunjukkan dalam buku ini.

Kini, dengan datangnya musim dingin yang baru, kita dapat mencoba belajar dari beberapa pelajaran, dan menghindari pasang surut ilusi yang tidak masuk akal dan kekecewaan yang berlebihan. Jangan lupa bahwa musim dingin AI seharusnya bukan musim dingin yang penuh peluang. Musim dingin ini tentu saja bukan musim dingin yang penuh risiko dan tantangan etika. Kita perlu bertanya pada diri sendiri apakah solusi AI benar-benar akan *menggantikan* solusi sebelumnya seperti yang telah dilakukan mobil dengan kereta mendiversifikasinya seperti yang dilakukan sepeda motor dengan sepeda atau *melengkapi* dan *memperluasnya* seperti yang telah dilakukan jam tangan pintar digital dengan jam tangan analog.

Bagaimana tingkat penerimaan atau preferensi sosial dalam cara AI bertahan di musim dingin yang baru? Apakah kita benar-benar akan mengenakan kacamata aneh untuk hidup di dunia virtual atau augmented yang diciptakan oleh AI? Pertimbangkan bahwa saat ini banyak orang enggan memakai kacamata meskipun mereka sangat membutuhkannya, hanya karena alasan estetika. Lalu, adakah solusi AI yang layak dalam kehidupan sehari-hari? Apakah keterampilan, kumpulan data, infrastruktur, dan model bisnis yang diperlukan sudah tersedia untuk menjadikan aplikasi AI berhasil? Para futurolog menganggap pertanyaan-pertanyaan ini membosankan.

Mereka menyukai satu ide tunggal yang sederhana, yang menafsirkan dan mengubah segalanya, yang dapat dijabarkan secara ringkas dalam sebuah buku sederhana yang membuat pembaca merasa cerdas, sebuah buku yang harus dibaca semua orang hari ini dan diabaikan oleh semua orang di masa mendatang. Ini adalah pola makan buruk makanan cepat saji yang buruk untuk pikiran dan kutukan buku terlaris di bandara. Kita perlu menolak penyederhanaan yang berlebihan. Kali ini mari kita berpikir lebih dalam dan ekstensif tentang apa yang kita lakukan dan rencanakan dengan AI. Latihan ini disebut filsafat, bukan futurologi.

Volume ini dimaksudkan untuk berkontribusi pada latihan berpikir yang lebih lambat dan lebih mendalam tersebut. Volume ini mengumpulkan beberapa hasil paling signifikan dari penelitian tentang etika AI yang dilakukan oleh anggota *Digital Ethics Lab* (DELab), kelompok penelitian OII yang saya pimpin di Universitas Oxford, juga bekerja sama dengan rekan-rekan lainnya. Bab-babnya telah muncul sebelumnya di berbagai jurnal internasional yang telah melalui penelaahan sejawat, tetapi belum pernah diterbitkan secara bersamaan. Demi konsistensi, isinya tidak diubah, hanya formatnya. Harapannya, pembaca tidak hanya merasa nyaman, tetapi juga terbantu secara intelektual dengan memiliki seluruh koleksi di satu tempat, untuk melihat pola dan perkembangan penalaran serta kesimpulan. Seiring dengan semakin terspesialisasinya perdebatan etika tentang AI, yang menjadi arus utama, dan berorientasi praktis, harapannya adalah bab dalam buku ini dapat membantu membangun fondasi yang kokoh untuk studi lebih lanjut. Realitas harapan ini hanya pembaca yang dapat menilainya.

1.2 KERANGKA KERJA TERPADU LIMA PRINSIP AI DALAM MASYARAKAT

Kecerdasan Buatan (AI) kini memainkan peran penting dalam berbagai aspek kehidupan, memberikan dampak yang signifikan terhadap masyarakat. Seiring dengan perkembangan ini, banyak organisasi dan institusi meluncurkan inisiatif untuk merumuskan prinsip-prinsip etika guna memastikan bahwa penerapan AI memberikan manfaat sosial yang luas. Namun, maraknya prinsip-prinsip yang diajukan justru menimbulkan kebingungan dan beban tersendiri, memunculkan tantangan yang dikenal sebagai proliferasi prinsip. Bab ini mengulas persoalan tersebut dengan melakukan analisis terhadap sejumlah prinsip etika AI yang paling dikenal secara global. Tujuannya adalah untuk menilai apakah terdapat konsensus terhadap prinsip-prinsip utama AI etis, atau justru terjadi perbedaan pandangan yang mendasar mengenai apa yang dimaksud dengan AI yang etis.

Hasil analisis menunjukkan bahwa, meskipun terdapat keragaman dalam perumusan, sebagian besar prinsip tersebut memiliki tingkat tumpang tindih yang tinggi. Berdasarkan temuannya, bab ini mengusulkan sebuah kerangka kerja menyeluruh yang terdiri dari lima prinsip inti untuk AI etis. Empat di antaranya merujuk pada prinsip-prinsip bioetika yang sudah mapan, yaitu: kebaikan, non-maleficence (tidak merugikan), otonomi, dan keadilan. Sebagai pelengkap, ditambahkan satu prinsip baru yang penting dalam konteks AI, yaitu eksplikabilitas sebuah konsep yang menggabungkan pemahaman epistemologis (bagaimana AI bekerja) dan tanggung jawab etis (siapa yang bertanggung jawab atas cara kerja tersebut). Bab ini juga membahas keterbatasan dari kerangka yang diusulkan serta mengevaluasi implikasinya terhadap penyusunan regulasi, standar teknis, dan praktik terbaik dalam pengembangan serta penerapan AI etis di berbagai sektor.

Kecerdasan Buatan (AI) telah memberikan dampak yang besar bagi masyarakat. Pertanyaan kuncinya adalah bagaimana, di mana, kapan, dan oleh siapa dampak AI akan dirasakan. Akibatnya, banyak organisasi telah meluncurkan berbagai inisiatif untuk menetapkan prinsip-prinsip etika dalam penerapan AI yang bermanfaat secara sosial. Sayangnya, banyaknya prinsip yang diusulkan berpotensi menjadi terlalu banyak dan membingungkan, sehingga menimbulkan dua masalah potensial. Entah berbagai rangkaian prinsip etika untuk AI serupa, yang menyebabkan pengulangan dan redundansi yang tidak perlu, atau, jika berbeda secara signifikan, akan mengakibatkan kebingungan dan ambiguitas. Hasil terburuknya adalah 'pasar untuk prinsip' di mana para pemangku kepentingan mungkin tergoda untuk 'berbelanja' prinsip yang paling menarik.

Bagaimana masalah 'proliferasi prinsip' ini dapat dipecahkan? Dalam makalah ini, kami melaporkan hasil analisis mendalam dari beberapa rangkaian prinsip etika AI yang paling terkenal. Kami menilai apakah prinsip-prinsip ini konvergen, dengan serangkaian prinsip yang disepakati, atau divergen, dengan perbedaan pendapat yang signifikan mengenai apa yang dimaksud dengan 'AI etis'. Analisis kami menemukan tingkat tumpang tindih yang tinggi di antara serangkaian prinsip yang kami analisis. Kami kemudian mengidentifikasi kerangka kerja menyeluruh yang terdiri dari lima prinsip inti untuk AI etis. Dalam diskusi selanjutnya, kami mencatat batasan dan menilai implikasi kerangka kerja etis ini bagi upaya di masa mendatang

untuk menciptakan hukum, aturan, standar, dan praktik terbaik bagi AI etis dalam berbagai konteks.

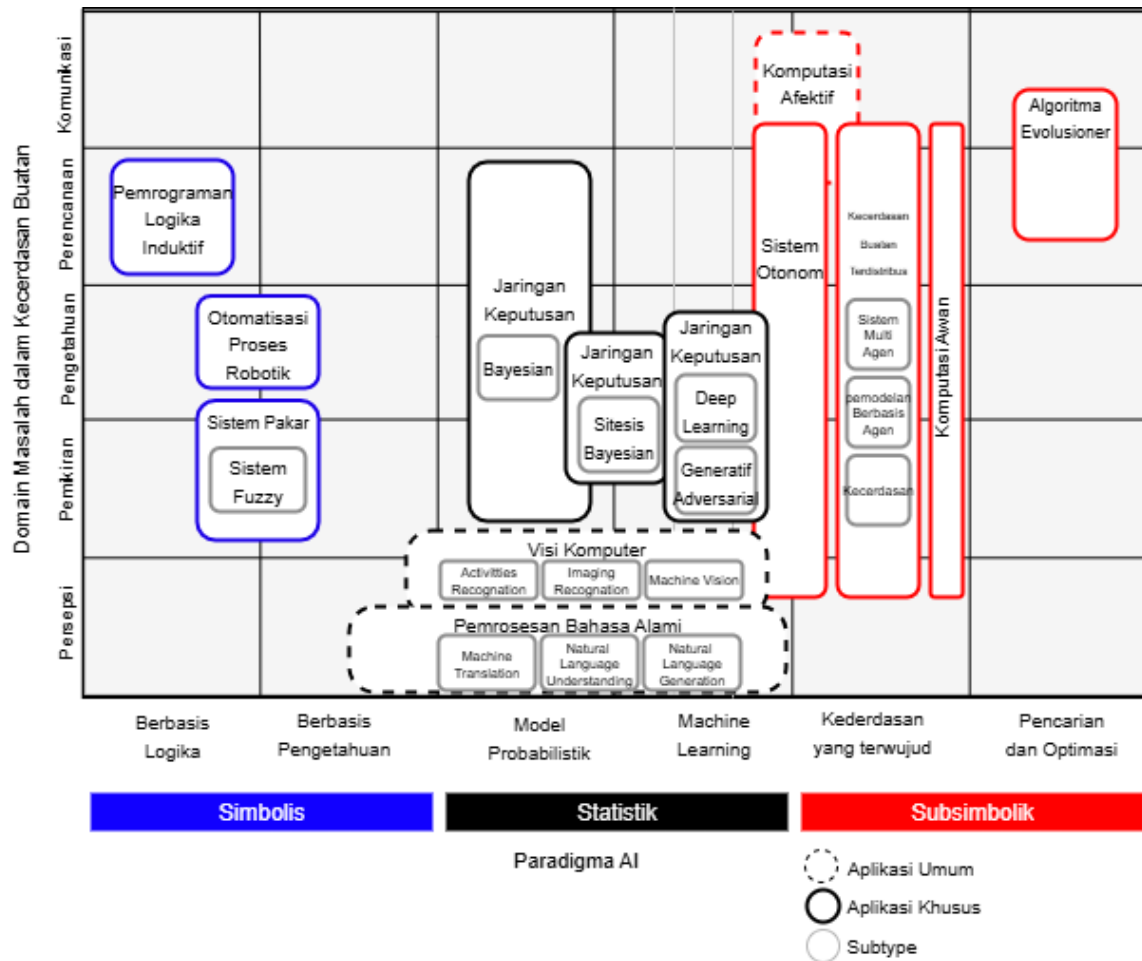
1.3 DEFINISI KECERDASAN BUATAN

AI telah didefinisikan dalam banyak cara. Saat ini, AI terdiri dari beberapa cabang tekno-ilmiah, yang dirangkum dengan baik dalam Gambar 1.1 (lihat juga artikel karya Dick dan Jordan dalam edisi ini untuk analisis yang mencerahkan). Secara keseluruhan, paradigma AI masih memenuhi definisi klasik yang diberikan oleh John McCarthy, Marvin Minsky, Nathaniel Rochester, dan Claude Shannon dalam Proposal penting mereka untuk Proyek Penelitian Musim Panas Dartmouth tentang Kecerdasan Buatan, dokumen pendiri dan peristiwa selanjutnya yang menetapkan bidang baru AI pada tahun 1955:

Untuk tujuan saat ini, masalah kecerdasan buatan dianggap sebagai masalah membuat mesin berperilaku dengan cara yang akan disebut cerdas jika manusia berperilaku demikian.

kontrafaktual: jika manusia berperilaku seperti itu, perilaku itu akan disebut cerdas. Ini tidak berarti bahwa mesin itu cerdas, atau bahkan berpikir. Skenario terakhir adalah kekeliruan, dan berbau takhayul. Hanya karena mesin pencuci piring membersihkan piring sebaik (atau bahkan lebih baik daripada) saya, bukan berarti ia membersihkannya seperti yang saya lakukan, atau membutuhkan kecerdasan apa pun untuk menyelesaikan tugasnya. Pemahaman kontrafaktual yang sama tentang AI mendasari uji Turing, yang, dalam hal ini, memeriksa kemampuan mesin untuk melakukan tugas sedemikian rupa sehingga hasilnya tidak dapat dibedakan dari hasil kerja agen manusia yang bekerja untuk mencapai tugas yang sama.

Definisi klasik memungkinkan kita untuk mengonseptualisasikan AI sebagai sumber daya yang terus berkembang, agensi yang interaktif, otonom, dan seringkali belajar mandiri (dalam pengertian pembelajaran mesin, lihat Gambar 1.1), yang dapat menangani tugas-tugas yang jika tidak demikian akan membutuhkan kecerdasan dan intervensi manusia agar berhasil dilakukan. Singkatnya, AI didefinisikan berdasarkan hasil dan tindakan yang direkayasa, sehingga, selanjutnya, kita akan memperlakukan AI sebagai sumber daya agensi cerdas yang siap pakai. Hal ini cukup umum untuk mencakup berbagai cara AI dibahas dalam dokumen-dokumen yang kami analisis di sisa artikel ini.



Gambar 1.1 Peta Pengetahuan AI (AIKM).

1.4 KERANGKA KERJA TERPADU LIMA PRINSIP UNTUK AI YANG ETIS

Penetapan kecerdasan buatan sebagai bidang penelitian akademis dimulai sejak tahun 1950-an. Perdebatan etika hampir sama tuanya. Namun, baru dalam beberapa tahun terakhir kemajuan yang mengesankan dalam kapabilitas dan aplikasi sistem AI telah mempertajam peluang dan risiko AI bagi masyarakat. Meningkatnya tuntutan akan refleksi dan kebijakan yang jelas tentang dampak AI terhadap masyarakat telah menghasilkan banyak inisiatif. Setiap inisiatif tambahan menghasilkan pernyataan prinsip, nilai, atau prinsip tambahan untuk memandu pengembangan dan adopsi AI. Risikonya adalah pengulangan dan tumpang tindih yang tidak perlu, jika berbagai rangkaian prinsip serupa, atau kebingungan dan ambiguitas, jika berbeda. Dalam kedua kemungkinan tersebut, pengembangan hukum, aturan, standar, dan praktik terbaik untuk memastikan AI bermanfaat secara sosial mungkin tertunda karena kebutuhan untuk menavigasi segudang prinsip dan deklarasi yang ditetapkan oleh serangkaian inisiatif yang terus berkembang.

Saatnya telah tiba untuk analisis komparatif atas dokumen-dokumen ini, termasuk penilaian apakah dokumen-dokumen tersebut konvergen atau divergen dan, jika yang pertama, apakah suatu kerangka kerja terpadu dapat disintesis. Untuk analisis komparatif ini,

kami mengidentifikasi enam inisiatif penting yang dibentuk demi kepentingan AI yang bermanfaat secara sosial:

1. Prinsip-Prinsip AI Asilomar, yang dikembangkan di bawah naungan Future of Life Institute, bekerja sama dengan para peserta konferensi tingkat tinggi Asilomar pada Januari 2017.
2. Deklarasi Montreal untuk AI yang Bertanggung Jawab, yang dikembangkan di bawah naungan Universitas Montreal, setelah Forum tentang Pengembangan AI yang Bertanggung Jawab Secara Sosial pada November 2017.
3. Prinsip-Prinsip Umum yang ditawarkan dalam versi kedua Desain yang Selaras Secara Etis: Sebuah Visi untuk Memprioritaskan Kesejahteraan Manusia dengan Sistem Otonom dan Cerdas. Risalah global yang bersumber dari masyarakat ini menerima kontribusi dari 250 pemimpin pemikiran global untuk mengembangkan prinsip dan rekomendasi bagi pengembangan dan perancangan sistem otonom dan cerdas yang etis, dan diterbitkan pada Desember 2017.
4. Prinsip-prinsip Etika yang ditawarkan dalam Pernyataan tentang Kecerdasan Buatan, Robotika, dan Sistem 'Otonom', diterbitkan oleh Kelompok Etika Eropa dalam Sains dan Teknologi Baru Komisi Eropa, pada Maret 2018.
5. 'Lima prinsip utama untuk kode AI' yang ditawarkan dalam laporan Komite Kecerdasan Buatan Dewan Perwakilan Rakyat Inggris, AI di Inggris: siap, bersedia, dan mampu?, yang diterbitkan pada bulan April 2018.
6. Prinsip-prinsip Kemitraan AI, sebuah organisasi multi-pemangku kepentingan yang terdiri dari akademisi, peneliti, organisasi masyarakat sipil, perusahaan yang membangun dan memanfaatkan teknologi AI, dan kelompok-kelompok lain.

Setiap rangkaian prinsip memenuhi tiga kriteria dasar: terbaru, diterbitkan dalam 3 tahun terakhir; relevan secara langsung dengan AI dan dampaknya terhadap masyarakat secara keseluruhan (dengan demikian tidak termasuk dokumen yang spesifik untuk domain, industri, atau sektor tertentu); dan sangat bereputasi, diterbitkan oleh organisasi multi-pemangku kepentingan yang berwibawa dengan setidaknya cakupan nasional. Jika digabungkan, mereka menghasilkan 47 prinsip. Secara keseluruhan, kami menemukan tingkat koherensi dan tumpang tindih antara enam set prinsip yang mengesankan dan meyakinkan. Konvergensi ini dapat ditunjukkan paling jelas dengan membandingkan set prinsip dengan empat prinsip inti yang umum digunakan dalam bioetika: kebaikan, non-maleficence, otonomi, dan keadilan. Perbandingan itu seharusnya tidak mengejutkan.

Dari semua bidang etika terapan, bioetika adalah yang paling mirip dengan etika digital dalam menangani secara ekologis bentuk-bentuk baru agen, pasien, dan lingkungan. Namun, meskipun keempat prinsip bioetika beradaptasi dengan sangat baik terhadap tantangan etika baru yang ditimbulkan oleh kecerdasan buatan, mereka tidak menawarkan terjemahan yang sempurna. Seperti yang akan kita lihat, makna yang mendasari masing-masing prinsip tersebut diperdebatkan, dengan istilah yang serupa sering digunakan untuk mengartikan hal yang berbeda. Keempat prinsip tersebut juga tidak lengkap.

Berdasarkan analisis komparatif kami, kami berpendapat bahwa diperlukan prinsip baru sebagai tambahan: eksplikabilitas, yang dipahami sebagai penggabungan antara keterpahaman (bagi non-pakar, misalnya, pasien atau pelanggan bisnis, dan bagi para pakar, misalnya, perancang produk atau insinyur) dan akuntabilitas. Namun, konvergensi yang kami temukan di antara rangkaian prinsip-prinsip yang berbeda ini juga menuntut kehati-hatian. Kami menjelaskan alasan kehati-hatian ini di bagian berikut, tetapi pertama-tama, kami akan memperkenalkan lima prinsip tersebut.

Kebaikan: Meningkatkan Kesejahteraan, Menjaga Martabat, dan Melestarikan Planet

Prinsip menciptakan teknologi AI yang bermanfaat bagi umat manusia diungkapkan dengan cara yang berbeda di keenam dokumen, tetapi mungkin merupakan prinsip yang paling mudah dipatuhi dari empat prinsip bioetika tradisional. Prinsip Montreal dan IEEE sama-sama menggunakan istilah "kesejahteraan"; bagi Montreal, "pengembangan AI pada akhirnya harus meningkatkan kesejahteraan semua makhluk hidup," sementara IEEE menyatakan perlunya "memprioritaskan kesejahteraan manusia sebagai hasil dalam semua desain sistem." AIUK dan Asilomar sama-sama menggolongkan prinsip ini sebagai "kebaikan bersama": AI harus "dikembangkan untuk kebaikan bersama dan manfaat umat manusia," menurut AIUK.

Kemitraan tersebut menjelaskan niat untuk "memastikan bahwa teknologi AI bermanfaat dan memberdayakan sebanyak mungkin orang", sementara EGE menekankan prinsip "martabat manusia" dan "keberlanjutan." Prinsip "keberlanjutan"-nya mengartikulasikan interpretasi beneficence yang mungkin paling luas, dengan argumen bahwa "teknologi AI harus sejalan dengan memastikan prasyarat dasar bagi kehidupan di planet kita, kesejahteraan umat manusia yang berkelanjutan, dan pelestarian lingkungan yang baik untuk generasi mendatang." Secara keseluruhan, keunggulan beneficence dengan tegas menggarisbawahi pentingnya mempromosikan kesejahteraan manusia dan planet ini dengan AI.

Non-maleficence: Privasi, Keamanan, dan 'Kehati-hatian Kapabilitas'

Meskipun 'hanya berbuat baik' (beneficence) dan 'tidak merugikan' (non-maleficence) mungkin tampak setara secara logis, keduanya tidak, dan mewakili prinsip yang berbeda. Meskipun keenam dokumen tersebut mendorong terciptanya AI yang beneficence, masing-masing juga memperingatkan terhadap berbagai konsekuensi negatif dari penggunaan teknologi AI yang berlebihan atau disalahgunakan. Yang menjadi perhatian khusus adalah pencegahan pelanggaran privasi pribadi, yang tercantum sebagai prinsip dalam lima dari enam set. Beberapa dokumen menekankan penghindaran penyalahgunaan teknologi AI dengan cara lain.

Prinsip-prinsip Asilomar memperingatkan ancaman perlombaan senjata AI dan peningkatan diri AI yang rekursif, sementara Kemitraan juga menegaskan pentingnya AI beroperasi "dalam batasan yang aman." Sementara itu, dokumen IEEE mengutip perlunya "menghindari penyalahgunaan," dan Deklarasi Montreal berpendapat bahwa mereka yang mengembangkan AI "harus memikul tanggung jawab mereka dengan melawan risiko yang timbul dari inovasi teknologi mereka." Namun dari berbagai peringatan ini, tidak sepenuhnya jelas apakah orang-orang yang mengembangkan AI, atau teknologi itu sendiri, yang harus

didorong untuk tidak melakukan tindakan berbahaya; dengan kata lain, apakah Frankenstein atau monsternya yang harus kita waspadai. Inti dari dilema ini adalah pertanyaan tentang otonomi.

Otonomi: Kekuatan untuk Memutuskan (untuk Memutuskan)

Ketika kita mengadopsi AI dan agensi cerdasnya, kita dengan sukarela menyerahkan sebagian kekuatan pengambilan keputusan kita kepada artefak teknologi. Dengan demikian, menegaskan prinsip otonomi dalam konteks AI berarti mencapai keseimbangan antara kekuatan pengambilan keputusan yang kita miliki untuk diri kita sendiri dan kekuatan yang kita delegasikan kepada agen-agen buatan. Risikonya adalah bahwa pertumbuhan otonomi buatan dapat melemahkan perkembangan otonomi manusia.

Oleh karena itu, tidak mengherankan bahwa prinsip otonomi dinyatakan secara eksplisit dalam empat dari enam dokumen. Deklarasi Montreal mengartikulasikan perlunya keseimbangan antara pengambilan keputusan yang dipimpin oleh manusia dan mesin, dengan menyatakan bahwa "pengembangan AI harus mendorong otonomi [cetak miring ditambahkan] semua manusia". EGE berpendapat bahwa sistem otonom "tidak boleh mengganggu kebebasan manusia untuk menetapkan standar dan norma mereka sendiri," sementara AIUK mengadopsi sikap yang lebih sempit bahwa "kekuatan otonom untuk menyakiti, menghancurkan, atau menipu manusia tidak boleh diberikan kepada AI." Dokumen Asilomar juga mendukung prinsip otonomi, sejauh "manusia harus memilih bagaimana dan apakah akan mendelegasikan keputusan kepada sistem AI, untuk mencapai tujuan yang dipilih manusia."

Oleh karena itu, jelas bahwa otonomi manusia harus dipromosikan dan bahwa otonomi mesin harus dibatasi dan dibuat secara intrinsik dapat dibalik, jika otonomi manusia perlu dilindungi atau dipulihkan (pertimbangkan kasus seorang pilot yang dapat mematikan autopilot dan mendapatkan kembali kendali penuh atas pesawat). Hal ini memperkenalkan gagasan yang dapat kita sebut 'meta-otonomi', atau model 'putuskan-untuk-mendelegasikan': manusia harus mempertahankan kekuasaan untuk memutuskan keputusan mana yang akan diambil: menjalankan kebebasan untuk memilih jika diperlukan, dan menyerahkannya dalam kasus-kasus di mana alasan-alasan yang lebih utama, seperti efikasi, dapat lebih besar daripada hilangnya kendali atas pengambilan keputusan. Setiap pendelegasian juga harus tetap dapat diabaikan secara prinsip (yaitu, memutuskan untuk memutuskan lagi).

Keadilan: Mempromosikan Kesejahteraan, Menjaga Solidaritas, Menghindari Ketidakadilan

Keputusan untuk membuat atau mendelegasikan keputusan tidak terjadi dalam ruang hampa. Kapasitas ini juga tidak didistribusikan secara merata di seluruh masyarakat. Konsekuensi dari disparitas otonomi ini dibahas dalam prinsip keadilan. Pentingnya 'keadilan' secara eksplisit disebutkan dalam Deklarasi Montreal, yang menyatakan bahwa "pengembangan AI harus mempromosikan keadilan dan berupaya menghilangkan semua jenis diskriminasi," sementara Prinsip-Prinsip Asilomar mencakup perlunya "manfaat bersama" dan "kemakmuran bersama" dari AI. Berdasarkan prinsipnya yang bernama "Keadilan, kesetaraan, dan solidaritas," EGE berpendapat bahwa AI harus "berkontribusi pada keadilan global dan akses yang setara terhadap manfaat" teknologi AI. EGE juga memperingatkan risiko bias dalam

kumpulan data yang digunakan untuk melatih sistem AI, dan—yang unik di antara dokumen-dokumen lainnya—menyarankan perlunya mempertahankan diri dari ancaman terhadap "solidaritas," termasuk "sistem bantuan bersama seperti dalam asuransi sosial dan layanan kesehatan."

Di tempat lain, 'keadilan' masih memiliki makna lain (terutama dalam arti kewajaran), yang beragam terkait dengan penggunaan AI untuk mengoreksi kesalahan masa lalu seperti menghapus diskriminasi yang tidak adil, mendorong keberagaman, dan mencegah munculnya ancaman baru terhadap keadilan. Beragamnya cara keadilan dicirikan mengisyaratkan kurangnya kejelasan yang lebih luas tentang AI sebagai reservoir 'agensi cerdas' buatan manusia. Sederhananya, apakah kita (manusia) pasien, yang menerima 'pengobatan' AI, dan dokter yang meresepkannya? Atau keduanya? Pertanyaan ini hanya dapat dijawab dengan memperkenalkan prinsip kelima yang muncul dari analisis kami.

Keterjelasan: Mengaktifkan Prinsip-Prinsip Lain Melalui Kecerdasan dan Akuntabilitas

Jawaban singkat untuk pertanyaan apakah 'kita' adalah pasien atau dokter adalah bahwa sebenarnya kita bisa menjadi salah satunya, tergantung pada keadaan dan siapa 'kita' dalam kehidupan sehari-hari. Situasinya secara inheren tidak setara: sebagian kecil umat manusia saat ini terlibat dalam pengembangan serangkaian teknologi yang telah mengubah kehidupan sehari-hari hampir semua orang.

Realitas pahit ini tidak luput dari perhatian para penulis yang dokumennya kami analisis. Semuanya merujuk pada kebutuhan untuk memahami dan mempertanggungjawabkan proses pengambilan keputusan AI. Istilah yang berbeda mengungkapkan prinsip ini: "transparansi" dalam Asilomar dan EGE; "transparansi" dan "akuntabilitas" dalam IEEE; "keterbacaan" dalam AIUK; dan sebagai "dapat dipahami dan ditafsirkan" oleh Kemitraan. Masing-masing prinsip ini menangkap sesuatu yang tampaknya baru tentang AI: bahwa cara kerjanya seringkali tidak terlihat atau tidak dapat dipahami oleh semua orang kecuali (paling banter) pengamat yang paling ahli.

Penambahan prinsip 'explicability,' yang menggabungkan baik arti epistemologis 'intelligibility' (sebagai jawaban atas pertanyaan 'bagaimana cara kerjanya?') dan dalam arti etis 'akuntabilitas' (sebagai jawaban atas pertanyaan 'siapa yang bertanggung jawab atas cara kerjanya?'), adalah bagian penting yang hilang dari teka-teki etika AI. Ini melengkapi empat prinsip lainnya: agar AI menjadi baik dan tidak jahat, kita harus dapat memahami kebaikan atau bahaya yang sebenarnya dilakukannya terhadap masyarakat, dan dengan cara apa; agar AI dapat mempromosikan dan tidak membatasi otonomi manusia, 'keputusan kita tentang siapa yang harus memutuskan' harus diinformasikan oleh pengetahuan tentang bagaimana AI akan bertindak, bukan kita; dan agar AI menjadi adil, kita harus tahu siapa yang harus dimintai pertanggungjawaban jika terjadi hasil yang serius dan negatif, yang pada gilirannya memerlukan pemahaman yang memadai tentang mengapa hasil ini muncul.

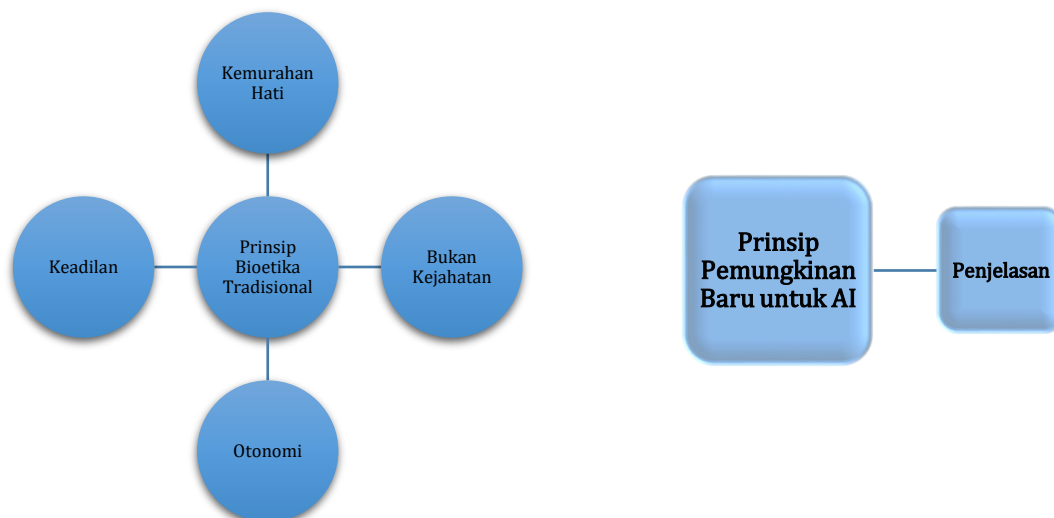
Pandangan Sinoptik

Secara keseluruhan, kelima prinsip ini mencakup keseluruhan dari 47 prinsip yang terdapat dalam enam dokumen penting yang dikaji oleh para ahli yang kami analisis. Lebih

lanjut, setiap prinsip terdapat dalam hampir setiap pernyataan prinsip yang kami analisis (lihat Tabel 1.1 di bawah).

Tabel 1.1 Lima prinsip dalam enam dokumen yang dianalisis dan kemunculannya dalam tiga dokumen terbaru

	Kebajikan (Beneficence)	Tidak Merugikan (Nonmaleficence)	Otonomi (Autonomy)	Keadilan (Justice)	Keterjelasan (Explicability)
AIUK	•	•	•	•	•
Asilomar	•	•	•	•	•
EGE	•	•	•	•	•
IEEE	•	•			•
Montreal	•	•	•	•	•
Partnership	•	•		•	•
AI4People	•	•	•	•	•
EC HLEG	•	•	•	•	•
OECD	•	•	•	•	•



Gambar 1.2 Kerangka etika dari lima prinsip utama AI yang muncul dari analisis

Kelima prinsip ini kemudian membentuk kerangka etika yang dapat digunakan untuk menyusun kebijakan, praktik terbaik, dan rekomendasi lainnya. Kerangka prinsip ini ditunjukkan pada Gambar 1.2.

1.5 ETIKA AI: DARI MANA DAN UNTUK SIAPA?

Penting untuk dicatat bahwa masing-masing dari enam rangkaian prinsip etika AI yang kami analisis muncul baik dari inisiatif berskala global maupun dari dalam demokrasi liberal Barat. Agar kerangka kerja ini dapat diterapkan secara lebih luas, niscaya akan lebih baik jika didukung oleh perspektif wilayah dan budaya yang saat ini kurang atau belum terwakili dalam

sampel kami. Yang menarik dalam hal ini adalah peran Tiongkok, yang telah menjadi rumah bagi perusahaan rintisan AI paling berharga di dunia, menikmati berbagai keunggulan struktural dalam mengembangkan AI, dan yang pemerintahnya telah menyatakan ambisinya untuk memimpin dunia dalam teknologi AI mutakhir pada tahun 2030 (Dewan Negara Tiongkok 2017). Dalam Pemberitahuan Dewan Negara tentang AI dan di tempat lain, pemerintah Tiongkok telah menyatakan minatnya untuk mempertimbangkan lebih lanjut dampak sosial dan etika AI. Antusiasme terhadap penggunaan teknologi juga tidak hanya dirasakan oleh pemerintah, tetapi juga dirasakan oleh masyarakat umum lebih banyak di Tiongkok dan India daripada di Eropa atau AS, sebagaimana ditunjukkan oleh penelitian survei representatif terbaru.

Seorang eksekutif di perusahaan teknologi besar Tiongkok, Tencent, baru-baru ini menyarankan agar Uni Eropa berfokus pada pengembangan AI yang memiliki "manfaat maksimal bagi kehidupan manusia, meskipun teknologi tersebut tidak kompetitif untuk bersaing di pasar Amerika atau Tiongkok". Hal ini juga didukung oleh klaim bahwa etika mungkin merupakan "solusi ajaib Eropa" dalam "pertempuran AI global". Kami tidak sependapat. Etika bukanlah hak milik satu benua atau budaya. Setiap perusahaan, lembaga pemerintah, dan institusi akademis yang merancang, mengembangkan, atau menerapkan AI memiliki kewajiban untuk melakukannya sesuai dengan kerangka etika yang sejalan dengan kerangka yang kami sajikan di sini, yang diperluas untuk menggabungkan beragam perspektif secara geografis, budaya, dan sosial. Demikian pula, undang-undang, peraturan, standar, dan praktik terbaik untuk membatasi atau mengendalikan AI termasuk semua yang saat ini sedang dipertimbangkan oleh badan pengatur, legislatif, dan kelompok industri juga akan mendapat manfaat dari keterlibatan erat dengan kerangka prinsip etika yang terpadu.

1.6 KESIMPULAN

Dari Prinsip ke Praktik

Jika kerangka kerja yang disajikan dalam artikel ini memberikan gambaran umum yang koheren dan cukup komprehensif tentang prinsip-prinsip etika inti AI, maka kerangka kerja tersebut dapat berfungsi sebagai arsitektur yang di dalamnya hukum, aturan, standar teknis, dan praktik terbaik dikembangkan untuk sektor, industri, dan yurisdiksi tertentu. Dalam konteks ini, kerangka kerja tersebut dapat memainkan peran yang memungkinkan (misalnya, penggunaan AI untuk membantu mencapai Tujuan Pembangunan Berkelanjutan Perserikatan Bangsa-Bangsa), dan peran yang membatasi.

Kerangka kerja ini memang memainkan peran penting dalam karya AI4People, forum global pertama di Eropa tentang dampak sosial AI, yang baru-baru ini mengadopsinya untuk mengusulkan 20 rekomendasi konkret bagi 'Masyarakat AI yang Baik' kepada Komisi Eropa. Sejak saat itu, hal ini telah diadopsi secara luas oleh Pedoman Etika untuk AI yang Dapat Dipercaya yang diterbitkan oleh Kelompok Pakar Tingkat Tinggi Komisi Eropa tentang Kecerdasan Buatan, yang pada gilirannya telah memengaruhi Rekomendasi Dewan OECD tentang Kecerdasan Buatan, yang menjangkau 42 negara (lihat Tabel 1.1).

Pengembangan dan penggunaan AI berpotensi memberikan dampak positif dan negatif bagi masyarakat, untuk mengurangi atau memperkuat ketimpangan yang ada, untuk menyembuhkan masalah lama, atau untuk menimbulkan masalah baru. Memetakan arah yang lebih disukai secara sosial tidak hanya akan bergantung pada regulasi yang dirancang dengan baik dan standar umum, tetapi juga pada penggunaan kerangka kerja prinsip-prinsip etika, yang di dalamnya tindakan konkret dapat ditempatkan. Kami percaya bahwa kerangka kerja yang disajikan di sini yang muncul dari perdebatan saat ini akan berfungsi sebagai arsitektur yang berharga untuk mengamankan hasil sosial yang positif dari teknologi AI dan beralih dari prinsip-prinsip yang baik ke praktik yang baik.

BAB 2

PELUANG DAN RISIKO PENGGUNAAN AI

Bab ini mengangkat hasil dari AI4People, sebuah inisiatif selama satu tahun yang bertujuan membangun landasan bagi terbentuknya “Masyarakat AI yang Baik”. Inisiatif ini berfokus pada bagaimana kecerdasan buatan (AI) dapat dikembangkan dan diterapkan secara etis untuk mendukung kesejahteraan sosial secara luas. Dalam pembahasan ini, kami mengidentifikasi peluang serta risiko utama yang ditimbulkan AI bagi masyarakat, sekaligus menekankan perlunya pendekatan etika yang kuat dalam proses pengembangan dan implementasinya. Sebagai dasar normatif, kami menyajikan sintesis dari lima prinsip etika utama yang seharusnya menjadi pijakan dalam pengembangan dan penggunaan AI.

Selain itu, bab ini juga mengusulkan 20 rekomendasi praktis yang dapat diterapkan oleh berbagai pihak, baik pembuat kebijakan di tingkat nasional dan supranasional, maupun pemangku kepentingan lainnya, seperti sektor industri, akademisi, dan masyarakat sipil. Rekomendasi-rekomendasi ini dirancang untuk menilai, mengembangkan, memberi insentif, dan mendukung penerapan AI yang etis dan bermanfaat secara sosial. Jika diimplementasikan secara menyeluruh, rekomendasi ini dapat menjadi fondasi yang kuat dalam membangun kerangka kerja dan kebijakan untuk menciptakan Masyarakat AI yang Baik sebuah masyarakat yang memanfaatkan teknologi secara bertanggung jawab demi kebaikan bersama.

2.1 PENDAHULUAN

AI bukanlah utilitas lain yang perlu diregulasi setelah matang. AI merupakan kekuatan yang dahsyat, sebuah bentuk baru agensi cerdas, yang telah membentuk kembali kehidupan kita, interaksi kita, dan lingkungan kita. AI4People dibentuk untuk membantu mengarahkan kekuatan dahsyat ini menuju kebaikan masyarakat, semua orang di dalamnya, dan lingkungan yang kita bagi bersama. Buku Putih ini merupakan hasil upaya kolaboratif Komite Ilmiah AI4People yang terdiri dari 12 pakar dan diketuai oleh Luciano Floridi untuk mengusulkan serangkaian rekomendasi bagi pengembangan Masyarakat AI yang Baik.

Buku Putih ini merangkum tiga hal: *peluang* dan *risiko* terkait yang ditawarkan oleh teknologi AI untuk meningkatkan martabat manusia dan memajukan kesejahteraan manusia; *prinsip-prinsip* yang seharusnya mendasari adopsi AI; dan dua puluh *rekomendasi* spesifik yang, jika diadopsi, akan memungkinkan semua pemangku kepentingan untuk memanfaatkan peluang, menghindari atau setidaknya meminimalkan dan menyeimbangkan risiko, menghormati prinsip-prinsip tersebut, dan dengan demikian mengembangkan Masyarakat AI yang Baik.

Buku ini disusun berdasarkan empat bagian lagi setelah pendahuluan ini. Bagian 2.2 menyatakan peluang inti untuk mempromosikan martabat manusia dan kesejahteraan manusia yang ditawarkan oleh AI, beserta risiko-risikonya. Bagian 2.3 menawarkan gambaran singkat dan mendalam tentang keuntungan bagi organisasi dalam mengambil pendekatan etis terhadap pengembangan dan penggunaan AI. Bagian 2.4 merumuskan 5 prinsip etika untuk

AI, berdasarkan analisis yang ada, yang seharusnya mendasari adopsi AI yang etis di masyarakat luas. Terakhir, Bagian 2.5 menawarkan 20 rekomendasi untuk tujuan mengembangkan Masyarakat AI yang Baik di Eropa. Sejak peluncuran AI4People pada Februari 2018, Komite Ilmiah telah bertindak kolaboratif untuk mengembangkan rekomendasi di bagian akhir makalah ini. Melalui karya ini, kami berharap dapat berkontribusi pada fondasi Masyarakat AI yang Baik yang dapat kita semua bagi bersama.

2.2 PELUANG DAN RISIKO AI BAGI MASYARAKAT

Bahwa AI akan berdampak besar pada masyarakat tidak lagi dipertanyakan. Perdebatan saat ini justru berkisar pada seberapa jauh dampak ini akan positif atau negatif, bagi siapa, dengan cara apa, di tempat mana, dan dalam jangka waktu berapa. Dengan kata lain, kita dapat dengan aman menyingkirkan pertanyaan apakah AI akan berdampak; pertanyaan yang relevan sekarang adalah *oleh siapa, bagaimana, di mana, dan kapan* dampak positif atau negatif ini akan dirasakan. Untuk membingkai pertanyaan-pertanyaan ini dengan cara yang lebih substantif dan praktis, di sini kami memperkenalkan apa yang kami anggap sebagai empat peluang utama bagi masyarakat yang ditawarkan AI.

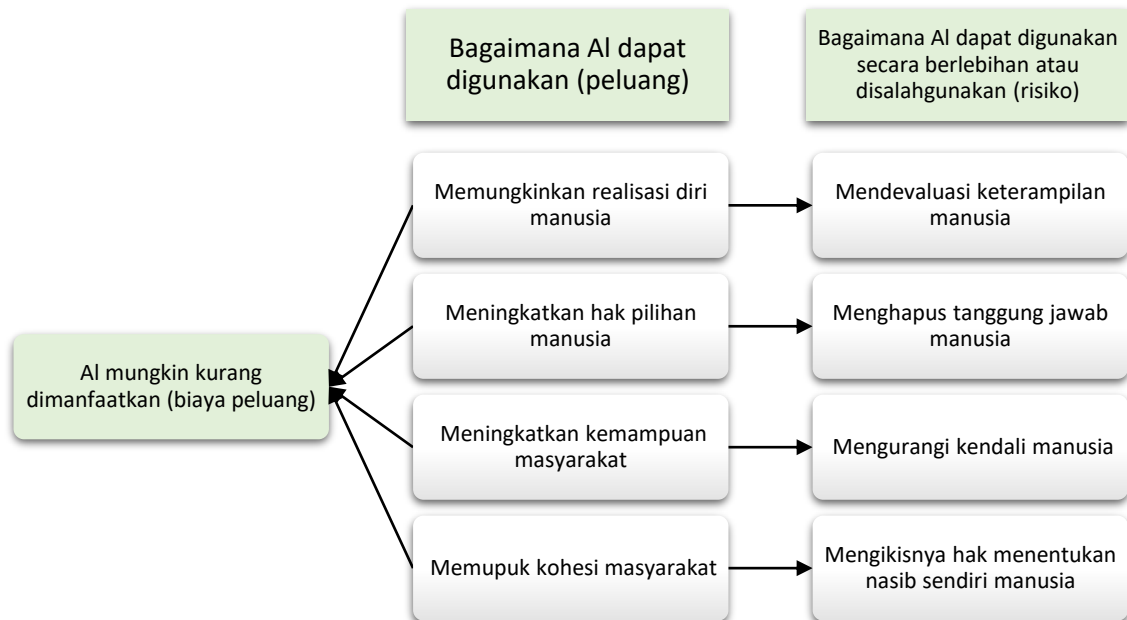
Keempatnya karena membahas empat poin fundamental dalam pemahaman tentang martabat dan perkembangan manusia: *siapa kita bisa menjadi* (realisasi diri otonom); *apa yang dapat kita lakukan* (agensi manusia); *apa yang dapat kita capai* (kemampuan individu dan masyarakat); *dan bagaimana kita dapat berinteraksi satu sama lain dan dunia* (kohesi masyarakat). Dalam setiap kasus, AI dapat *digunakan* untuk mengembangkan sifat manusia dan potensinya, sehingga menciptakan peluang; *kurang dimanfaatkan*, sehingga menciptakan biaya peluang; *terlalu sering digunakan* dan *disalahgunakan*, sehingga menciptakan risiko.

Sebagaimana ditunjukkan oleh terminologi tersebut, asumsinya adalah bahwa *penggunaan* AI identik dengan inovasi yang baik dan penerapan positif dari teknologi ini. Namun, ketakutan, ketidaktahuan, kekhawatiran yang salah tempat, atau reaksi yang berlebihan dapat menyebabkan masyarakat *kurang memanfaatkan* teknologi AI di bawah potensi penuhnya, untuk apa yang secara umum dapat digambarkan sebagai alasan yang salah. Hal ini dapat menyebabkan biaya peluang yang signifikan. Ini mungkin termasuk, misalnya, regulasi yang terlalu ketat atau salah paham, kurangnya investasi, atau reaksi publik yang serupa dengan yang dihadapi oleh tanaman rekayasa genetika.

Akibatnya, manfaat yang ditawarkan oleh teknologi AI mungkin tidak sepenuhnya disadari oleh masyarakat. Bahaya-bahaya ini sebagian besar muncul dari konsekuensi yang tidak diinginkan dan biasanya berkaitan dengan niat baik yang salah sasaran. Namun, kita juga harus mempertimbangkan risiko yang terkait dengan *penggunaan berlebihan* yang tidak disengaja atau *penyalahgunaan* teknologi AI yang disengaja, yang didasarkan, misalnya, pada insentif yang tidak selaras, keserakahan, geopolitik yang berseteru, atau niat jahat.

Segala hal, mulai dari penipuan email hingga perang siber skala penuh, dapat dipercepat atau diintensifkan oleh penggunaan teknologi AI yang berbahaya. Dan kejahatan baru mungkin saja terjadi. Kemungkinan kemajuan sosial yang diwakili oleh peluang-peluang yang disebutkan di atas harus dipertimbangkan dengan risiko bahwa manipulasi berbahaya

akan dimungkinkan atau ditingkatkan oleh AI. Namun, risiko yang lebih luas adalah AI mungkin kurang dimanfaatkan karena takut akan penggunaan berlebihan atau penyalahgunaan. Kami merangkum risiko-risiko ini pada Gambar 2.1 di bawah ini, dan menawarkan penjelasan yang lebih rinci dalam teks berikut.



Gambar 2.1 Tinjauan umum dari empat peluang inti yang ditawarkan oleh AI, empat risiko yang terkait, dan biaya peluang dari kurangnya penggunaan AI

Memungkinkan Realisasi Diri Manusia, Tanpa Meremehkan Kemampuan Manusia

AI dapat memungkinkan realisasi diri, yang kami maksud adalah kemampuan manusia untuk berkembang sesuai karakteristik, minat, potensi kemampuan atau keterampilan, aspirasi, dan proyek kehidupan mereka sendiri. Sebagaimana penemuan, seperti mesin cuci, telah membebaskan manusia terutama perempuan dari pekerjaan rumah tangga yang membosankan, otomatisasi "cerdas" pada aspek-aspek kehidupan lainnya yang biasa dapat membebaskan lebih banyak waktu untuk kegiatan budaya, intelektual, dan sosial, serta pekerjaan yang lebih menarik dan memuaskan.

Lebih banyak AI dapat dengan mudah berarti lebih banyak kehidupan manusia yang dihabiskan dengan lebih cerdas. Risiko dalam kasus ini bukanlah keusangan beberapa keterampilan lama dan munculnya keterampilan baru itu *sendiri*, melainkan kecepatan terjadinya hal ini dan distribusi biaya dan manfaat yang tidak merata. Devaluasi keterampilan lama yang sangat cepat dan dengan demikian disruptif cepat pasar kerja dan sifat pekerjaan dapat dilihat pada tingkat individu dan masyarakat. Pada tingkat individu, pekerjaan sering kali terkait erat dengan identitas pribadi, harga diri, dan peran atau kedudukan sosial, semua faktor yang mungkin terpengaruh secara negatif oleh redundansi, bahkan mengesampingkan potensi kerugian ekonomi yang parah.

Lebih lanjut, pada tingkat masyarakat, penurunan keterampilan dalam domain yang sensitif dan padat keterampilan, seperti diagnosis perawatan kesehatan atau penerbangan,

dapat menciptakan kerentanan berbahaya jika terjadi malfungsi AI atau serangan musuh. Membina pengembangan AI untuk mendukung kemampuan dan keterampilan baru, sambil mengantisipasi dan memitigasi dampaknya terhadap yang lama akan membutuhkan studi yang cermat dan ide-ide yang berpotensi radikal, seperti proposal untuk beberapa bentuk "pendapatan dasar universal", yang semakin populer dan digunakan secara eksperimental. Pada akhirnya, kita membutuhkan solidaritas antargenerasi antara mereka yang kurang beruntung hari ini dan mereka yang diuntungkan besok, untuk memastikan bahwa transisi yang mengganggu antara masa kini dan masa depan akan seadil mungkin, untuk semua orang.

Apa yang Dapat Kita Lakukan: Meningkatkan Agensi Manusia, Tanpa Menghilangkan Tanggung Jawab Manusia

AI menyediakan sumber daya "agensi cerdas" yang terus bertambah. Dengan memanfaatkan kecerdasan manusia, sumber daya tersebut dapat meningkatkan agensi manusia secara signifikan. Kita dapat melakukan lebih banyak, lebih baik, dan lebih cepat, berkat dukungan yang diberikan oleh AI. Dalam pengertian "Kecerdasan Tertambah" ini, AI dapat dibandingkan dengan dampak yang ditimbulkan oleh mesin terhadap kehidupan kita. Semakin banyak orang yang akan menikmati peluang dan manfaat dari sumber daya agensi cerdas yang "siap pakai", semakin baik pula masyarakat kita nantinya.

Oleh karena itu, tanggung jawab sangatlah penting, mengingat jenis AI yang kita kembangkan, bagaimana kita menggunakannya, dan apakah kita berbagi keuntungan dan manfaatnya dengan semua orang. Tentu saja, risiko yang terkait adalah ketiadaan tanggung jawab tersebut. Hal ini mungkin terjadi bukan hanya karena kita memiliki kerangka sosio-politik yang salah, tetapi juga karena mentalitas "kotak hitam", yang menganggap sistem AI untuk pengambilan keputusan berada di luar pemahaman manusia, dan karenanya berada di luar kendali. Kekhawatiran ini tidak hanya berlaku untuk kasus-kasus yang menonjol, seperti kematian yang disebabkan oleh kendaraan otonom, tetapi juga untuk penggunaan yang lebih umum tetapi masih signifikan, seperti dalam keputusan otomatis tentang pembebasan bersyarat atau kelayakan kredit.

Namun, hubungan antara tingkat dan kualitas agensi yang dinikmati orang dan seberapa banyak agensi yang kita delegasikan ke sistem otonom bukanlah zero-sum, baik secara pragmatis maupun etis. Bahkan, jika dikembangkan dengan cermat, AI menawarkan peluang untuk *meningkatkan dan melipatgandakan* kemungkinan agensi manusia. Pertimbangkan contoh "moralitas terdistribusi" dalam sistem antarmanusia seperti pinjaman peer-to-peer. Agensi manusia pada akhirnya dapat didukung, disempurnakan, dan diperluas dengan menanamkan "kerangka kerja yang memfasilitasi", yang dirancang untuk meningkatkan kemungkinan hasil yang baik secara moral, dalam serangkaian fungsi yang kita delegasikan ke sistem AI. Sistem AI dapat, jika dirancang secara efektif, memperkuat dan memperkuat sistem moral bersama.

Apa yang Dapat Kita Capai: Meningkatkan Kapabilitas Masyarakat, Tanpa Mengurangi Kendali Manusia

Kecerdasan buatan menawarkan segudang peluang untuk meningkatkan dan memperkaya kapabilitas individu dan masyarakat luas. Baik dengan mencegah dan

menyembuhkan penyakit maupun mengoptimalkan transportasi dan logistik, penggunaan teknologi AI menghadirkan segudang kemungkinan untuk menciptakan kembali masyarakat dengan secara radikal meningkatkan kemampuan kolektif manusia. Lebih banyak AI dapat mendukung koordinasi yang lebih baik, dan karenanya tujuan yang lebih ambisius. Kecerdasan manusia yang diperkuat oleh AI dapat menemukan solusi baru untuk masalah lama dan baru, mulai dari distribusi sumber daya yang lebih adil atau lebih efisien hingga pendekatan konsumsi yang lebih berkelanjutan. Justru karena teknologi semacam itu berpotensi begitu kuat dan disruptif, teknologi tersebut juga menghadirkan risiko yang sepadan.

Semakin lama, kita mungkin tidak perlu lagi 'berada di dalam atau di tengah-tengah' (yaitu, sebagai bagian dari proses atau setidaknya mengendalikannya), jika kita dapat mendelegasikan tugas kita kepada AI. Namun, jika kita mengandalkan penggunaan teknologi AI untuk meningkatkan kemampuan kita sendiri dengan cara yang salah, kita mungkin mendelegasikan tugas-tugas penting dan terutama keputusan-keputusan kepada sistem-sistem otonom yang seharusnya tetap berada di bawah pengawasan dan pilihan manusia, setidaknya sebagian.

Hal ini pada gilirannya dapat mengurangi kemampuan kita untuk memantau kinerja sistem-sistem ini (dengan tidak lagi berada dalam 'loop') atau mencegah atau memperbaiki kesalahan atau kerugian yang muncul ('post-loop'). Potensi kerugian ini juga mungkin terakumulasi dan mengakar, karena semakin banyak fungsi yang didelegasikan kepada sistem-sistem buatan. Oleh karena itu, sangat penting untuk mencapai keseimbangan antara mengejar peluang ambisius yang ditawarkan oleh AI untuk meningkatkan kehidupan manusia dan apa yang dapat kita capai, di satu sisi, dan, di sisi lain, memastikan bahwa kita tetap mengendalikan perkembangan-perkembangan besar ini dan dampaknya.

Bagaimana Kita Dapat Berinteraksi: Memupuk Kohesi Masyarakat, Tanpa Mengikis Hak Penentuan Nasib Sendiri Manusia

Dari perubahan iklim dan resistensi antimikroba hingga proliferasi nuklir dan fundamentalisme, permasalahan global semakin memiliki tingkat kompleksitas koordinasi yang tinggi, yang berarti bahwa permasalahan tersebut hanya dapat ditangani dengan sukses jika semua pemangku kepentingan bersama-sama merancang dan memiliki solusinya serta bekerja sama untuk mewujudkannya. AI, dengan solusinya yang intensif data dan berbasis algoritmik, dapat sangat membantu dalam menangani kompleksitas koordinasi tersebut, mendukung kohesi dan kolaborasi masyarakat yang lebih besar. Misalnya, upaya untuk mengatasi perubahan iklim telah mengungkap tantangan dalam menciptakan respons yang kohesif, baik di dalam maupun di antara masyarakat.

Skala tantangan ini sedemikian rupa sehingga kita mungkin perlu segera memutuskan antara merekayasa iklim secara langsung dan merancang kerangka kerja masyarakat untuk mendorong pengurangan emisi berbahaya secara drastis. Pilihan terakhir ini mungkin didukung oleh sistem algoritmik untuk menumbuhkan kohesi masyarakat. Sistem semacam itu tidak akan dipaksakan dari luar; Ini akan menjadi hasil dari pilihan yang kita buat sendiri, tidak seperti pilihan kita untuk tidak membeli coklat jika sebelumnya kita memilih untuk berdiet, atau menyetel alarm untuk bangun. "Mendorong diri sendiri" untuk berperilaku dengan cara

yang lebih disukai secara sosial adalah bentuk dorongan terbaik, dan satu-satunya yang mempertahankan otonomi. Ini adalah hasil dari keputusan dan pilihan manusia, tetapi dapat bergantung pada solusi AI untuk diimplementasikan dan difasilitasi.

Namun risikonya adalah sistem AI dapat mengikis penentuan nasib sendiri manusia, karena dapat menyebabkan perubahan yang tidak direncanakan dan tidak diinginkan dalam perilaku manusia untuk mengakomodasi rutinitas yang membuat otomatisasi berfungsi dan kehidupan manusia lebih mudah. Kekuatan prediktif AI dan dorongan tanpa henti, meskipun tidak disengaja, seharusnya melayani penentuan nasib sendiri manusia dan memupuk kohesi masyarakat, bukan merusak martabat manusia atau kemajuan manusia.

Secara keseluruhan, keempat peluang ini, dan tantangannya yang terkait, melukiskan gambaran yang beragam tentang dampak AI terhadap masyarakat dan orang-orang di dalamnya. Menerima adanya trade-off, memanfaatkan peluang sambil berupaya mengantisipasi, menghindari, atau meminimalkan risiko secara langsung akan meningkatkan prospek teknologi AI dalam mempromosikan martabat dan kemajuan manusia. Setelah menguraikan potensi manfaat bagi individu dan masyarakat luas dari pendekatan AI yang terlibat secara etis, di bagian selanjutnya kami menyoroti "keuntungan ganda" bagi organisasi yang mengambil pendekatan tersebut.

2.3 KEUNTUNGAN GANDA DARI PENDEKATAN ETIS TERHADAP AI

Memastikan hasil AI yang lebih disukai secara sosial bergantung pada penyelesaian ketegangan antara mengintegrasikan manfaat dan memitigasi potensi bahaya AI, singkatnya, secara bersamaan menghindari penyalahgunaan dan kurangnya penggunaan teknologi ini. Dalam konteks ini, nilai pendekatan etis terhadap teknologi AI menjadi lebih jelas. Kepatuhan terhadap hukum hanyalah perlu (setidaknya yang diwajibkan), tetapi secara signifikan tidak cukup (bukan yang maksimal yang dapat dilakukan).

Dengan analogi, ini adalah perbedaan antara bermain sesuai aturan, dan bermain dengan baik, sehingga seseorang dapat memenangkan permainan. Mengadopsi pendekatan etis terhadap AI memberikan apa yang kami definisikan di sini sebagai "keuntungan ganda". Di satu sisi, etika memungkinkan organisasi untuk memanfaatkan nilai sosial yang dimungkinkan oleh AI. Ini adalah keuntungan karena mampu mengidentifikasi dan memanfaatkan peluang baru yang dapat diterima atau disukai secara sosial.

Di sisi lain, etika memungkinkan organisasi untuk mengantisipasi dan menghindari atau setidaknya meminimalkan kesalahan yang merugikan. Inilah keuntungan dari pencegahan dan mitigasi tindakan yang ternyata tidak dapat diterima secara sosial dan karenanya ditolak, meskipun secara hukum tidak perlu dipertanyakan lagi. Hal ini juga menurunkan biaya peluang dari pilihan yang tidak dibuat atau opsi yang tidak diambil karena takut akan kesalahan. Keunggulan ganda etika hanya dapat berfungsi dalam lingkungan kepercayaan publik dan tanggung jawab yang jelas secara lebih luas.

Penerimaan dan adopsi publik terhadap teknologi AI hanya akan terjadi jika manfaatnya dipandang bermakna dan risikonya potensial, namun dapat dicegah, diminimalkan, atau setidaknya sesuatu yang dapat dilindungi, melalui manajemen risiko

(misalnya asuransi) atau pemulihan. Sikap-sikap ini pada gilirannya akan bergantung pada keterlibatan publik dalam pengembangan teknologi AI, keterbukaan tentang cara kerjanya, dan mekanisme regulasi dan pemulihan yang mudah dipahami dan dapat diakses secara luas. Dengan demikian, pendekatan etis terhadap AI juga dapat dipandang sebagai sistem peringatan dini terhadap risiko yang dapat membahayakan seluruh organisasi. Nilai yang jelas bagi setiap organisasi dari keunggulan ganda pendekatan etis terhadap AI cukup membenarkan biaya keterlibatan, keterbukaan, dan kontestabilitas yang dibutuhkan oleh pendekatan semacam itu.

2.4 KERANGKA PRINSIP TERPADU UNTUK AI DALAM MASYARAKAT

AI4People bukanlah inisiatif pertama yang mempertimbangkan implikasi etis AI. Banyak organisasi telah menghasilkan pernyataan nilai atau prinsip yang harus memandu pengembangan dan penerapan AI di masyarakat. Daripada melakukan latihan serupa yang berpotensi redundan di sini, kami berusaha untuk memajukan dialog, secara konstruktif, dari prinsip ke kebijakan yang diusulkan, praktik terbaik, dan rekomendasi konkret untuk strategi baru.

Rekomendasi semacam itu tidak ditawarkan dalam ruang hampa. Namun, alih-alih menghasilkan serangkaian prinsip lain untuk berfungsi sebagai landasan etis bagi rekomendasi kami, kami menawarkan sintesis dari serangkaian prinsip yang ada yang dihasilkan oleh berbagai organisasi dan inisiatif multi-pemangku kepentingan yang bereputasi baik. Penjelasan yang lebih lengkap tentang ruang lingkup, pemilihan, dan metode penilaian serangkaian prinsip ini tersedia di Cowls dan Floridi (Akan Datang). Di sini, kami berfokus pada persamaan dan perbedaan penting yang dapat diamati di seluruh rangkaian prinsip ini, mengingat 20 rekomendasi yang ditawarkan di sisa makalah ini. Dokumen-dokumen yang kami nilai adalah:

1. Prinsip-prinsip AI Asilomar, yang dikembangkan di bawah naungan Future of Life Institute, bekerja sama dengan para peserta konferensi tingkat tinggi Asilomar pada Januari 2017.
2. Deklarasi Montreal untuk AI yang Bertanggung Jawab, yang dikembangkan di bawah naungan Universitas Montreal, setelah Forum tentang Pengembangan AI yang Bertanggung Jawab Secara Sosial pada November 2017.
3. Prinsip-prinsip Umum yang ditawarkan dalam versi kedua *Desain yang Selaras secara Etis: Sebuah Visi untuk Memprioritaskan Kesejahteraan Manusia dengan Sistem Otonom dan Cerdas*. Risalah global yang bersumber dari banyak pihak ini menerima kontribusi dari 250 pemimpin pemikiran global untuk mengembangkan prinsip dan rekomendasi bagi pengembangan dan perancangan sistem otonom dan cerdas yang etis, dan diterbitkan pada Desember 2017.
4. Prinsip-prinsip Etika yang ditawarkan dalam *Pernyataan tentang Kecerdasan Buatan, Robotika, dan Sistem 'Otonom'*, yang diterbitkan oleh Kelompok Etika Eropa dalam Sains dan Teknologi Baru Komisi Eropa, pada Maret 2018.

5. "lima prinsip menyeluruh untuk kode AI" yang ditawarkan dalam paragraf 417 laporan Komite Kecerdasan Buatan Dewan Perwakilan Rakyat Inggris, *AI di Inggris: siap, bersedia, dan mampu?*, yang diterbitkan pada April 2018.
6. Prinsip-Prinsip Kemitraan untuk AI, sebuah organisasi multipihak yang terdiri dari akademisi, peneliti, organisasi masyarakat sipil, perusahaan yang membangun dan memanfaatkan teknologi AI, dan kelompok-kelompok lain.

Secara keseluruhan, prinsip-prinsip tersebut menghasilkan 47 prinsip. Secara keseluruhan, kami menemukan tingkat koherensi dan tumpang tindih yang mengesankan dan meyakinkan antara keenam rangkaian prinsip tersebut.

Hal ini paling jelas ditunjukkan dengan membandingkan rangkaian prinsip tersebut dengan rangkaian empat prinsip inti yang umum digunakan dalam bioetika: kebaikan, non-maleficence, otonomi, dan keadilan. Perbandingan ini seharusnya tidak mengejutkan. Dari semua bidang etika terapan, bioetika adalah yang paling mirip dengan etika digital dalam menangani bentuk-bentuk baru agen, pasien, dan lingkungan secara ekologis (Floridi 2013). Keempat prinsip bioetika ini beradaptasi dengan sangat baik terhadap tantangan etika baru yang ditimbulkan oleh kecerdasan buatan. Namun, prinsip-prinsip tersebut tidak lengkap. Atas dasar analisis perbandingan berikut, kami berpendapat bahwa satu prinsip baru lagi dibutuhkan sebagai tambahan: *keterjelasan*, dipahami sebagai sesuatu yang menggabungkan antara keterpahaman dan akuntabilitas.

Kebaikan: Meningkatkan Kesejahteraan, Menjaga Martabat, dan Melestarikan Planet

Dari empat prinsip inti bioetika, kebaikan mungkin yang paling mudah diamati dari keenam prinsip yang kami sintesiskan di sini. Prinsip menciptakan teknologi AI yang bermanfaat bagi umat manusia diungkapkan dengan cara yang berbeda, tetapi biasanya berada di urutan teratas setiap daftar prinsip. Prinsip Montreal dan IEEE sama-sama menggunakan istilah "kesejahteraan": untuk Montreal, "pengembangan AI pada akhirnya harus meningkatkan kesejahteraan semua makhluk hidup"; sementara IEEE menyatakan perlunya "memprioritaskan kesejahteraan manusia sebagai hasil dalam semua desain sistem".

AIUK dan Asilomar sama-sama menggolongkan prinsip ini sebagai "kebaikan bersama": AI harus "dikembangkan untuk kebaikan bersama dan manfaat umat manusia", menurut AIUK. Kemitraan tersebut menjelaskan niat untuk "memastikan bahwa teknologi AI memberi manfaat dan memberdayakan sebanyak mungkin orang"; sementara EGE menekankan prinsip "martabat manusia" dan "keberlanjutan".

Prinsip "keberlanjutan"-nya mungkin mewakili interpretasi terluas dari semua konsep kebaikan, dengan argumen bahwa "teknologi AI harus sejalan dengan ... memastikan prasyarat dasar bagi kehidupan di planet kita, kesejahteraan umat manusia yang berkelanjutan, dan pelestarian lingkungan yang baik untuk generasi mendatang". Secara keseluruhan, keunggulan prinsip-prinsip kebaikan ini dengan tegas menggarisbawahi pentingnya mempromosikan kesejahteraan manusia dan planet ini.

Non-maleficence: Privasi, Keamanan, dan "Kehati-hatian Kemampuan"

Meskipun "hanya berbuat baik" (kebaikan) dan "tidak merugikan" (non-maleficence) tampak setara secara logis, baik dalam konteks bioetika maupun etika AI, keduanya mewakili

prinsip-prinsip yang berbeda, yang masing-masing memerlukan penjelasan. Meskipun mendorong kesejahteraan, pembagian manfaat, dan kemajuan kebaikan publik, keenam rangkaian prinsip tersebut juga memperingatkan terhadap berbagai potensi konsekuensi negatif dari penggunaan teknologi AI yang berlebihan atau disalahgunakan. Yang menjadi perhatian khusus adalah pencegahan pelanggaran privasi pribadi, yang tercantum sebagai prinsip dalam lima dari enam set, dan sebagai bagian dari prinsip-prinsip "hak asasi manusia" dalam dokumen IEEE. Dalam setiap kasus, privasi dicirikan sebagai sesuatu yang terkait erat dengan akses individu terhadap, dan kendali atas, bagaimana data pribadi digunakan.

Namun, pelanggaran privasi bukanlah satu-satunya bahaya yang harus dihindari dalam adopsi AI. Beberapa dokumen juga menekankan pentingnya menghindari penyalahgunaan teknologi AI dengan cara lain. Prinsip-prinsip Asilomar cukup spesifik dalam hal ini, mengutip ancaman perlombaan senjata AI dan peningkatan diri AI yang rekursif, serta perlunya "kehati-hatian" seputar "batas atas kemampuan AI di masa mendatang". Kemitraan juga menegaskan pentingnya AI beroperasi "dalam batasan yang aman". Sementara itu, dokumen IEEE mengutip perlunya "menghindari penyalahgunaan", sementara Deklarasi Montreal berpendapat bahwa mereka yang mengembangkan AI "harus memikul tanggung jawab mereka dengan melawan risiko yang timbul dari inovasi teknologi mereka", yang digaungkan oleh kebutuhan serupa EGE akan tanggung jawab.

Dari berbagai peringatan ini, tidak sepenuhnya jelas apakah orang-orang yang mengembangkan AI, atau teknologinya sendiri, yang harus didorong untuk tidak melakukan hal yang merugikan dengan kata lain, apakah Frankenstein atau monsternya yang harus kita waspadai terhadap kejahatannya. Yang juga membingungkan adalah pertanyaan tentang niat: mempromosikan non-kejahatan dapat dilihat sebagai upaya untuk mencegah timbulnya kerugian, baik yang tidak disengaja (apa yang kita sebut "penggunaan berlebihan") maupun yang disengaja (apa yang kita sebut "penyalahgunaan").

Dalam hal prinsip non-kejahatan, ini tidak perlu menjadi pertanyaan salah satu/atau: intinya hanyalah untuk mencegah timbulnya kerugian, baik dari niat manusia maupun perilaku mesin yang tidak terduga (termasuk dorongan yang tidak disengaja terhadap perilaku manusia dengan cara yang tidak diinginkan). Namun pertanyaan mendasar mengenai agensi, niat, dan kendali menjadi lebih rumit ketika kita mempertimbangkan prinsip berikutnya.

Otonomi: Kekuatan untuk Memutuskan (Apakah Akan Memutuskan)

Prinsip klasik bioetika lainnya adalah prinsip otonomi: gagasan bahwa individu memiliki hak untuk membuat keputusan sendiri tentang perawatan yang mereka terima atau tidak. Dalam konteks medis, prinsip otonomi ini paling sering terganggu ketika pasien tidak memiliki kapasitas mental untuk membuat keputusan demi kepentingan terbaik mereka sendiri; dengan demikian, otonomi dilepaskan tanpa disadari. Dengan AI, situasinya menjadi jauh lebih kompleks: ketika kita mengadopsi AI dan agensi cerdasnya, kita *dengan sukarela* menyerahkan sebagian kekuatan pengambilan keputusan kita kepada mesin. Dengan demikian, menegaskan prinsip otonomi dalam konteks AI berarti mencapai keseimbangan antara kekuatan pengambilan keputusan yang kita miliki untuk diri kita sendiri dan kekuatan yang kita delegasikan kepada agen buatan.

Prinsip otonomi dinyatakan secara eksplisit dalam empat dari enam dokumen. Deklarasi Montreal mengartikulasikan perlunya keseimbangan antara pengambilan keputusan yang dipimpin manusia dan mesin, yang menyatakan bahwa "*pengembangan AI harus mendorong otonomi semua manusia dan mengendalikan ... otonomi sistem komputer*". EGE berpendapat bahwa sistem otonom "tidak boleh mengganggu kebebasan manusia untuk menetapkan standar dan norma mereka sendiri serta mampu hidup sesuai dengannya", sementara AIUK mengadopsi sikap yang lebih sempit bahwa "kekuatan otonom untuk melukai, menghancurkan, atau menipu manusia tidak boleh diberikan kepada AI". Dokumen Asilomar juga mendukung prinsip otonomi, sejauh "manusia harus memilih bagaimana dan apakah akan mendelegasikan keputusan kepada sistem AI, untuk mencapai tujuan yang dipilih manusia".

Dokumen-dokumen ini mengungkapkan sentimen serupa dengan cara yang sedikit berbeda, menggemakan perbedaan yang telah diuraikan di atas antara kebaikan dan non-kejahatan: tidak hanya otonomi manusia yang harus dipromosikan, tetapi juga otonomi mesin harus dibatasi dan dibuat secara intrinsik dapat dibalik, jika otonomi manusia perlu dibangun kembali (pertimbangkan kasus seorang pilot yang dapat mematikan pilot otomatis dan mendapatkan kembali kendali penuh atas pesawat).

Secara keseluruhan, poin utamanya adalah melindungi nilai intrinsik pilihan manusia setidaknya untuk keputusan-keputusan penting dan, sebagai konsekuensinya, membatasi risiko pendelegasian terlalu banyak kepada mesin. Oleh karena itu, yang tampaknya paling penting di sini adalah apa yang bisa kita sebut "meta-otonomi", atau model "putuskan-untuk-mendelegasikan": manusia harus selalu memiliki kekuatan untuk *memutuskan keputusan mana yang akan diambil*, menggunakan kebebasan untuk memilih jika perlu, dan menyerahkannya dalam kasus-kasus di mana alasan-alasan yang lebih penting, seperti efisiensi, mungkin lebih besar daripada hilangnya kendali atas pengambilan keputusan.

Sebagaimana diantisipasi, setiap pendelegasian pada prinsipnya harus tetap dapat diabaikan (memutuskan untuk memutuskan lagi). Keputusan untuk membuat atau mendelegasikan keputusan tidak terjadi dalam ruang hampa. Kapasitas untuk memutuskan (untuk memutuskan, dan memutuskan lagi) ini juga tidak didistribusikan secara merata di seluruh masyarakat. Konsekuensi dari potensi disparitas otonomi ini dibahas dalam prinsip terakhir dari empat prinsip yang terinspirasi oleh bioetika.

Keadilan: Mempromosikan Kemakmuran dan Memelihara Solidaritas

Prinsip terakhir dari empat prinsip bioetika klasik adalah keadilan, yang biasanya dikaitkan dengan distribusi sumber daya, seperti pilihan pengobatan baru dan eksperimental atau sekadar ketersediaan umum layanan kesehatan konvensional. Sekali lagi, prinsip bioetika ini menemukan gaung yang jelas di seluruh prinsip AI yang kami analisis. Pentingnya "keadilan" secara eksplisit dikutip dalam Deklarasi Montreal, yang menyatakan bahwa "pengembangan AI harus mempromosikan keadilan dan berupaya menghilangkan semua jenis diskriminasi", sementara Prinsip Asilomar mencakup perlunya "manfaat bersama" dan "kemakmuran bersama" dari AI. Berdasarkan prinsipnya yang bernama "Keadilan, kesetaraan, dan solidaritas", EGE berpendapat bahwa AI harus "berkontribusi pada keadilan global dan akses yang setara terhadap manfaat" teknologi AI.

Prinsip ini juga memperingatkan risiko bias dalam kumpulan data yang digunakan untuk melatih sistem AI, dan yang unik di antara dokumen-dokumen lainnya mengemukakan perlunya mempertahankan diri dari ancaman terhadap "solidaritas", termasuk "sistem bantuan timbal balik seperti dalam asuransi sosial dan layanan kesehatan". Penekanan pada perlindungan sistem dukungan sosial mungkin mencerminkan geopolitik, mengingat EGE merupakan badan Eropa. Laporan AIUK berpendapat bahwa warga negara harus dapat "berkembang secara mental, emosional, dan ekonomi bersama kecerdasan buatan". Sementara itu, Kemitraan mengadopsi kerangka yang lebih hati-hati, berjanji untuk "menghormati kepentingan semua pihak yang mungkin terdampak oleh kemajuan AI".

Sebagaimana prinsip-prinsip lain yang telah dibahas, interpretasi tentang arti keadilan sebagai prinsip etika dalam konteks AI ini secara umum serupa, namun mengandung perbedaan yang halus. Dalam berbagai dokumen, keadilan berkaitan dengan berbagai hal, yaitu:

- (a) penggunaan AI untuk mengoreksi kesalahan masa lalu, seperti menghilangkan diskriminasi yang tidak adil;
- (b) memastikan bahwa penggunaan AI menciptakan manfaat yang dibagi (atau setidaknya dapat dibagi); dan
- (c) mencegah terciptanya kerugian baru, seperti meruntuhkan struktur sosial yang ada.

Yang juga perlu diperhatikan adalah perbedaan cara posisi AI, *vis-à-vis* manusia, dicirikan dalam kaitannya dengan keadilan. Di Asilomar dan EGE, teknologi AI itu sendirilah yang "harus memberi manfaat dan memberdayakan sebanyak mungkin orang" dan "berkontribusi pada keadilan global", sedangkan di Montreal, "*pengembangan AI*"-lah yang "harus mendorong keadilan" (cetak miring ditambahkan).

Sementara itu, di AIUK, manusia seharusnya berkembang hanya "bersama" AI. Tujuan kami di sini bukan untuk mempermasalahkan hal-hal semantik. Berbagai cara hubungan antara manusia dan AI dijelaskan dalam dokumen-dokumen ini mengisyaratkan kebingungan yang lebih luas tentang AI sebagai reservoir "agensi cerdas" buatan manusia. Sederhananya, dan untuk melanjutkan analogi bioetika kita, apakah kita (manusia) adalah pasien, yang menerima "pengobatan" AI, dan dokter yang meresepkannya? Atau keduanya? Tampaknya kita harus menjawab pertanyaan ini sebelum mencoba menjawab pertanyaan berikutnya, apakah pengobatan tersebut akan berhasil. Inilah justifikasi inti bagi identifikasi kami dalam dokumen-dokumen ini atas sebuah prinsip baru, yang tidak bersumber dari bioetika.

Keterangan: Mengaktifkan Prinsip-Prinsip Lain Melalui Kecerdasan dan Akuntabilitas

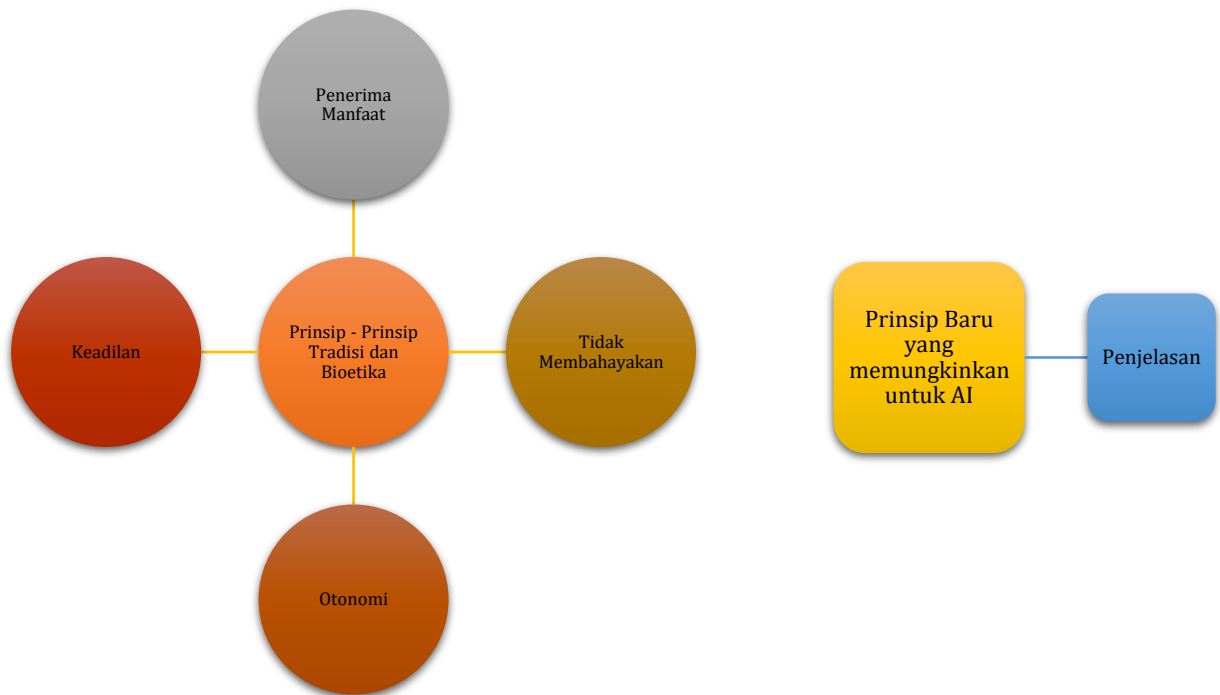
Jawaban singkat untuk pertanyaan apakah "kita" adalah pasien atau dokter adalah bahwa sebenarnya kita bisa menjadi salah satunya tergantung pada keadaan dan siapa "kita" dalam kehidupan sehari-hari. Situasinya pada dasarnya tidak setara: sebagian kecil umat manusia saat ini terlibat dalam perancangan dan pengembangan serangkaian teknologi yang telah mengubah kehidupan sehari-hari hampir semua orang. Realitas pahit ini tidak luput dari perhatian para penulis yang dokumennya kami analisis.

Secara keseluruhan, terdapat rujukan pada kebutuhan untuk *memahami* dan *mempertanggungjawabkan* proses pengambilan keputusan AI. Prinsip ini diungkapkan

menggunakan istilah yang berbeda: "transparansi" dalam Asilomar; "akuntabilitas" dalam EGE; baik "transparansi" maupun "akuntabilitas" dalam IEEE; "kecerdasan" dalam AIUK; dan "dapat dipahami dan ditafsirkan" dalam Kemitraan. Meskipun dijelaskan dengan cara yang berbeda, masing-masing prinsip ini menangkap sesuatu yang tampaknya baru tentang AI: bahwa cara kerjanya seringkali tidak terlihat atau tidak dapat dipahami oleh semua orang kecuali (paling banter) pengamat yang paling ahli.

Penambahan prinsip ini, yang kami sintesiskan sebagai "explicability" baik dalam arti epistemologis "intelligibility" (sebagai jawaban atas pertanyaan "bagaimana cara kerjanya?") dan dalam arti etis "accountability" (sebagai jawaban atas pertanyaan: "siapa yang bertanggung jawab atas cara kerjanya?"), oleh karena itu merupakan bagian penting yang hilang dari teka-teki ketika kita berusaha menerapkan kerangka bioetika pada etika AI. Ini melengkapi empat prinsip lainnya: agar AI bersifat dermawan dan tidak jahat, kita harus mampu memahami kebaikan atau kerugian yang sebenarnya ditimbulkannya terhadap masyarakat, dan dengan cara apa; agar AI dapat mempromosikan dan tidak membatasi otonomi manusia, "keputusan kita tentang siapa yang harus memutuskan" harus diinformasikan oleh pengetahuan tentang bagaimana AI akan bertindak, bukan kita; Agar AI adil, kita harus memastikan bahwa teknologinya atau, lebih tepatnya, orang-orang dan organisasi yang mengembangkannya dan menerapkannya bertanggung jawab jika terjadi hasil negatif, yang pada gilirannya membutuhkan pemahaman tentang mengapa hasil ini muncul.

Lebih luas lagi, kita harus menegosiasikan syarat-syarat hubungan antara kita dan teknologi transformatif ini, dengan alasan yang mudah dipahami oleh orang awam. Secara keseluruhan, kami berpendapat bahwa kelima prinsip ini menangkap makna dari masing-masing dari 47 prinsip yang terkandung dalam enam dokumen penting yang dikaji oleh para ahli, membentuk kerangka etika yang kami tawarkan rekomendasinya di bawah ini. Kerangka prinsip ini ditunjukkan pada Gambar 2.2.



Gambar 2.2 Kerangka etika untuk AI, dibentuk dari empat prinsip tradisional dan satu prinsip baru

2.5 REKOMENDASI UNTUK MASYARAKAT AI YANG BAIK

Bagian ini memperkenalkan Rekomendasi untuk Masyarakat AI yang Baik. Bagian ini terdiri dari dua bagian: Pembukaan, dan 20 Poin Aksi. Ada empat jenis Poin Aksi: *menilai*, *mengembangkan*, *memberi insentif*, dan *mendukung*. Beberapa rekomendasi dapat dilaksanakan secara langsung, oleh pembuat kebijakan nasional atau Eropa, bekerja sama dengan para pemangku kepentingan jika diperlukan. Untuk rekomendasi lainnya, pembuat kebijakan dapat memainkan peran pendukung bagi upaya yang dilakukan atau dipimpin oleh pihak ketiga.

Pembukaan

Kami percaya bahwa, untuk menciptakan Masyarakat AI yang Baik, prinsip-prinsip etika yang diidentifikasi di bagian sebelumnya harus tertanam dalam praktik AI yang baku. Secara khusus, AI harus dirancang dan dikembangkan dengan cara yang mengurangi ketimpangan dan memajukan pemberdayaan sosial, dengan menghormati otonomi manusia, dan meningkatkan manfaat yang dinikmati semua orang secara merata.

Sangat penting bahwa AI dapat dijelaskan, karena penjelasan merupakan alat penting untuk membangun kepercayaan publik terhadap, dan pemahaman tentang, teknologi ini. Kami juga percaya bahwa menciptakan Masyarakat AI yang Baik membutuhkan pendekatan multipihak, yang merupakan cara paling efektif untuk memastikan bahwa AI akan melayani kebutuhan masyarakat, dengan memungkinkan pengembang, pengguna, dan pembuat aturan untuk terlibat dan berkolaborasi sejak awal.

Kerangka budaya yang berbeda menginformasikan sikap terhadap teknologi baru. Dokumen ini merupakan pendekatan Eropa, yang dimaksudkan untuk melengkapi pendekatan

lain. Kami berkomitmen pada pengembangan teknologi AI dengan cara yang *menjamin kepercayaan masyarakat, melayani kepentingan publik, dan memperkuat tanggung jawab sosial bersama*. Akhirnya, serangkaian rekomendasi ini harus dilihat sebagai "dokumen hidup". Poin-poin Aksi dirancang agar dinamis, tidak hanya membutuhkan kebijakan tunggal atau investasi sekali pakai, melainkan upaya berkelanjutan agar dampaknya berkelanjutan.

Poin Aksi; Penilaian

1. *Menilai* kapasitas lembaga yang ada, seperti pengadilan sipil nasional, untuk memperbaiki kesalahan atau kerugian yang ditimbulkan oleh sistem AI. Penilaian ini harus mengevaluasi keberadaan fondasi pertanggungjawaban yang berkelanjutan dan disepakati mayoritas sejak tahap desain dan seterusnya untuk mengurangi kelalaian dan konflik.
2. *Menilai* tugas dan fungsi pengambilan keputusan mana yang tidak boleh didelegasikan kepada sistem AI, melalui penggunaan mekanisme partisipatif untuk memastikan keselarasan dengan nilai-nilai sosial dan pemahaman opini publik. Penilaian ini harus mempertimbangkan undang-undang yang ada dan didukung oleh dialog berkelanjutan antara semua pemangku kepentingan (termasuk pemerintah, industri, dan masyarakat sipil) untuk membahas bagaimana AI akan memengaruhi opini masyarakat.
3. *Menilai* apakah peraturan saat ini cukup berlandaskan etika untuk menyediakan kerangka legislatif yang dapat mengimbangi perkembangan teknologi. Ini dapat mencakup kerangka prinsip-prinsip utama yang akan diterapkan pada masalah yang mendesak dan/atau tidak terduga.
4. *Mengembangkan* kerangka kerja untuk meningkatkan keterjelasan sistem AI yang membuat keputusan penting secara sosial. Inti dari kerangka kerja ini adalah kemampuan individu untuk memperoleh penjelasan yang faktual, langsung, dan jelas tentang proses pengambilan keputusan, terutama jika terjadi konsekuensi yang tidak diinginkan. Hal ini kemungkinan memerlukan pengembangan kerangka kerja yang spesifik untuk berbagai industri, dan asosiasi profesional harus dilibatkan dalam proses ini, bersama para ahli di bidang sains, bisnis, hukum, dan etika.
5. *Mengembangkan* prosedur hukum yang tepat dan meningkatkan infrastruktur TI sistem peradilan untuk memungkinkan pengawasan keputusan algoritmik di pengadilan. Hal ini kemungkinan mencakup pembuatan kerangka kerja untuk keterjelasan AI sebagaimana ditunjukkan dalam Rekomendasi 4, yang khusus untuk sistem hukum. Contoh prosedur yang tepat dapat mencakup pengungkapan informasi komersial sensitif yang berlaku dalam litigasi kekayaan intelektual, dan jika pengungkapan menimbulkan risiko yang tidak dapat diterima, misalnya terhadap keamanan nasional konfigurasi sistem AI untuk mengadopsi solusi teknis secara default, seperti bukti tanpa pengetahuan (*zero-knowledge proofs*) guna mengevaluasi keandalannya.
6. *Mengembangkan* mekanisme audit untuk sistem AI guna mengidentifikasi konsekuensi yang tidak diinginkan, seperti bias yang tidak adil, dan (misalnya, bekerja sama dengan sektor asuransi) mekanisme solidaritas untuk menangani risiko berat di sektor-sektor

yang intensif AI. Risiko-risiko tersebut dapat dimitigasi melalui mekanisme multipihak di hulu. Pengalaman pra-digital menunjukkan bahwa, dalam beberapa kasus, dibutuhkan beberapa dekade sebelum masyarakat dapat mengejar ketertinggalan teknologi melalui penyeimbangan kembali hak dan perlindungan yang memadai untuk memulihkan kepercayaan. Semakin dini pengguna dan pemerintah terlibat sebagaimana dimungkinkan oleh TIK semakin pendek pula jeda ini.

7. *Mengembangkan* proses atau mekanisme pemulihan untuk memperbaiki atau memberikan kompensasi atas kesalahan atau keluhan yang disebabkan oleh AI. Untuk menumbuhkan kepercayaan publik terhadap AI, masyarakat membutuhkan mekanisme pemulihan yang dapat diakses secara luas dan andal atas kerugian yang ditimbulkan, biaya yang dikeluarkan, atau keluhan lain yang disebabkan oleh teknologi. Mekanisme semacam itu tentu akan melibatkan alokasi akuntabilitas yang jelas dan komprehensif kepada manusia dan/atau organisasi. Pelajaran dapat dipetik dari industri kedirgantaraan, misalnya, yang memiliki sistem yang terbukti mampu menangani konsekuensi yang tidak diinginkan secara menyeluruh dan serius. Pengembangan proses ini harus mengikuti penilaian kapasitas yang ada sebagaimana diuraikan dalam Rekomendasi 1. Jika terdapat kekurangan kapasitas, solusi kelembagaan tambahan harus dikembangkan di tingkat nasional dan/atau Uni Eropa, untuk memungkinkan masyarakat mencari ganti rugi. Solusi tersebut dapat mencakup:
 - ➡ “ombudsman AI” untuk memastikan audit atas dugaan penggunaan AI yang tidak adil atau tidak setara;
 - ➡ proses terpadu untuk mendaftarkan pengaduan yang serupa dengan pengajuan permohonan Kebebasan Informasi; dan
 - ➡ pengembangan mekanisme asuransi liabilitas, yang akan diwajibkan sebagai pelengkap wajib untuk kelas penawaran AI tertentu di Uni Eropa dan pasar lainnya. Hal ini akan memastikan bahwa keandalan relatif artefak bertenaga AI, terutama dalam robotika, tercermin dalam penetapan harga asuransi dan oleh karena itu dalam harga pasar produk pesaing.

Solusi apa pun yang dikembangkan, kemungkinan besar akan bergantung pada kerangka kerja untuk kejelasan yang diusulkan dalam Rekomendasi 4.

8. *Kembangkan* metrik yang disepakati untuk kepercayaan produk dan layanan AI, yang akan dilakukan baik oleh organisasi baru maupun oleh organisasi lama yang sesuai. Metrik ini akan berfungsi sebagai dasar bagi sistem yang memungkinkan perbandingan berbasis pengguna untuk semua penawaran AI yang dipasarkan. Dengan cara ini, indeks untuk AI yang tepercaya dapat dikembangkan dan ditandai, selain harga produk. "Indeks perbandingan kepercayaan" untuk AI ini akan meningkatkan pemahaman publik dan mendorong daya saing dalam pengembangan AI yang lebih aman dan bermanfaat secara sosial (misalnya, "IwantgreatAI.org"). Dalam jangka panjang, sistem semacam itu dapat menjadi dasar bagi sistem sertifikasi yang lebih luas untuk produk dan layanan yang layak, yang dikelola oleh organisasi yang disebutkan di sini, dan/atau oleh badan pengawas yang diusulkan dalam Rekomendasi 9. Organisasi tersebut juga dapat mendukung pengembangan kode etik (lihat

Rekomendasi 18). Lebih lanjut, pihak yang memiliki atau mengoperasikan input untuk sistem AI dan mendapatkan keuntungan darinya dapat ditugaskan untuk mendanai dan/atau membantu mengembangkan program literasi AI bagi konsumen, demi kepentingan terbaik mereka sendiri.

9. *Mengembangkan* badan pengawas Uni Eropa baru yang bertanggung jawab atas perlindungan kesejahteraan publik melalui evaluasi dan supervisi ilmiah terhadap produk, perangkat lunak, sistem, atau layanan AI. Hal ini mungkin serupa, misalnya, dengan Badan Obat-obatan Eropa. Sehubungan dengan itu, sistem pemantauan "pasca-rilis" untuk AI yang serupa dengan, misalnya, sistem yang tersedia untuk obat-obatan perlu dikembangkan, dengan kewajiban pelaporan bagi beberapa pemangku kepentingan dan mekanisme pelaporan yang mudah bagi pengguna lainnya.
10. Mengembangkan observatorium Eropa untuk AI. Misi observatorium ini adalah untuk memantau perkembangan, menyediakan forum untuk mendorong debat dan konsensus, menyediakan repositori untuk literatur dan perangkat lunak AI (termasuk konsep dan tautan ke literatur yang tersedia), dan mengeluarkan rekomendasi serta panduan tindakan langkah demi langkah.
11. *Mengembangkan* instrumen hukum dan templat kontrak untuk meletakkan fondasi bagi kolaborasi manusia-mesin yang lancar dan bermanfaat di lingkungan kerja. Membentuk narasi tentang 'Masa Depan Pekerjaan' berperan penting untuk memenangkan "hati dan pikiran". Sejalan dengan 'Eropa yang melindungi', gagasan "inovasi inklusif" dan untuk memperlancar transisi ke jenis pekerjaan baru, Dana Penyesuaian AI Eropa dapat dibentuk sejalan dengan Dana Penyesuaian Globalisasi Eropa.

Insentif

12. Berikan *insentif* finansial, di tingkat Uni Eropa, untuk pengembangan dan penggunaan teknologi AI di Uni Eropa yang secara sosial lebih disukai (bukan sekadar dapat diterima) dan ramah lingkungan (bukan sekadar berkelanjutan tetapi juga ramah lingkungan). Ini akan mencakup pengembangan metodologi yang dapat membantu menilai apakah proyek AI secara sosial lebih disukai dan ramah lingkungan. Dalam hal ini, mengadopsi 'pendekatan tantangan' (lihat tantangan DARPA) dapat mendorong kreativitas dan mendorong persaingan dalam pengembangan solusi AI spesifik yang etis dan demi kepentingan bersama.
13. Berikan *insentif* finansial untuk upaya penelitian Eropa yang berkelanjutan, meningkat, dan koheren, yang disesuaikan dengan fitur spesifik AI sebagai bidang penelitian ilmiah. Ini harus melibatkan misi yang jelas untuk memajukan AI demi kebaikan sosial, untuk berfungsi sebagai penyeimbang yang unik terhadap tren AI dengan lebih sedikit fokus pada peluang sosial.
14. Berikan *insentif* finansial untuk kerja sama dan debat lintas disiplin dan lintas sektor mengenai persimpangan antara teknologi, isu sosial, studi hukum, dan etika. Debat tentang tantangan teknologi mungkin tertinggal dari kemajuan teknis yang sebenarnya, tetapi jika diinformasikan secara strategis oleh kelompok multipemangku kepentingan

yang beragam, debat tersebut dapat mengarahkan dan mendukung inovasi teknologi ke arah yang tepat. Etika seharusnya membantu meraih peluang dan mengatasi tantangan, bukan hanya mendeskripsikannya. Dalam hal ini, penting bagi keberagaman untuk menanamkan desain dan pengembangan AI, dalam hal gender, kelas, etnis, disiplin, dan dimensi terkait lainnya, guna meningkatkan inklusivitas, toleransi, serta kekayaan gagasan dan perspektif.

15. Berikan *insentif* finansial untuk penyertaan pertimbangan etika, hukum, dan sosial dalam proyek penelitian AI. Secara paralel, berikan insentif untuk tinjauan rutin terhadap undang-undang guna menguji sejauh mana undang-undang tersebut mendorong inovasi yang positif secara sosial. Secara keseluruhan, kedua langkah ini akan membantu memastikan bahwa teknologi AI berlandaskan etika dan kebijakan berorientasi pada inovasi.
16. Berikan *insentif* finansial untuk pengembangan dan penggunaan zona khusus yang telah dideregulasi secara hukum di Uni Eropa untuk pengujian dan pengembangan empiris sistem AI. Zona-zona ini dapat berbentuk "laboratorium hidup" (atau *Tokku*), yang dibangun berdasarkan pengalaman "jalur uji coba" (atau *Teststrecken*) yang sudah ada. Selain menyelaraskan inovasi lebih dekat dengan tingkat risiko yang diinginkan masyarakat, eksperimen sandbox seperti ini berkontribusi pada pendidikan langsung dan promosi akuntabilitas serta akseptabilitas sejak dini. "Perlindungan sejak awal" merupakan hal yang intrinsik dalam kerangka kerja semacam ini.
17. Berikan *insentif* finansial untuk penelitian tentang persepsi dan pemahaman publik tentang AI dan aplikasinya, serta implementasi mekanisme konsultasi publik terstruktur untuk merancang kebijakan dan aturan terkait AI. Hal ini dapat mencakup penggalian opini publik secara langsung melalui metode penelitian tradisional, seperti jajak pendapat dan kelompok fokus, serta pendekatan yang lebih eksperimental, seperti memberikan contoh simulasi dilema etika yang ditimbulkan oleh sistem AI, atau eksperimen di laboratorium ilmu sosial. Agenda penelitian ini seharusnya tidak hanya berfungsi untuk mengukur opini publik, tetapi juga mengarah pada penciptaan bersama kebijakan, standar, praktik terbaik, dan aturan sebagai hasilnya.

Dukungan

18. *Mendukung* pengembangan kode etik pengaturan mandiri untuk profesi terkait data dan AI, dengan tugas-tugas etika spesifik. Hal ini sejalan dengan profesi lain yang peka sosial, seperti dokter atau pengacara, yaitu dengan sertifikasi 'AI etis' melalui label kepercayaan untuk memastikan bahwa orang-orang memahami manfaat AI etis dan oleh karena itu akan menuntutnya dari penyedia. Teknik manipulasi perhatian yang ada saat ini dapat dibatasi melalui instrumen pengaturan mandiri ini.
19. *Mendukung* kapasitas dewan direksi perusahaan untuk bertanggung jawab atas implikasi etis dari teknologi AI perusahaan. Misalnya, hal ini dapat mencakup peningkatan pelatihan untuk dewan yang ada dan potensi pengembangan komite etik dengan wewenang audit internal. Hal ini dapat dikembangkan dalam struktur sistem dewan satu tingkat dan dua tingkat yang ada, dan/atau bersamaan dengan

pengembangan bentuk wajib "dewan peninjau etika perusahaan" yang akan diadopsi oleh organisasi yang mengembangkan atau menggunakan sistem AI, untuk mengevaluasi proyek awal dan penerapannya dengan memperhatikan prinsip-prinsip dasar.

20. *Mendukung* penyusunan kurikulum pendidikan dan kegiatan penyadaran publik seputar dampak sosial, hukum, dan etika Kecerdasan Buatan. Ini dapat mencakup:

- ❖ kurikulum untuk sekolah, yang mendukung inklusi ilmu komputer di antara disiplin ilmu dasar yang diajarkan;
- ❖ inisiatif dan program kualifikasi di perusahaan yang bergerak di bidang teknologi AI, untuk mendidik karyawan tentang dampak sosial, hukum, dan etika dari bekerja berdampingan dengan AI;
- ❖ rekomendasi tingkat Eropa untuk memasukkan etika dan hak asasi manusia dalam program gelar ilmuwan data dan AI serta kurikulum ilmiah dan teknik lainnya yang berkaitan dengan sistem komputasi dan AI;
- ❖ pengembangan program serupa untuk masyarakat luas, dengan fokus khusus pada mereka yang terlibat di setiap tahap manajemen teknologi, termasuk pegawai negeri sipil, politisi, dan jurnalis;
- ❖ keterlibatan dengan inisiatif yang lebih luas seperti acara ITU AI for Good dan LSM yang bekerja untuk Tujuan Pembangunan Berkelanjutan PBB.

2.6 KESIMPULAN

Eropa, dan dunia pada umumnya, menghadapi kemunculan teknologi yang sangat menjanjikan bagi banyak aspek kehidupan manusia, namun tampaknya juga menimbulkan ancaman besar. Buku Putih ini dan khususnya Rekomendasi di bagian sebelumnya berupaya mendorong para pemangku kepentingan ke arah hasil yang lebih baik secara etis dan sosial dari pengembangan, perancangan, dan penerapan teknologi AI. Berdasarkan identifikasi kami terhadap peluang inti dan risiko AI bagi masyarakat serta serangkaian lima prinsip etika yang kami sintesiskan untuk memandu adopsinya, kami merumuskan 20 Poin Aksi dalam semangat kolaborasi dan demi menciptakan respons yang *konkret* dan *konstruktif* terhadap tantangan sosial paling mendesak yang ditimbulkan oleh AI.

Dengan laju perubahan teknologi yang pesat, kita mungkin tergoda untuk memandang proses politik dalam demokrasi liberal saat ini sebagai sesuatu yang kuno, ketinggalan zaman, dan tidak lagi mampu melestarikan nilai-nilai dan memajukan kepentingan masyarakat dan semua orang di dalamnya. Kami tidak sependapat. Dengan Rekomendasi yang kami tawarkan di sini, termasuk pembentukan pusat, lembaga, kurikulum, dan infrastruktur lainnya, kami telah memperjuangkan program pembuatan kebijakan dan inovasi teknologi yang ambisius, inklusif, dan berkeadilan. Program ini kami yakini akan berkontribusi dalam mengamankan manfaat dan memitigasi risiko AI, bagi semua orang, dan bagi dunia yang kita tinggali bersama.

BAB 3

ATURAN DALAM MEMBANGUN AI YANG TERPERCAYA

Bab ini membahas pentingnya panduan etika dalam pengembangan kecerdasan buatan (AI), dengan menyoroti kontribusi signifikan dari dokumen Pedoman Etika untuk AI yang Terpercaya yang diterbitkan oleh Komisi Eropa. Dokumen ini menawarkan tolok ukur yang jelas dan terstruktur untuk mengevaluasi apakah suatu sistem AI dikembangkan secara bertanggung jawab. Lebih dari itu, pedoman ini juga membuka peluang bagi terciptanya dukungan internasional terhadap solusi AI yang tidak hanya bermanfaat bagi umat manusia, tetapi juga memperhatikan keberlanjutan lingkungan.

Perkembangan AI telah membawa perubahan besar dalam berbagai aspek kehidupan manusia. Oleh karena itu, menjadi sangat penting untuk memastikan bahwa perubahan tersebut diarahkan dengan cara yang tepat dan bertanggung jawab. Salah satu tantangan utama yang dihadapi saat ini adalah bagaimana merancang sistem AI yang memiliki tingkat otonomi tertentu, tanpa mengorbankan nilai-nilai kemanusiaan. Potensi AI yang bersifat transformatif menuntut adanya keseimbangan etis baru antara otonomi manusia dan otonomi buatan. Dengan demikian, pendekatan etika yang komprehensif menjadi landasan utama dalam memastikan bahwa AI dikembangkan tidak hanya dengan kecanggihan teknologi, tetapi juga dengan komitmen terhadap nilai-nilai moral dan sosial yang universal.

3.1 PERENCANAAN YANG CERMAT, ALIH-ALIH PENGUJIAN BETA

Sebagai jenis baru keagenan cerdas dan otonomi, AI dapat membawa manfaat yang sangat besar secara individu, sosial, dan lingkungan. Ini bisa menjadi kekuatan untuk kebaikan di dunia yang semakin kompleks dan membutuhkan solusi canggih untuk menangani masalah berskala besar dan saling terkait 17 Tujuan Pembangunan Berkelanjutan PBB menunjukkan bahwa umat manusia sedang berjuang dengan banyak tantangan, di banyak bidang vital, dan tidaklah bijaksana untuk tidak memanfaatkan solusi AI.

Proses dan keputusan apa yang akan didelegasikan ke sistem AI, efek seperti apa yang akan ditimbulkan oleh trade-off antara manusia dan agensi buatan, dan bentuk penilaian, pengendalian, revisi, dan perbaikan apa yang harus diterapkan, merupakan pertanyaan krusial yang tidak boleh dijawab melalui coba-coba. AI tidak boleh diuji beta pada manusia atau lingkungan. Pengembangan AI membutuhkan pertimbangan dan konsensus sosio-politik, dengan mempertimbangkan strategi jangka panjang tentang jenis AI apa yang harus dikembangkan, untuk tujuan apa, untuk siapa, dan sesuai dengan prioritas etika apa. Ini adalah tujuan utama dari laporan pedoman etika dari Komisi Eropa (EC).

Laporan, yang diterbitkan pada 8 April 2019 setelah beberapa versi dan lebih dari 500 konsultasi publik, disusun oleh Kelompok Pakar Tingkat Tinggi (HLEG) yang independen. HLEG ditunjuk oleh EC pada Juni 2018 dan terdiri dari 52 pakar (pengungkapan: saya salah satunya), dengan keahlian relevan dari akademisi, masyarakat sipil, dan industri. Pekerjaan HLEG diharapkan untuk menginformasikan kebijakan dan undang-undang Uni Eropa (UE) tentang AI,

untuk mendukung implementasi strategi UE tentang AI, dan untuk berfungsi sebagai kelompok pengarah untuk pekerjaan Aliansi AI Eropa. Pedoman ini mendukung pendekatan yang bertanggung jawab terhadap pengembangan AI, yang harus (1) sah, menghormati semua hukum dan peraturan yang berlaku; (2) etis, menghormati prinsip dan nilai etika (Gambar 3.1 merangkum prinsip-prinsip yang mendasari pedoman, yang diinformasikan oleh penelitian AI4People) dan (3) tangguh, baik secara teknis maupun dalam hal lingkungan sosialnya.

Karena AI akan menjadi semakin penting dan meluas, AI harus bekerja dengan andal, dengan cara yang dapat dipercaya oleh siapa pun demi kebaikan umat manusia dan seluruh lingkungan. Alternatifnya adalah AI dapat disalahgunakan, digunakan secara berlebihan, atau kurang dimanfaatkan. Ketidakpastian etika melahirkan pengambilan risiko yang sembrono dan kehati-hatian yang berlebihan. Inilah mengapa pedoman ini sangat penting. Pedoman ini merupakan langkah yang baik ke arah yang benar, yaitu kerangka kerja yang jelas, bersama, dan disukai secara sosial untuk AI yang etis.

3.2 ETIKA PERTAMA YANG MENINFORMASIKAN LEGISLASI

Pedoman ini telah dipuji dan disambut baik oleh banyak pihak, tetapi juga dikritik karena dianggap lemah, karena merupakan bagian dari strategi pengaturan diri semata, yang tidak ditegakkan secara hukum, dan tidak bermanfaat, karena terlalu umum, dan bergabung dengan begitu banyak inisiatif lain yang sejauh ini berdampak kecil. Kritik-kritik ini dan yang serupa dapat dibantah. Pertama, pedoman ini memuat prinsip dan klarifikasi yang kuat, dalam hal ekspektasi sosial, dan konsisten dengan keadaan terkini perdebatan tentang etika AI. Tentu saja, baik hukum maupun etika tentang AI diperlukan. Pedoman ini mengandaikan dan selaras dengan legislasi Uni Eropa. Uni Eropa berada di garis depan perdebatan internasional tentang AI, juga berkat Peraturan Perlindungan Data Umum (GDPR). Etika dapat berkontribusi pada pembentukan legislasi baru (misalnya, tentang sistem pengenalan wajah) atau bertindak sebagai panduan jika tidak ada.

Terkadang, etika diperlukan untuk menafsirkan undang-undang yang ada (misalnya, GDPR). Di lain waktu, etika mungkin merekomendasikan untuk tidak melakukan sesuatu yang tidak dilarang oleh undang-undang (misalnya, menyerahkan keputusan medis sepenuhnya kepada algoritma tanpa pengawasan atau penjelasan), atau merekomendasikan untuk melakukan sesuatu yang tidak diwajibkan oleh undang-undang (misalnya, merancang algoritma yang meminimalkan dampak lingkungan dari pemanas sentral rumah tangga).

Dalam semua kasus ini, kepatuhan terhadap hukum diperlukan tetapi tidak cukup, dan, seperti yang diakui oleh pedoman, itu harus dilengkapi dengan pendekatan 'etika lunak' pasca-kepatuhan, karena hukum memberikan aturan main, tetapi tidak menunjukkan cara bermain dengan baik sesuai aturan. Kedua, memang: pedoman tersebut tidak terlalu orisinal atau inovatif, tetapi itu akan mengherankan dan mungkin sedikit mengkhawatirkan, setelah lebih dari setengah abad diskusi tentang topik tersebut. Faktanya, saat ini terdapat lebih dari 70 kerangka kerja dan daftar prinsip tentang etika AI.

Maraknya deklarasi ini menimbulkan inkonsistensi dan kebingungan di antara para pemangku kepentingan mengenai dokumen mana yang lebih disukai. Hal ini juga memberi

tekanan pada aktor swasta dan publik yang mengembangkan atau menerapkan solusi AI untuk membuat deklarasi mereka sendiri karena takut dianggap tertinggal, sehingga semakin menambah kegaduhan. Dan hal ini berisiko menciptakan supermarket prinsip dan nilai, tempat aktor swasta dan publik dapat mencari jenis etika yang paling tepat untuk membenarkan perilaku mereka, alih-alih merevisi perilaku mereka agar konsisten dengan kerangka etika yang diterima secara sosial.

Pedoman ini mengatasi tantangan-tantangan ini karena merupakan standar yang paling komprehensif dan otoritatif yang tersedia di Uni Eropa, menawarkan kerangka acuan yang jelas dan kosakata konseptual yang umum. Pedoman ini dirancang untuk menetapkan tolok ukur tentang apa yang mungkin atau mungkin tidak memenuhi syarat, mulai sekarang, sebagai AI yang tepercaya.

3.3 LANGKAH LANJUTAN UNTUK PANGGUNG GLOBAL

Dalam beberapa kasus, pendekatan regulasi mungkin terlalu prematur, terlalu preskriptif, atau menghambat inovasi yang berharga. Pendekatan etis menghasilkan ekspektasi yang lebih fleksibel dan tetap menuntut. Penting untuk diingat bahwa penerbitan pedoman ini juga hanyalah langkah awal. Pedoman ini akan berkontribusi untuk menginformasikan legislasi dan kebijakan Uni Eropa, tetapi juga merupakan peta jalan bagi transformasi cepat yang dimungkinkan oleh teknologi AI.

Pada Juni 2019, HLEG akan mengeluarkan rekomendasinya untuk agenda penelitian AI Uni Eropa, dan tentang bagaimana Uni Eropa dapat memperkuat daya saingnya dalam pengembangan dan penerapan AI, sejalan dengan pedoman tersebut. Dan musim panas ini, Komisi Eropa akan meluncurkan proyek percontohan untuk menguji pedoman tersebut, bekerja sama dengan para pemangku kepentingan, guna mengidentifikasi potensi perbaikan dan mempromosikan penerapan praktis. HLEG akan meninjau hasilnya pada awal 2020 dan menyempurnakan lebih lanjut keluarannya.

Dalam jangka panjang, Komisi Eropa “ingin membawa pendekatan etika AI ini ke panggung global memperkuat kerja sama dengan mitra yang sepaham seperti Jepang, Kanada, atau Singapura G7 dan G20” (Komisi Eropa 2019b). Beberapa kritikus mengakui semua ini, tetapi tetap keberatan bahwa seseorang tidak dapat menjadi pemimpin dalam AI yang etis tanpa menjadi pemimpin AI terlebih dahulu. Namun, ‘berinovasi dulu, perbaiki kemudian’ adalah kesalahan yang, dalam kasus AI, juga bisa sangat merugikan dan dapat menyebabkan reaksi publik terhadap AI, serupa dengan reaksi terhadap tanaman rekayasa genetika di masa lalu.

Bencana perubahan iklim dan masalah platform media sosial yang mengganggu demokrasi seharusnya mengajarkan kita untuk merencanakan inovasi dengan lebih cermat. Inilah sebabnya mengapa Uni Eropa ingin menentukan strategi jangka panjang di mana etika menjadi pendorong inovasi yang menawarkan keunggulan kompetitif, dan yang memastikan bahwa hak dan nilai-nilai fundamental dipupuk, kepentingan publik terpenuhi, dan lingkungan alam berkembang. Etika yang pertama adalah pendekatan yang tepat untuk menetapkan

standar global AI. Era "bergerak cepat dan memecahkan masalah" telah berakhir. Kini saatnya untuk "bersegera perlahan" (*festina lente*) dalam pengembangan AI.

	Tujuh hal penting untuk mencapai AI yang tepercaya
	AI yang tepercaya harus mematuhi semua hukum dan peraturan yang berlaku, serta serangkaian persyaratan; Daftar penilaian spesifik bertujuan untuk membantu memverifikasi penerapan setiap persyaratan utama:
1	Agensi dan pengawasan manusia: Sistem AI harus memungkinkan masyarakat yang setara dengan mendukung agensi manusia dan hak-hak dasar, dan tidak mengurangi, membatasi, atau menyesatkan otonomi manusia.
2	Ketahanan dan keamanan: AI yang tepercaya membutuhkan algoritma yang aman, andal, dan cukup tangguh untuk menangani kesalahan atau inkonsistensi selama semua fase siklus hidup sistem AI.
3	Privasi dan tata kelola data: Warga negara harus memiliki kendali penuh atas data mereka sendiri, sementara data yang menyangkut mereka tidak akan digunakan untuk merugikan atau mendiskriminasi mereka.
4	Transparansi: Ketertelusuran sistem AI harus dipastikan.
5	Keragaman, non-diskriminasi, dan keadilan: Sistem AI harus mempertimbangkan seluruh rentang kemampuan, keterampilan, dan persyaratan manusia, serta memastikan aksesibilitas.
6	Kesejahteraan masyarakat dan lingkungan: Sistem AI harus digunakan untuk meningkatkan perubahan sosial yang positif dan meningkatkan keberlanjutan serta tanggung jawab ekologis.
7	Akuntabilitas: Mekanisme harus diterapkan untuk memastikan tanggung jawab dan akuntabilitas untuk sistem AI dan hasilnya.
	Komisi Eropa, ref. IP/19/1893

BAB 4

KEBIJAKAN, ETIKA, DAN REGULASI TERHADAP AI

Pada Juli 2017, Dewan Negara Tiongkok merilis dokumen strategis bertajuk Rencana Pengembangan Kecerdasan Buatan Generasi Baru (新一代人工智能发展规划), yang menetapkan visi ambisius Tiongkok untuk menjadi pemimpin global dalam bidang kecerdasan buatan (AI) pada tahun 2030. Strategi ini mencakup tujuan untuk membangun industri AI senilai Rp.2,5 triliun, sekaligus mendorong peran Tiongkok dalam merumuskan norma dan standar etika global terkait AI.

Berbagai analisis sebelumnya umumnya berfokus pada aspek teknis atau komponen kebijakan tertentu dari strategi AI Tiongkok. Namun, dalam bab ini, perhatian diarahkan pada latar belakang sosial-politik serta dinamika perdebatan kebijakan yang melatarbelakangi lahirnya strategi tersebut. Fokus utama kami adalah pada area-area strategis yang menjadi prioritas investasi pemerintah Tiongkok dalam pengembangan AI, serta perdebatan etika yang menyertainya dan turut memengaruhi batasan penggunaan teknologi tersebut.

Dengan menelaah secara kritis kerangka kebijakan dan dokumen-dokumen resmi yang relevan, bab ini bertujuan untuk memberikan pemahaman yang lebih menyeluruh tentang bagaimana kebijakan AI Tiongkok dirancang, dijalankan, dan dinegosiasikan dalam konteks sosial-politik yang kompleks. Pendekatan ini diharapkan dapat membantu pembaca memahami bukan hanya arah teknologis, tetapi juga dimensi etis dan ideologis di balik kebangkitan Tiongkok sebagai kekuatan utama dalam ekosistem AI global.

4.1 PENDAHULUAN

Pada bulan Maret 2016, AI Google DeepMind yang dirancang untuk bermain permainan papan Go (AlphaGo) mengalahkan Lee Sedol, seorang pemain Go profesional Korea Selatan. Saat itu, Sedol meraih kemenangan terbanyak kedua dalam kejuaraan Go internasional, namun kalah dari AlphaGo dengan skor empat banding satu. Meskipun pertandingan tersebut mendapat liputan di Barat, pertandingan tersebut merupakan acara besar di Tiongkok, di mana lebih dari 280 juta orang menontonnya secara langsung.

Dua orang dalam pemerintahan menggambarkan pertandingan ini sebagai 'momen Sputnik' bagi pengembangan Kecerdasan Buatan (AI) di Tiongkok. Meskipun sebelumnya telah ada inisiatif kebijakan AI di negara tersebut, kemenangan AlphaGo berkontribusi pada peningkatan fokus, sebagaimana ditunjukkan oleh 'Rencana Pengembangan Kecerdasan Buatan Generasi Baru' (AIDP) 2017. AIDP menetapkan tujuan strategis dan menguraikan tujuan menyeluruh untuk menjadikan Tiongkok pemimpin dunia dalam AI pada tahun 2030.

Sejumlah kecil laporan telah mencoba menilai plausibilitas strategi AI Tiongkok mengingat kemampuan teknis Tiongkok saat ini. Laporan lainnya berupaya memahami area pengembangan tertentu, misalnya, keamanan atau pertumbuhan ekonomi. Namun, untuk memahami implikasi dan arah AIDP yang luas, tidaklah cukup hanya menganalisis elemen-

elemen spesifik secara terpisah atau hanya mempertimbangkan kemampuan teknis. Sebaliknya, analisis yang lebih komprehensif dan kritis terhadap kekuatan pendorong di balik strategi AI Tiongkok, ekonomi politiknya, kekhususan budayanya, dan perdebatan kebijakan relevan saat ini, diperlukan untuk memahami strategi AI Tiongkok. Inilah tugas yang kami lakukan dalam Bab ini.

Untuk memberikan pemahaman kontekstual ini, Bagian 4.2 memetakan tata kelola AI yang relevan di Tiongkok. Kami berpendapat bahwa, meskipun inisiatif kebijakan sebelumnya telah menyatakan niat untuk mengembangkan AI, upaya-upaya ini terpecah-pecah dan memandang AI sebagai salah satu dari banyak alat dalam mencapai tujuan yang berbeda. Sebaliknya, AIDP adalah upaya tata kelola tingkat nasional pertama yang berfokus secara eksplisit pada pengembangan AI sebagai strategi terpadu.

Setelah ini, Bagian 4.3 menganalisis intervensi dan dampak AIDP pada tiga bidang strategis yang diidentifikasi dalam dokumen tersebut, yaitu: *persaingan internasional*, *pertumbuhan ekonomi*, dan *tata kelola sosial*. Bagian 4.4 berfokus pada tujuan Tiongkok untuk mengembangkan *norma dan standar etika untuk AI*. Di sana kami berpendapat bahwa, meskipun perdebatan masih dalam tahap awal, keinginan untuk mendefinisikan batasan normatif untuk penggunaan AI yang dapat diterima ada dan mendesak.

Secara keseluruhan, artikel ini berupaya memberikan pemahaman yang rinci dan kritis tentang alasan di balik, dan lintasan strategi AI Tiongkok saat ini. Hal ini menekankan bahwa pemerintah Tiongkok menyadari potensi manfaat, risiko praktis, dan tantangan etika yang dihadirkan oleh AI, dan bahwa arah strategi AI Tiongkok sebagian besar akan ditentukan oleh interaksi faktor-faktor ini dan sejauh mana kepentingan pemerintah dapat lebih diutamakan daripada masalah etika. Bagian 4.5 menyimpulkan makalah ini dengan merangkum temuan-temuan utama analisis kami.

4.2 TATA KELOLA AI DI TIONGKOK

Sejak 2013, Tiongkok telah menerbitkan beberapa dokumen kebijakan tingkat nasional yang mencerminkan niat untuk mengembangkan dan menerapkan AI di berbagai sektor. Misalnya, pada tahun 2015, Dewan Negara merilis pedoman tentang aksi 'Internet+' Tiongkok. Pedoman ini bertujuan untuk mengintegrasikan internet ke dalam semua elemen ekonomi dan masyarakat. Dokumen tersebut dengan jelas menyatakan pentingnya mengembangkan industri AI yang sedang berkembang dan berinvestasi dalam penelitian dan pengembangan ("*Internet Plus*" 2015). Pada tahun yang sama, rencana sepuluh tahun '*Made in China 2025*' dirilis, dengan tujuan untuk mengubah Tiongkok menjadi pemain dominan dalam manufaktur berteknologi tinggi global, termasuk AI. Contoh penting lainnya adalah Rencana Lima Tahun ke-13 Komite Sentral Partai Komunis Tiongkok, yang diterbitkan pada Maret 2016.

Dokumen tersebut menyebutkan AI sebagai salah satu dari enam bidang penting untuk mengembangkan industri-industri yang sedang berkembang di negara tersebut, dan sebagai faktor penting dalam merangsang pertumbuhan ekonomi. Jika dibaca bersama-sama, dokumen-dokumen ini menunjukkan bahwa telah ada upaya sadar untuk mengembangkan dan menggunakan AI di Tiongkok selama beberapa waktu, bahkan sebelum 'momen Sputnik'.

Namun, sebelum 2016, AI hanya disajikan sebagai salah satu teknologi di antara banyak teknologi lainnya, yang dapat berguna dalam mencapai berbagai tujuan kebijakan. Hal ini berubah dengan dirilisnya AIDP.

Rencana Pengembangan Kecerdasan Buatan Generasi Baru (AIDP)

Dirilis pada Juli 2017 oleh Dewan Negara (yang merupakan badan administratif utama di Tiongkok), 'Rencana Pengembangan Kecerdasan Buatan Generasi Baru' (AIDP) bertindak sebagai dokumen terpadu yang menguraikan tujuan kebijakan AI Tiongkok. Media Tiongkok menyebutnya sebagai 'tahun pertama strategi pengembangan AI Tiongkok. Tujuan utama kebijakan ini, sebagaimana diartikulasikan oleh AIDP, adalah menjadikan Tiongkok pusat inovasi AI dunia pada tahun 2030, dan menjadikan AI 'penggerak utama bagi peningkatan industri dan transformasi ekonomi Tiongkok.

AIDP juga menunjukkan pentingnya penggunaan AI di berbagai sektor yang lebih luas, termasuk pertahanan dan kesejahteraan sosial, serta berfokus pada kebutuhan untuk mengembangkan standar dan norma etika dalam penggunaan AI. Secara keseluruhan, Rencana ini menyediakan strategi AI yang komprehensif, dan menantang kekuatan-kekuatan terkemuka lainnya di banyak bidang utama.

AIDP menjabarkan tiga langkah kunci, yang masing-masing berisi serangkaian tujuan, beberapa di antaranya didefinisikan secara ketat, sementara yang lainnya kurang jelas. Tujuan-tujuan tersebut dirangkum sebagai berikut dan pada Gambar 5.1 di bawah ini:

1. Pada tahun 2020, Tiongkok bertujuan untuk mempertahankan daya saing dengan kekuatan-kekuatan besar lainnya dan mengoptimalkan lingkungan pengembangan AI-nya. Secara moneter, Tiongkok bermaksud menciptakan industri AI senilai lebih dari Rp.800 miliar. Terakhir, Tiongkok berupaya menetapkan norma, kebijakan, dan regulasi etika awal untuk bidang-bidang vital AI.
2. Pada tahun 2025, Tiongkok menargetkan untuk mencapai 'terobosan besar' (sebagaimana dinyatakan dalam dokumen) dalam teori dasar AI dan menjadi yang terdepan di dunia dalam beberapa aplikasi ('beberapa teknologi dan aplikasi mencapai tingkat terdepan di dunia'). Secara keseluruhan, Tiongkok menargetkan peningkatan nilai industri inti AI-nya menjadi lebih dari Rp. 1 Triliun, dan berencana untuk memperluas, dan mengesahkan, standar etika untuk AI.
3. Pada tahun 2030, Tiongkok berupaya menjadi pusat inovasi AI dunia. Pada saat itu, pertumbuhan industri inti AI diperkirakan akan meningkat lebih dari dua kali lipat dan bernilai Rp.15 triliun, dan peningkatan lebih lanjut dalam undang-undang dan standar juga diharapkan, untuk menghadapi tantangan baru yang muncul.

Implementasi AIDP

Rencana ini akan dipandu oleh Komite Penasihat Strategi AI yang baru, yang dibentuk pada November 2017, dan akan dikoordinasikan oleh Kementerian Sains dan Teknologi (MIST), bersama dengan Kantor Promosi Rencana AI, dan badan-badan terkait lainnya. Meskipun badan-badan ini akan memberikan panduan terpusat, Rencana ini tidak dimaksudkan untuk bertindak sebagai inisiatif yang diberlakukan secara terpusat. AIDP justru berfungsi sebagai

stempel persetujuan untuk mengurangi risiko dan secara aktif memberi insentif kepada proyek-proyek lokal yang memanfaatkan AI.

Menyadari hal ini penting: AIDP adalah strategi ambisius yang ditetapkan oleh pemerintah pusat, tetapi inovasi dan transformasi yang sesungguhnya diharapkan akan didorong oleh sektor swasta dan pemerintah daerah. Dengan kata lain, lebih tepat untuk memandang AIDP sebagai 'daftar keinginan' yang sangat berinsentif, untuk mendorong dan mengoordinasikan para pemangku kepentingan terkait lainnya, alih-alih sebagai arahan terpusat. Inilah sebabnya rencana tiga tahun yang mempromosikan AIDP (2018–2020) menekankan koordinasi antarprovinsi dan dengan pemerintah daerah.

Terkait sektor swasta, Tiongkok telah memilih 'juara nasional AI': bisnis yang didukung oleh pemerintah untuk berfokus pada pengembangan sektor-sektor AI tertentu. Misalnya, Baidu ditugaskan untuk mengembangkan kendaraan otonom, Alibaba untuk pengembangan kota pintar, dan Tencent untuk visi komputer untuk diagnosis medis. Didukung sebagai juara nasional melibatkan kesepakatan di mana perusahaan swasta setuju untuk berfokus pada tujuan strategis pemerintah. Sebagai imbalannya, perusahaan-perusahaan ini menerima penawaran kontrak preferensial, akses yang lebih mudah ke keuangan, dan terkadang perlindungan pangsa pasar. Meskipun perusahaan lain dapat bersaing di bidang-bidang ini, secara historis status 'juara nasional' telah membantu perusahaan-perusahaan besar mendominasi sektor mereka masing-masing.

Dengan demikian, 'tim nasional' AI yang baru ini berbeda dari juara nasional yang disponsori negara sebelumnya karena mereka telah sukses secara internasional di bidangnya masing-masing, terlepas dari perlakuan istimewa ini. Lebih lanjut, terdapat persaingan domestik yang ketat di area-area tempat juara nasional telah terpilih. Hal ini menunjukkan bahwa persaingan mungkin tidak terhambat dengan cara tradisional. Misalnya, semua perusahaan yang terpilih sebagai juara nasional AI sedang mengembangkan teknologi di area kota pintar yang ditunjuk Alibaba.

Sejalan dengan ini, patronase tidak melarang perusahaan kecil mendapatkan keuntungan dari struktur insentif finansial. Perusahaan rintisan teknologi di Tiongkok sering menerima dukungan dan subsidi pemerintah untuk mengembangkan teknologi AI. Sebagai contoh, Zhongguancun Innovation Town adalah ruang kerja inkubator yang dibangun khusus dan disubsidi pemerintah yang menyediakan serangkaian layanan untuk membantu perusahaan rintisan teknologi Tiongkok agar berhasil, sering kali di sektor-sektor tempat juara nasional telah terpilih. Terakhir, ada juga kasus-kasus di mana tidak ada dukungan khusus. Misalnya, sementara AIDP mempromosikan pengadilan pintar, dengan keinginan yang dinyatakan untuk mengembangkan AI untuk pengumpulan bukti, analisis kasus, dan pembacaan dokumen hukum, hingga April 2020 belum ada juara nasional yang dipilih untuk mengembangkan aplikasi AI untuk administrasi peradilan.



Gambar 4.1 Visualisasi AIDP Tiongkok

Mengenai pemerintah daerah, struktur politik di Tiongkok menciptakan sistem insentif untuk memenuhi tujuan kebijakan pemerintah nasional. Batasan jangka pendek untuk politisi provinsi, dan promosi berdasarkan kinerja ekonomi memberikan insentif yang kuat untuk mengikuti inisiatif pemerintah yang ditetapkan oleh pusat. Dengan demikian, pemerintah daerah menjadi tempat yang subur untuk menguji dan mengembangkan kebijakan pemerintah pusat. Kekuatan sistem insentif ini dapat dilihat dari keputusan yang dibuat oleh pemerintah kota Tianjin untuk menetapkan dana Rp.10 miliar untuk pengembangan AI, sekitar waktu yang sama dengan penerbitan AIDP.

Pada saat yang sama, penting untuk menyadari bagaimana tidak adanya tinjauan akuntabilitas yang efektif atas pengeluaran pemerintah daerah menciptakan masalah dalam sistem ini. Khususnya, hal ini telah memfasilitasi pola pikir di mana politisi daerah tahu bahwa pemerintah pusat akan menyelamatkan mereka untuk proyek-proyek yang gagal, yang menyebabkan manajemen anggaran yang buruk. Contoh nyata dari hal ini adalah inisiatif pembangunan pelabuhan skala besar yang dikembangkan oleh pemerintah provinsi di provinsi-provinsi pesisir timur yang lebih didasarkan pada prestise daripada alasan ekonomi apa pun, dan yang menyebabkan kelebihan kapasitas dan persaingan yang tidak teratur.

Struktur insentif ini mengandung perbedaan yang halus. Sebuah tim nasional telah dipilih untuk memimpin penelitian dan pengembangan di beberapa area strategis yang telah ditentukan. Selain perusahaan-perusahaan terpilih ini, hanya ada sedikit pedoman khusus yang diberikan kepada industri dan agen pemerintah daerah mengenai item mana yang harus dikejar dalam 'daftar keinginan' AIDP. Hal ini memungkinkan perusahaan untuk memilih teknologi yang ingin mereka kembangkan, dan memberi pemerintah daerah pilihan mitra sektor swasta untuk mengintegrasikan AI ke dalam infrastruktur atau tata kelola kota. Dokumentasi selanjutnya telah menekankan pentingnya memperkuat organisasi dan implementasi, termasuk antara provinsi dan kementerian, namun belum jelas bagaimana

koordinasi ini akan berfungsi dalam praktiknya. Dengan demikian, AIDP dapat berfungsi sebagai 'daftar keinginan', tetapi panduan, insentif, dan risiko yang tepat berbeda-beda, bergantung pada jenis pemangku kepentingan.

AIDP tidak boleh dibaca secara terpisah ketika mempertimbangkan strategi AI Tiongkok, tetapi AIDP memberikan indikasi yang paling transparan dan berpengaruh tentang kekuatan pendorong di balik strategi AI Tiongkok. Karena signifikansi AIDP (dalam hal kebijakan) dan pentingnya (dalam hal strategi), di sisa artikel ini, kami akan menggunakannya sebagai kerangka organisasi untuk menjelaskan pendorong dan batasan etika yang membentuk pendekatan Tiongkok terhadap AI.

4.3 FOKUS STRATEGIS AI TIONGKOK

AIDP memberikan perspektif longitudinal mengenai situasi strategis Tiongkok terkait AI, termasuk *kapabilitas komparatifnya*, *peluang* yang ditawarkan, dan potensi risikonya. Dengan pendekatan yang mengutamakan teknologi, mungkin tergoda untuk memfokuskan perhatian pada kapabilitas AI yang telah disebutkan sebelumnya, untuk mendapatkan wawasan tentang jenis teknologi yang diinvestasikan Tiongkok.

Namun, hal ini kemungkinan hanya menawarkan perspektif jangka pendek dan akan segera menjadi usang seiring pesatnya kemajuan inovasi teknologi. Lebih lanjut, hal ini tidak akan banyak menjelaskan mengapa Tiongkok berupaya mengembangkan sektor AI yang kuat dalam beberapa dekade mendatang. Untuk tujuan ini, akan lebih bermanfaat untuk mencoba memahami fokus strategis Tiongkok dari pendekatan yang mengutamakan kebijakan, dengan menganalisis area-area di mana Tiongkok menganggap AI menghadirkan peluang.

Di bagian ini, kami berfokus pada area-area yang sangat penting bagi Tiongkok ini, tentang bagaimana dan apa yang diharapkan Tiongkok dari pengembangan AI di masing-masing area tersebut, dan pada beberapa risiko yang dirasakan di masing-masing area ini. AIDP menyoroti tiga area di mana AI dapat membuat perbedaan substansial di Tiongkok: *persaingan internasional*, *pembangunan ekonomi*, dan *tata kelola sosial*. Ketiganya saling terkait erat, tetapi demi kejelasan, kami akan menganalisisnya secara terpisah, dan mengontekstualisasikan masing-masing dengan membahas literatur relevan seputar latar belakang politik yang lebih luas dan perdebatan kebijakan kontemporer.

Persaingan Internasional

AIDP menyatakan bahwa AI telah menjadi fokus baru persaingan internasional dan bahwa 'pengembangan AI sebuah strategi utama untuk meningkatkan daya saing nasional dan melindungi keamanan nasional'. AIDP menekankan bahwa Tiongkok harus memanfaatkan peluang strategis yang diberikan oleh AI untuk membuat 'perkembangan yang melampaui batas' dalam kapabilitas militer. Meskipun Tiongkok dan AS sering digambarkan sebagai rival geopolitik, anggaran militer kedua kekuatan tersebut tetap berbeda secara signifikan. Tiongkok memiliki anggaran militer terbesar kedua di dunia, dengan alokasi Rp.175 Triliun pada tahun 2019, tetapi pengeluarannya masih hanya sepertiga dari anggaran AS. Alih-alih mengungguli AS dalam persenjataan konvensional, Tiongkok menganggap investasi dalam AI sebagai

peluang untuk membuat terobosan radikal dalam teknologi militer dan dengan demikian menyalip AS.

Upaya untuk menggunakan teknologi guna menantang hegemoni AS bukanlah hal baru dalam strategi militer Tiongkok. Sejak akhir 1990-an, negara tersebut telah mengikuti kebijakan '*shashoujian*' (杀手锏), yang secara kasar diterjemahkan sebagai 'kartu truf'. Alih-alih bersaing langsung dengan AS, Tiongkok telah berupaya mengembangkan kapabilitas asimetris, yang dapat memberikan keuntungan penting dalam peperangan dan pencegahan yang kredibel di masa damai. Strategi kartu truf ini berupaya menggunakan teknologi yang tidak lazim untuk menyerang kelemahan musuh guna mendapatkan inisiatif dalam perang. Pendekatan kartu truf ini digaungkan oleh mantan Ketua Partai, Jiang Zemin, yang menekankan bahwa teknologi harus menjadi fokus utama militer, terutama teknologi yang "paling ditakuti musuh".

Satu area di mana Tiongkok telah mengembangkan taktik asimetris ini adalah perang siber, di mana kapabilitas telah dikembangkan untuk menargetkan jaringan kritis pertempuran militer AS, jika diperlukan. Bersamaan dengan ini, bukti menunjukkan penggunaan serangan siber yang terus-menerus untuk mengumpulkan intelijen ilmiah, teknologi, dan komersial. Posisi Tiongkok mengenai kapabilitas ini ambivalen. Di satu sisi, Tiongkok telah secara resmi mempromosikan inisiatif internasional untuk mengatur aktivitas yang dijalankan negara yang bermusuhan di dunia maya, dan untuk mengisi kesenjangan regulasi yang ada untuk perilaku negara dalam domain ini.

Misalnya, Tiongkok ikut mensponsori *Kode Etik Internasional untuk Keamanan Informasi* di Majelis Umum PBB pada September 2011, yang mengupayakan komitmen untuk tidak menggunakan teknologi informasi dalam tindakan agresi dan telah memberikan dukungan berkelanjutan untuk dialog oleh Kelompok Ahli Pemerintah PBB dalam mencegah konflik siber. Di sisi lain, Tiongkok juga telah menjalankan operasi siber yang menargetkan infrastruktur AS dan bertujuan untuk mengekstraksi informasi komersial dan ilmiah serta memperoleh intelijen yang relevan terhadap beberapa negara, termasuk Australia, Filipina, Hong Kong, dan AS.

Keinginan untuk melampaui AS digaungkan dalam pernyataan dari para pemimpin politik dan militer Tiongkok. Misalnya, Presiden Xi Jinping menyatakan pada tahun 2017 bahwa 'dalam situasi persaingan militer internasional yang semakin ketat, hanya para inovator yang menang'. Sentimen ini diamini oleh Letnan Jenderal Liu Guozhi, wakil Kongres Nasional ke-19 dan direktur Komite Sains dan Teknologi Komisi Militer Pusat, yang menyatakan dalam sebuah wawancara bahwa AI menghadirkan peluang langka untuk mengambil jalan pintas guna mencapai inovasi dan melampaui para pesaing.

Secara paralel, akademisi yang berafiliasi dengan Tentara Pembebasan Rakyat (PLA) menyoroti bahwa AI akan digunakan untuk memprediksi situasi medan perang dan mengidentifikasi pendekatan yang optimal, memfasilitasi 'kemenangan sebelum perang'. Beberapa anggota PLA melangkah lebih jauh dari ini dengan mengantisipasi 'singularitas' medan perang, di mana AI melampaui pengambilan keputusan manusia. Pernyataan-

pernyataan ini menekankan keyakinan, yang tersebar luas di seluruh kalangan militer dan pertahanan Tiongkok, tentang pentingnya memanfaatkan teknologi yang sedang berkembang termasuk AI untuk mencapai keunggulan militer yang kompetitif.

Seiring dengan perkembangan ekonomi dan militer Tiongkok, fokus strategi militer negara tersebut juga semakin matang. Selama beberapa tahun terakhir, strategi Tiongkok telah menyatu di sekitar upaya untuk mengembangkan 'senjata konsep baru' untuk melampaui kemampuan militer AS. Ini tidak terbatas pada AI saja, dan berlaku untuk investasi Tiongkok di bidang teknologi militer baru lainnya, seperti persenjataan hipersonik. Oleh karena itu, upaya Tiongkok untuk menggunakan teknologi guna mendapatkan keunggulan dalam urusan militer tidak boleh dilihat sebagai sesuatu yang baru, tetapi dipahami dalam konteks sejarah yang lebih luas untuk menemukan cara-cara inovatif untuk menantang hegemoni AS.

Meskipun dorongan untuk pengembangan lompatan menandai kelanjutan dari kebijakan sebelumnya, terdapat indikasi kuat yang bersamaan bahwa para pejabat Tiongkok juga khawatir AI akan menyebabkan perlombaan senjata dan potensi eskalasi militer. Pernyataan para pejabat senior tampaknya menunjukkan keyakinan pada kerja sama dan pengendalian senjata untuk memitigasi risiko yang ditimbulkan oleh pengembangan militer AI. Secara khusus, tiga risiko utama menjadi pusat perdebatan:

- (i) keterlibatan dan pengendalian manusia setelah senjata berbasis AI dikerahkan;
- (ii) tidak adanya norma yang terdefinisi dengan baik untuk perilaku negara dan penggunaan senjata AI; yang pada gilirannya meningkatkan
- (iii) kemungkinan mispersepsi atau eskalasi konflik yang tidak disengaja.

Kekhawatiran ini mendasari dukungan Tiongkok untuk membatasi penggunaan senjata otonom, sebagaimana diungkapkan pada Konvensi ke-5 tentang Senjata Konvensional Tertentu (*"Makalah Posisi Tiongkok"* 2016) dan, baru-baru ini, keinginan untuk melarang senjata mematikan otonom. Terlepas dari kekhawatiran (i)–(iii), penting untuk menekankan bahwa Tiongkok adalah aktor yang menerapkan strategi paling agresif dalam mengembangkan AI untuk keperluan militer di antara kekuatan militer utama.

Menelaah lebih lanjut tindakan Tiongkok di panggung internasional sangatlah menarik. Larangan yang dianjurkan Tiongkok hanya mencakup *penggunaan*, bukan *pengembangan atau produksi* sistem senjata mematikan otonom. Dengan demikian, larangan tersebut tidak akan mencegah keberadaan senjata mematikan otonom yang berfungsi sebagai pencegah, sama seperti Tiongkok memiliki doktrin 'tidak boleh digunakan pertama kali' (NFU) untuk senjata nuklir. Lebih lanjut, definisi otonomi yang dianut Tiongkok sangat sempit, hanya mencakup senjata yang sepenuhnya otonom.

Beberapa komentator berpendapat bahwa penjajaran antara kekhawatiran yang hati-hati tentang pengerahan, di satu sisi, dan pendekatan agresif terhadap pengembangan, di sisi lain, dapat dijelaskan oleh upaya Tiongkok untuk memberikan tekanan pada militer lain yang masyarakat demokratisnya lebih sensitif terhadap kontroversi penggunaan senjata otomatis. Klaim ini masuk akal: kelanjutan propaganda mungkin menjadi bagian dari penjelasannya. Misalnya, Tiongkok adalah negara nuklir pertama yang berjanji 'tidak akan menggunakan senjata nuklir untuk pertama kalinya' (sejauh ini hanya India yang memiliki janji serupa;

negara-negara lain, termasuk AS dan Inggris, telah berjanji untuk menggunakan senjata nuklir hanya untuk tujuan defensif). Namun, alih-alih menawarkan komitmen yang tulus terhadap NFU, janji ini dimaksudkan sebagai alat propaganda internal dan eksternal, yang akan dielakkan oleh semantik jika diperlukan.

Secara keseluruhan, fokus Tiongkok pada AI militer dapat dianggap sebagai kelanjutan dari strategi jangka panjang, yang mengutamakan pengembangan (dengan ancaman pengerahan) teknologi untuk mendapatkan keuntungan militer. Masih terdapat pengakuan yang disadari, oleh beberapa aktor di Tiongkok, bahwa pengembangan AI menghadirkan risiko yang sangat besar untuk memicu perlombaan senjata atau menyebabkan eskalasi yang tidak disengaja karena otonomi teknologi ini. Namun pada tingkat politik, upaya untuk membatasi penggunaan AI militer secara internasional juga dapat dilihat sebagai bagian dari strategi propaganda.

Pembangunan Ekonomi

Pembangunan ekonomi merupakan peluang strategis kedua yang secara eksplisit disebutkan dalam AIDP. Dinyatakan bahwa AI akan menjadi kekuatan pendorong di balik babak baru transformasi industri, yang akan 'menyuntikkan energi kinetik baru ke dalam pertumbuhan ekonomi Tiongkok'. Rekonstruksi aktivitas ekonomi ditargetkan di semua sektor, dengan manufaktur, pertanian, logistik, dan keuangan menjadi contoh yang dipromosikan dalam AIDP.

Pertumbuhan pesat Tiongkok sering disebut sebagai 'keajaiban ekonomi', karena negara tersebut beralih dari ekonomi dengan pertumbuhan lambat menjadi menikmati beberapa tingkat pertumbuhan tertinggi di dunia selama lebih dari dua dekade. Sejumlah faktor memfasilitasi pertumbuhan ekonomi ini, salah satunya adalah dividen demografi. Tenaga kerja yang besar, dikombinasikan dengan populasi yang bergantung kecil, mendorong tingkat tabungan yang tinggi dan investasi yang besar.

Perubahan struktural, termasuk pergeseran sadar dari ekonomi yang didominasi pertanian ke ekonomi manufaktur, dan pembukaan pasar, merupakan faktor-faktor tambahan yang krusial. Pada tahun 2012, pertumbuhan angkatan kerja Tiongkok turun hingga mendekati nol, dan pergeserannya dari ekonomi pertanian ke ekonomi manufaktur sebagian besar telah matang. Tren ini telah mendorong para pembuat kebijakan Tiongkok untuk menyadari bahwa model pembangunan alternatif diperlukan untuk mempertahankan tingkat pertumbuhan yang tinggi. Model ini bertumpu pada pergeseran dari investasi besar-besaran di industri ke pertumbuhan yang didorong oleh masyarakat yang inovatif. Baru-baru ini, sains dan teknologi telah diajukan sebagai sarana krusial untuk mencapai jenis pertumbuhan inovatif ini.

Beberapa komentator berpendapat bahwa mempertahankan tingkat pertumbuhan yang tinggi ini sangat penting bagi Tiongkok karena adanya perdagangan implisit kebebasan politik oleh warga negara demi pertumbuhan ekonomi dan eborjuisemen. Penelitian telah menyoroti bahwa dukungan terhadap partai dan keinginan yang relatif lemah untuk demokrasi berasal dari kepuasan terhadap pekerjaan dan aspek-aspek material kehidupan, terutama di kalangan kelas menengah.

Pertumbuhan ekonomi yang melambat kemungkinan akan menimbulkan ketidakpuasan di kalangan masyarakat dan membuat ciri-ciri inheren dalam sistem politik Tiongkok, seperti korupsi, kurang dapat ditoleransi. Kurangnya saluran demokratis untuk frustrasi ini dapat menurunkan dukungan keseluruhan yang saat ini diterima pemerintah. Beberapa pihak berpendapat bahwa hal ini menciptakan dinamika 'demokratisasi atau mati', namun hal ini mungkin tidak layak, mengingat kendali politik Tiongkok.

Dengan latar belakang ini, sebuah laporan oleh PwC menunjukkan bahwa Tiongkok adalah negara yang paling diuntungkan dari AI, dengan peningkatan PDB hingga 26% pada tahun 2030. Perkiraan juga menunjukkan bahwa AI dapat memfasilitasi peningkatan lapangan kerja sebesar 12% selama dua dekade mendatang ("Dampak Bersih AI terhadap Pekerjaan di Tiongkok" 2018). Karena potensi manfaat ini, Presiden Xi sering menekankan pentingnya AI bagi pembangunan ekonomi negara secara keseluruhan. Tiongkok telah lama mengupayakan potensi manfaat ekonomi AI secara konkret dan proaktif. Misalnya, telah terjadi peningkatan 500% dalam pemasangan pembaruan robotik tahunan sejak 2012. Angka ini sangat mencengangkan, terutama jika dibandingkan dengan angka di atas 100% di Eropa, yang setara dengan lebih dari dua kali lipat jumlah pemasangan robot di Tiongkok dibandingkan di Eropa.

AI bisa menjadi pedang bermata dua, karena manfaat dan peningkatan yang dihasilkan AI disertai risiko, antara lain, disrupsi pasar tenaga kerja. Hal ini merupakan kekhawatiran yang dinyatakan secara eksplisit dalam AIDP. Meskipun laporan PwC yang disebutkan sebelumnya memprediksi bahwa otomatisasi akan meningkatkan jumlah lapangan kerja bersih di Tiongkok, disrupsi kemungkinan akan menyebar secara tidak merata. Otomatisasi yang lebih cerdas akan berdampak paling langsung pada pekerjaan berketerampilan rendah dan menengah, sekaligus menciptakan peluang bagi peran teknis berketerampilan tinggi.

Tiongkok telah aktif dalam upayanya untuk beradaptasi dengan risiko terkait AI tersebut, terutama dengan perombakan pendidikan yang dipromosikan oleh 'Rencana Reformasi dan Pembangunan Pendidikan Jangka Menengah dan Panjang Nasional (2010–2020)'. Rencana ini bertujuan untuk mendukung tenaga kerja terampil yang dibutuhkan di era informasi. Senada dengan itu, Tiongkok secara khusus mengatasi kekurangan keterampilan AI dengan menawarkan program pendidikan tinggi tentang subjek tersebut. Oleh karena itu, Tiongkok tampaknya lebih siap daripada negara-negara berpenghasilan menengah lainnya dalam menghadapi tantangan jangka panjang otomatisasi.

Meskipun upaya-upaya ini akan membantu mengembangkan keahlian yang dibutuhkan dalam jangka menengah dan panjang, upaya-upaya tersebut tidak banyak membantu meringankan perubahan struktural jangka pendek. Perkiraan menunjukkan bahwa, pada tahun 2030, otomatisasi di sektor manufaktur mungkin telah menggantikan seperlima dari seluruh pekerjaan di sektor ini, setara dengan 100 juta pekerja. Perubahan-perubahan ini sudah berlangsung, dengan robot telah menggantikan hingga 40% pekerja di beberapa perusahaan di provinsi-provinsi manufaktur ekspor Tiongkok, yaitu Zhejiang, Jiangsu, dan Guangdong. Di kota Dongguan di selatan saja, laporan menunjukkan bahwa 200.000 pekerja telah digantikan oleh robot. Jika hal ini dipadukan dengan peringkat internasional Tiongkok yang rendah dalam program transisi tenaga kerja untuk pelatihan kejuruan, dapat disimpulkan

bahwa konsekuensi jangka pendek dari transformasi yang dipimpin AI kemungkinan besar adalah gangguan yang signifikan terhadap tenaga kerja, yang berpotensi memperburuk kesenjangan yang semakin meningkat di Tiongkok.

Tata Kelola Sosial

Tata kelola sosial, atau lebih tepatnya dalam bahasa Mandarin disebut 'konstruksi sosial', adalah area ketiga di mana AI dipromosikan sebagai peluang strategis bagi Tiongkok. Bersamaan dengan perlambatan ekonomi, Tiongkok menghadapi tantangan sosial yang muncul, yang menghambat upayanya untuk menjadi 'masyarakat yang cukup sejahtera'. Populasi yang menua dan keterbatasan lingkungan serta sumber daya lainnya merupakan contoh nyata yang diberikan dalam AIDP tentang permasalahan sosial yang dihadapi Tiongkok. Dengan demikian, AIDP menguraikan tujuan penggunaan AI dalam berbagai layanan publik untuk membuat tata kelola layanan sosial lebih presisi dan, dengan demikian, memitigasi tantangan tersebut serta meningkatkan taraf hidup masyarakat.

Tiongkok telah mengalami beberapa perubahan struktural tercepat dibandingkan negara mana pun dalam 40 tahun terakhir. Tiongkok telah bergeser dari ekonomi terencana menjadi ekonomi pasar dan dari masyarakat pedesaan menjadi masyarakat perkotaan. Perubahan-perubahan ini telah membantu memfasilitasi pembangunan ekonomi, tetapi juga menimbulkan sejumlah masalah sosial. Salah satu tantangan sosial paling mendesak yang dihadapi Tiongkok adalah tidak adanya sistem kesejahteraan yang mapan. Di bawah ekonomi terencana, pekerja dijamin mendapatkan tunjangan dari lahir hingga meninggal, termasuk jaminan kerja dan tunjangan kesejahteraan, yang disediakan melalui perusahaan negara setempat atau kolektif pedesaan. Pergerakan Tiongkok menuju ekonomi pasar sosialis sejak tahun 1990-an telah mempercepat pergeseran ketentuan ini dari perusahaan dan kolektif lokal ke lembaga negara dan masyarakat. Dalam praktiknya, Tiongkok kesulitan mengembangkan program pensiun dan asuransi kesehatan yang matang, sehingga menciptakan kesenjangan dalam jaring pengaman sosial. Meskipun beberapa inisiatif telah diperkenalkan untuk mengatasi masalah ini, negara tersebut kesulitan untuk menerapkannya.

Degradasi lingkungan yang serius yang terjadi selama pembangunan pesat Tiongkok merupakan elemen lain yang memprihatinkan. Selama sebagian besar periode pembangunan Tiongkok, fokusnya adalah pada pertumbuhan ekonomi, dengan sedikit atau tanpa insentif yang diberikan untuk perlindungan lingkungan. Akibatnya, eksternalitas negatif yang signifikan dan beberapa bencana alam akibat manusia telah terjadi yang terbukti merugikan masyarakat. Salah satu yang paling menonjol adalah kualitas udara yang sangat buruk, yang telah dikaitkan dengan peningkatan risiko penyakit dan sekarang menjadi penyebab kematian keempat di Tiongkok. Di samping itu, 40% sungai di Tiongkok tercemar oleh industri, menyebabkan 60.000 kematian dini per tahun. Degradasi lingkungan sebesar ini merusak kesehatan penduduk, menurunkan kualitas hidup, dan semakin membebani infrastruktur kesejahteraan yang ada.

Sentralitas dari kekhawatiran ini dapat dilihat pada Kongres Partai Nasional ke-19 pada tahun 2017, di mana Presiden Xi menyatakan bahwa 'kontradiksi utama' di Tiongkok telah berubah. Meskipun 'kontradiksi' sebelumnya berfokus pada 'kebutuhan material dan budaya rakyat yang terus meningkat dan produksi sosial yang terbelakang,' Xi menyatakan 'apa yang

kita hadapi sekarang adalah kontradiksi antara pembangunan yang tidak seimbang dan tidak memadai dengan kebutuhan rakyat yang terus meningkat akan kehidupan yang lebih baik'. Setelah bertahun-tahun berfokus pada pertumbuhan ekonomi yang tak terkendali, pernyataan Presiden Xi menekankan pergeseran yang lebih luas dalam pendekatan Tiongkok dalam menangani konsekuensi liberalisasi ekonomi.

Pernyataan-pernyataan ini tercermin dalam beberapa rencana pemerintah, termasuk Prakarsa Dewan Negara, 'Tiongkok Sehat 2030', yang berupaya merombak sistem perawatan kesehatan. Tren serupa dapat dilihat dalam upaya Tiongkok untuk membersihkan lingkungannya, dengan rencana tiga tahun baru yang dibangun berdasarkan inisiatif-inisiatif relevan sebelumnya. Tiongkok baru-baru ini berfokus pada AI sebagai cara untuk mengatasi masalah-masalah ini dan meningkatkan kesejahteraan warga negara. Telah ditunjukkan bahwa strategi pembangunan utama Tiongkok bergantung pada solusi yang didorong oleh data besar. Misalnya, 'Tiongkok Sehat 2030' secara eksplisit menekankan pentingnya teknologi dalam mencapai strategi reformasi perawatan kesehatan Tiongkok, dan menekankan peralihan dari pengobatan ke pencegahan, dengan pengembangan AI sebagai sarana untuk mencapai tujuan. Pendekatan ini juga membentuk perlindungan lingkungan, di mana Presiden Xi telah mempromosikan 'perlindungan lingkungan digital' (数字环保). Dalam hal ini, AI digunakan untuk memprediksi dan memitigasi tingkat polusi udara, serta untuk meningkatkan pengelolaan dan pemilahan sampah ("Pengelolaan Sampah Bertenaga AI Sedang Berlangsung di Tiongkok" 2019).

Administrasi peradilan merupakan bidang lain yang telah dikembangkan pemerintah Tiongkok dengan menggunakan AI untuk meningkatkan tata kelola sosial. Di bawah Xi Jinping, terdapat tujuan eksplisit untuk memprofesionalkan sistem hukum, yang masih terkendala oleh kurangnya transparansi, isu proteksionisme lokal, dan campur tangan pejabat lokal dalam kasus pengadilan. Berbagai reformasi telah diperkenalkan dalam upaya untuk membatasi praktik-praktik ini, termasuk pengalihan tanggung jawab pengelolaan pengadilan lokal dari pemerintah daerah ke pemerintah provinsi, pembuatan basis data tempat hakim dapat melaporkan upaya campur tangan politisi lokal, dan sistem registrasi kasus yang mempersulit pengadilan untuk menolak kasus yang kompleks atau kontroversial. Yang menarik, ketika berfokus pada AI, adalah *Beberapa Pendapat Mahkamah Rakyat Agung tentang Peningkatan Sistem Akuntabilitas Yudisial* (2015), yang mengharuskan hakim untuk merujuk kasus serupa dalam penalaran yudisial mereka.

Lebih lanjut, hal itu menetapkan bahwa keputusan yang bertentangan dengan kasus serupa sebelumnya harus memicu mekanisme pengawasan dengan hakim yang lebih senior. Untuk membantu hakim meminimalkan inkonsistensi, upaya telah dilakukan untuk memperkenalkan teknologi AI yang memfasilitasi pembuatan 'putusan serupa dalam kasus serupa'. Dalam hal teknologi, dua jenis sistem menyeluruh telah muncul. Yang pertama adalah 'sistem dorong kasus serupa', di mana AI digunakan untuk mengidentifikasi putusan dari kasus serupa dan memberikan hakim ini untuk referensi. Jenis sistem ini telah diperkenalkan oleh, antara lain, Pengadilan Tinggi Rakyat Hainan yang juga telah mendorong penggunaan sistem

AI di pengadilan tingkat bawah di seluruh provinsi. Teknologi kedua adalah 'peringatan penilaian abnormal' yang akan mendeteksi jika suatu putusan berbeda dari kasus-kasus serupa, dan jika demikian, akan memperingatkan atasan hakim, yang mendorong intervensi. Penerimaan terhadap penggunaan teknologi ini beragam, dengan masyarakat yang menerima prospek berkurangnya korupsi sementara hakim mengapresiasi berkurangnya beban kerja. Namun, beberapa ahli teori hukum mengkritik dampak tidak manusiawi dari penggunaan teknologi dalam penjatuhan hukuman dan kerugian yang dapat ditimbulkannya terhadap 'hermeneutika hukum, teknik penalaran hukum, pelatihan profesional, dan kepribadian etis hakim'.

Ke depannya, fokus pada penggunaan AI di Tiongkok dalam tata kelola tampaknya kemungkinan besar akan berpusat pada Sistem 'Kredit Sosial' yang banyak diberitakan, yang didasarkan pada pengembangan perangkat yang diperlukan untuk mengatasi masalah sosial Tiongkok yang mendesak. Untuk melakukan ini, sistem ini secara luas bertujuan untuk meningkatkan kapasitas tata kelola negara, mempromosikan kredibilitas lembaga negara, dan membangun basis kredit keuangan yang layak. Saat ini, Sistem Kredit Sosial bukanlah satu sistem nasional yang terpadu tetapi terdiri dari daftar hitam nasional yang menyusun data dari berbagai lembaga pemerintah, sistem kredit sosial individual yang dijalankan oleh pemerintah daerah, dan inisiatif perusahaan swasta. Sistem ini terpecah-pecah dan, dalam banyak kasus, uji coba lokal kurang canggih secara teknis, dengan beberapa versi hanya mengandalkan kertas dan pena. Meskipun demikian, target ambisius dari Sistem Kredit Sosial memberikan contoh yang menarik tentang niat pemerintah untuk mengandalkan teknologi digital, untuk tata kelola sosial dan juga untuk regulasi perilaku warganya yang lebih rinci.

Tata Kelola Moral

Tata kelola/konstruksi sosial di Tiongkok tidak hanya mencakup aspek material dan lingkungan, tetapi juga perilaku warga negara. Para ahli berpendapat bahwa disrupsi periode Maois yang diikuti oleh 'keterbukaan' telah menciptakan kekosongan moral di Tiongkok. Kekhawatiran ini juga dirasakan oleh masyarakat Tiongkok, dengan Ipsos Mori menemukan bahwa kekhawatiran atas 'kemerosotan moral' di Tiongkok dua kali lebih tinggi daripada rata-rata global.

Hal ini telah diakui oleh pemerintah Tiongkok, dengan para pejabat tinggi, termasuk Presiden Xi, yang mengemukakan gagasan tentang 'standar moral minimum' dalam masyarakat. Tujuan ini tidak terbatas pada memastikan tata kelola yang 'baik' dalam pengertian tradisional; tetapi juga mencakup pengaturan perilaku warga negara dan peningkatan integritas moral mereka, yang dianggap sebagai tugas dalam kewenangan pemerintah. Dari sudut pandang pemerintah, AI dapat digunakan untuk tujuan ini.

AIDP menyoroti potensi AI untuk memahami kognisi dan psikologi kelompok (2017). Niat untuk mengandalkan AI dalam tata kelola moral dapat dilihat dalam undang-undang selanjutnya, dengan contoh paling jelas mungkin adalah 'Garis Besar Pembentukan Sistem Kredit Sosial' Dewan Negara, yang dirilis pada tahun 2014. Dokumen ini menggarisbawahi bahwa Sistem Kredit Sosial tidak hanya bertujuan untuk mengatur tindakan keuangan dan korporasi bisnis dan warga negara, tetapi juga perilaku sosial individu. Dokumen ini

menguraikan beberapa tantangan sosial yang ingin diatasi oleh rencana tersebut, termasuk penghindaran pajak, kekhawatiran keamanan pangan, dan ketidakjujuran akademis. Sebagaimana disoroti, upaya saat ini untuk menerapkan sistem ini masih belum konsisten, namun beberapa di antaranya telah memasukkan unsur moral, seperti mempermalukan debitur bermasalah di depan umum.

Contoh konkret lebih lanjut tentang bagaimana Tiongkok telah memanfaatkan AI dalam tata kelola sosial dapat dilihat di bidang keamanan dalam negeri dan kepolisian. Tiongkok telah menjadi yang terdepan dalam pengembangan kota pintar, dengan sekitar separuh kota pintar dunia berada di Tiongkok. Mayoritas sumber daya yang dicurahkan untuk mengembangkan kota-kota ini berfokus pada teknologi pengawasan, seperti pengenalan wajah dan komputasi awan untuk kepolisian biasa. Penggunaan program pengawasan 'kontraterorisme' yang canggih di wilayah otonom Xinjiang memberikan bukti yang lebih jelas dan lebih bermasalah tentang upaya pemerintah untuk menggunakan AI untuk pengawasan internal. Teknologi ini tidak terbatas pada pengenalan wajah, tetapi juga mencakup aplikasi telepon seluler untuk melacak penduduk Uighur setempat, yang digambarkan oleh pemerintah sebagai calon pembangkang atau teroris.

Ketika pernyataan pemerintah dibaca secara paralel dengan perkembangan ini, tampaknya beberapa bentuk Sistem Kredit Sosial akan memainkan peran sentral dalam masa depan tata kelola Tiongkok yang didukung AI, yang menempatkan hak-hak warga negara di bawah tekanan devaluasi yang tajam. Misalnya, sebagian besar warga negara menghasilkan jejak data yang besar, dan hampir semua transaksi sehari-hari di kota-kota bersifat non-tunai dan dilakukan dengan aplikasi seluler, penyedia internet memberlakukan 'pendaftaran nama asli', yang menghubungkan semua aktivitas daring dengan individu, yang memungkinkan pemerintah untuk mengidentifikasi dan memiliki akses ke profil digital semua warga negara yang menggunakan layanan internet seluler.

Risiko yang signifikan dan mungkin terkait dengan penerapan AI untuk tata kelola berasal dari jalinan aspek material tata kelola sosial dengan pengawasan dan kontrol moral. Artikel-artikel di media Barat sering menekankan sifat bermasalah dari 'Sistem Kredit Sosial', karena nada otoriter dari kontrol yang meluas ini. Contoh ketidakpuasan publik dengan fitur-fitur spesifik dari sistem kredit sosial yang dikelola secara lokal tampaknya mendukung sudut pandang ini.

Dalam beberapa kasus, bahkan ada kasus reaksi publik yang menyebabkan revisi dalam kriteria pemeringkatan untuk sistem kredit sosial lokal. Sebaliknya, beberapa komentator menekankan bahwa, di dalam negeri, sistem kredit sosial nasional mungkin diterima secara positif sebagai respons terhadap persepsi kemerosotan moral di Tiongkok, dan keinginan yang menyertainya untuk membangun kepercayaan yang lebih besar; bahkan, telah disarankan bahwa sistem tersebut mungkin lebih tepat disebut sistem 'Kepercayaan Sosial'. Ketika melihat hukuman yang dijatuhkan oleh sistem kredit sosial, beberapa tindakan, termasuk memasukkan warga negara yang berperilaku buruk di kereta api ke dalam daftar hitam untuk bepergian, telah menerima respons positif di media sosial Tiongkok. Sensor pemerintah dan

efek mengerikan dapat menjadi alasan dukungan ini, tetapi saat ini tidak ada bukti sensor yang secara khusus menargetkan unggahan tentang sistem kredit sosial.

Upaya juga telah dilakukan untuk memahami opini publik tentang sistem secara keseluruhan, alih-alih hanya kontroversi spesifik atau kasus-kasus daftar hitam. Sebuah survei nasional oleh seorang akademisi Barat tentang sistem kredit sosial Tiongkok menemukan tingkat persetujuan yang tinggi di antara masyarakat. Meskipun demikian, permasalahan dalam metodologi makalah ini, khususnya penerjemahan kata 'Sistem Kredit Sosial', menunjukkan bahwa mungkin lebih tepat untuk mempertimbangkan kurangnya kesadaran secara umum, daripada sentimen dukungan yang meluas ("Beyond Black Mirror" 2019). Poin-poin ini menunjukkan bahwa masih terlalu dini untuk mengukur sentimen publik di Tiongkok seputar pengembangan Sistem Kredit Sosial.

Penting untuk menyadari bahwa meskipun penerapan Sistem Kredit Sosial saat ini relatif monoton, ke depannya, risiko dan tantangan etika yang substansial tetap ada terkait kriteria untuk dimasukkan ke dalam daftar hitam atau menerima skor rendah, serta pengecualian yang dapat ditimbulkannya. Dalam hal yang pertama, daftar hitam nasional terdiri dari mereka yang telah melanggar hukum dan peraturan yang berlaku, dengan alasan yang jelas untuk pencantumannya. Namun, dokumen hukum yang menjadi dasar daftar ini seringkali tidak terdefinisi dengan baik dan berfungsi dalam sistem hukum yang berada di bawah Partai Komunis Tiongkok. Akibatnya, undang-undang, seperti yang melarang penyebaran informasi yang secara serius mengganggu ketertiban sosial, dapat digunakan untuk menghukum individu atas tindakan yang secara politis tidak diinginkan, termasuk kebebasan berbicara. Namun, lebih tepat untuk menganggap ini sebagai masalah struktur politik-hukum dan bukan sistem kredit sosial itu sendiri. Isu etika fundamental tentang pendekatan pengawasan yang tidak dapat diterima masih belum tertangani.

Sehubungan dengan sistem berbasis skor lokal yang tidak semata-mata bergantung pada ilegalitas, kriteria penilaian bisa jadi lebih samar lagi. Misalnya, skor kredit sosial di Fuzhou memperhitungkan 'kekuatan pekerjaan', yang didasarkan pada 'pekerja keras/teliti dan teliti' yang didefinisikan secara longgar. Ini bermasalah secara etika karena standar inklusi yang tidak transparan dan sewenang-wenang yang diperkenalkan untuk memberikan orang manfaat tertentu. Bersamaan dengan inklusi adalah pengecualian yang dapat disebabkan oleh sistem ini. Saat ini, sebagian besar sistem kredit sosial dikendalikan oleh entitas yang terpisah dan tidak terhubung satu sama lain, membatasi hukuman yang berlebihan. Meskipun demikian, nota kesepahaman muncul antara sistem kredit sosial dan perusahaan swasta untuk mengecualikan mereka yang masuk daftar hitam dari kegiatan seperti terbang (Arsène 2019). Akibatnya, penting untuk menekankan bahwa sementara Sistem Kredit Sosial masih berkembang, kriteria inklusi dan potensi pengecualian yang ditimbulkan menimbulkan pertanyaan etika yang serius.

4.4 PERDEBATAN TENTANG ETIKA DIGITAL DAN AI DI TIONGKOK

Selain menetapkan tujuan material, AIDP menguraikan keinginan khusus bagi Tiongkok untuk menjadi pemimpin dunia dalam mendefinisikan norma dan standar etika untuk AI.

Setelah AIDP dirilis, pemerintah, badan publik, dan industri di Tiongkok relatif lambat dalam mengembangkan kerangka kerja etika AI. Namun, belakangan ini terdapat lonjakan upaya untuk mendefinisikan prinsip-prinsip etika. Pada Maret 2019, Kementerian Sains dan Teknologi Tiongkok membentuk Komite Ahli Tata Kelola Kecerdasan Buatan Generasi Baru Nasional. Pada Juni 2019, badan ini merilis delapan prinsip tata kelola AI. Prinsip-prinsip tersebut menekankan bahwa, di atas segalanya, pengembangan AI harus dimulai dari peningkatan kesejahteraan bersama umat manusia. Penghormatan terhadap hak asasi manusia, privasi, dan keadilan juga digarisbawahi dalam prinsip-prinsip tersebut. Terakhir, prinsip-prinsip tersebut menyoroti pentingnya transparansi, tanggung jawab, kolaborasi, dan kelincahan untuk menghadapi risiko baru dan yang muncul.

Sejalan dengan publikasi ini, Administrasi Standardisasi Republik Rakyat Tiongkok, badan tingkat nasional yang bertanggung jawab untuk mengembangkan standar teknis, merilis buku putih tentang standar AI. Buku putih ini berisi diskusi tentang isu-isu keselamatan dan etika yang terkait dengan teknologi tersebut. Tiga prinsip utama untuk menetapkan persyaratan etika teknologi AI diuraikan. Pertama, prinsip *kepentingan manusia* menyatakan bahwa tujuan akhir AI adalah untuk memberikan manfaat bagi kesejahteraan manusia. Kedua, prinsip tanggung jawab menekankan perlunya menetapkan *akuntabilitas* sebagai persyaratan untuk pengembangan dan penerapan sistem dan solusi AI. *Transparansi* juga termasuk dalam prinsip ini, yang mendukung persyaratan untuk memahami prinsip-prinsip operasional sistem AI. Ketiga, prinsip *konsistensi hak dan tanggung jawab* menekankan bahwa, di satu sisi, data harus dicatat dengan benar dan pengawasan dilakukan, tetapi di sisi lain, entitas komersial harus dapat melindungi kekayaan intelektual mereka.

Badan-badan afiliasi pemerintah dan perusahaan swasta juga telah mengembangkan prinsip-prinsip etika AI mereka sendiri. Misalnya, Akademi Kecerdasan Buatan Beijing, sebuah badan penelitian dan pengembangan yang mencakup perusahaan-perusahaan terkemuka Tiongkok dan universitas-universitas Beijing, didirikan pada November 2018. Badan ini kemudian merilis 'Prinsip-Prinsip AI Beijing' yang harus diikuti untuk penelitian dan pengembangan, penggunaan, dan tata kelola AI. Mirip dengan prinsip-prinsip yang diajukan oleh Komite Ahli AIDP, Prinsip-Prinsip Beijing berfokus pada berbuat baik bagi kemanusiaan, menggunakan AI 'dengan benar', dan memiliki pandangan ke depan untuk memprediksi dan beradaptasi dengan ancaman di masa depan. Di sektor swasta, kerangka etika yang paling terkenal datang dari CEO Tencent, Pony Ma. Kerangka kerja ini menekankan pentingnya AI yang tersedia, andal, mudah dipahami, dan terkendali. Akhirnya, Asosiasi Kecerdasan Buatan Tiongkok (CAII) belum menetapkan prinsip-prinsip etika, tetapi telah membentuk komite etika AI pada pertengahan tahun 2018 dengan tujuan ini.

Prinsip-prinsip yang disebutkan di atas memiliki beberapa kesamaan dengan prinsip-prinsip yang didukung di negara-negara maju, namun perbedaan institusional dan budaya berarti bahwa hasilnya kemungkinan akan sangat berbeda. Etika AI Tiongkok perlu dipahami dalam konteks budaya, ideologi, dan opini publik negara tersebut. Meskipun analisis komparatif yang lengkap berada di luar cakupan artikel ini, dapat diantisipasi, misalnya, bahwa prinsip-prinsip yang muncul dari Tiongkok lebih menekankan tanggung jawab sosial dan

hubungan kelompok dan komunitas, dengan fokus yang relatif lebih sedikit pada hak-hak individualistik, sehingga menggemakan diskusi sebelumnya tentang etika Konfusianisme di media sosial.

Pada bagian-bagian berikut, kami akan berfokus pada perdebatan tentang etika AI sebagaimana munculnya kaitannya dengan privasi dan etika medis, karena keduanya merupakan dua area yang paling matang di mana seseorang dapat memahami secara lebih umum 'pendekatan Tiongkok' saat ini terhadap etika digital. Analisis kedua area ini tidak dimaksudkan untuk memberikan peta lengkap dari semua perdebatan tentang masalah etika terkait AI di Tiongkok. Sebaliknya, hal ini dapat berfungsi untuk menyoroti beberapa isu kontroversial yang muncul, dan memberikan pemahaman yang lebih luas tentang jenis batasan yang mungkin dibuat di Tiongkok ketika agenda normatif di negara tersebut ditetapkan.

Privasi

Semua prinsip AI etis yang diuraikan di atas menyebutkan pentingnya melindungi privasi. Namun, terdapat perdebatan sengit di Tiongkok mengenai jenis data apa saja yang harus dilindungi. Tiongkok secara historis memiliki peraturan perlindungan data yang lemah—yang memungkinkan pengumpulan dan pembagian informasi pribadi dalam jumlah besar oleh aktor publik dan swasta—dan minimnya perlindungan terhadap privasi individu. Pada tahun 2018, Robin Li, salah satu pendiri Baidu, menyatakan bahwa "masyarakat Tiongkok lebih terbuka atau kurang sensitif terhadap masalah privasi. Jika mereka mampu menukar privasi dengan kenyamanan, keamanan, dan efisiensi, dalam banyak kasus, mereka bersedia melakukannya".

Sudut pandang ini—yang sesuai dengan tanggapan yang tampaknya tidak terlalu negatif terhadap Sistem Kredit Sosial—telah menyebabkan beberapa komentator Barat salah menafsirkan persepsi publik tentang privasi dalam evaluasi mereka terhadap strategi AI Tiongkok. Namun, pemahaman Li tentang privasi bukanlah pemahaman yang diterima secara luas, dan pernyataannya memicu reaksi keras di media sosial Tiongkok. Kekhawatiran terhadap privasi ini mencerminkan data survei dari Internet Society of China, dengan 54% menyatakan bahwa mereka menganggap masalah pelanggaran data pribadi sebagai 'parah'. Ketika mempertimbangkan beberapa kasus penyalahgunaan data, angka ini tidak mengejutkan. Misalnya, survei Asosiasi Konsumen Tiongkok mengungkapkan bahwa 85% orang telah mengalami kebocoran data dalam beberapa bentuk. Dengan demikian, bertentangan dengan apa yang mungkin disimpulkan dari beberapa pernyataan penting, ada sentimen kekhawatiran umum dalam masyarakat Tiongkok atas penyalahgunaan informasi pribadi.

Menanggapi kekhawatiran serius ini, Tiongkok telah menerapkan langkah-langkah perlindungan privasi, yang menyebabkan seorang komentator menyebut negara itu sebagai 'pemimpin kejutan Asia dalam perlindungan data'. Inti dari upaya ini adalah Spesifikasi Keamanan Informasi Pribadi (Spesifikasi), sebuah standar privasi yang dirilis pada Mei 2018. Standar ini dimaksudkan untuk menguraikan aturan privasi yang lebih luas, yang ditetapkan dalam Undang-Undang Keamanan Siber 2017. Secara khusus, standar ini berfokus pada perlindungan data pribadi dan memastikan bahwa orang-orang diberdayakan untuk

mengendalikan informasi mereka sendiri. Sejumlah ketentuan dalam standar tersebut sangat mencakup semuanya, termasuk definisi luas tentang informasi pribadi yang sensitif, yang mencakup fitur-fitur seperti kerusakan reputasi. Bahasa yang digunakan dalam standar tersebut membuat seorang komentator berpendapat bahwa beberapa ketentuan lebih memberatkan daripada ketentuan GDPR.

Terlepas dari bukti sebelumnya, sifat standar tersebut berarti bahwa standar tersebut tidak benar-benar sebanding dengan GDPR. Di satu sisi, alih-alih menjadi bagian dari undang-undang yang dapat ditegakkan secara formal, Spesifikasi tersebut hanyalah standar nasional 'sukarela' yang dibuat oleh Komite Teknis Standardisasi Keamanan Informasi Nasional Tiongkok (TC260). Atas dasar inilah salah satu penyusun menyatakan bahwa standar ini tidak sebanding dengan GDPR, karena hanya dimaksudkan sebagai pendamping panduan untuk undang-undang perlindungan data sebelumnya, seperti undang-undang Keamanan Siber 2017. Di sisi lain, masih ada ketegangan yang sulit untuk diselesaikan karena, meskipun benar bahwa standar hanya bersifat sukarela, standar di Tiongkok memiliki pengaruh substantif untuk menegakkan tujuan kebijakan pemerintah, juga melalui skema sertifikasi. Dengan demikian, pada bulan Juni 2018, standar sertifikasi untuk langkah-langkah privasi ditetapkan, dengan perusahaan-perusahaan seperti Alipay dan Tencent Cloud menerima sertifikasi. Lebih lanjut, Spesifikasi tersebut menetapkan kekhususan Undang-Undang Keamanan Siber yang dapat ditegakkan, dengan Baidu dan AliPay keduanya terpaksa merombak kebijakan data mereka karena tidak 'mematuhi semangat Standar Keamanan Informasi Pribadi'.

Pada kenyataannya, kelemahan dalam undang-undang privasi Tiongkok lebih disebabkan oleh banyaknya celah di dalamnya, kelemahan sistem peradilan Tiongkok, dan kekuatan pemerintah yang berpengaruh, yang seringkali merupakan otoritas terakhir, yang tidak dimintai pertanggungjawaban melalui mekanisme demokrasi. Secara khusus, terdapat pengecualian yang signifikan dan bermasalah untuk pengumpulan dan penggunaan data, termasuk yang berkaitan dengan keamanan, kesehatan, atau 'kepentingan publik yang signifikan' yang samar dan dapat ditafsirkan secara fleksibel. Celah-celah besar inilah yang paling mengungkapkan kebijakan data Tiongkok. Dapat dikatakan bahwa beberapa perlindungan konsumen yang luas memang ada, tetapi sebenarnya hal ini tidak mencakup pemerintah. Dengan demikian, kekuatan perlindungan privasi kemungkinan besar ditentukan oleh keputusan pemerintah terkait pengumpulan dan penggunaan data, alih-alih kendala hukum dan praktis. Hal ini mengkhawatirkan.

Penting untuk menyadari bahwa GDPR Uni Eropa memuat dasar 'kepentingan publik' yang serupa untuk memproses data pribadi secara sah jika persetujuan atau anonimisasi tidak praktis, dan bahwa ketentuan-ketentuan ini tidak didefinisikan dengan baik dalam undang-undang dan sering diabaikan dalam praktik. Namun, perbedaan krusial dan mencolok antara contoh Tiongkok dan Uni Eropa berkaitan dengan sistem hukum yang mendasari kedua pendekatan tersebut. Cabang yudikatif Uni Eropa memiliki pengaruh substantif, termasuk kapasitas untuk menafsirkan undang-undang dan menggunakan mekanisme peninjauan kembali yudisial untuk menentukan keabsahan undang-undang secara lebih luas. Sebaliknya, menurut *sistem hukum* Tiongkok, peradilan tunduk pada pengawasan dan campur tangan dari

legislatif, yang secara *de jure* memiliki supremasi legislatif; hal ini memberikan *kendali de facto* kepada Partai. Dengan demikian, kekuatan perlindungan privasi di Tiongkok mungkin dan seringkali ditentukan oleh keputusan pemerintah terkait pengumpulan dan penggunaan data, alih-alih oleh batasan hukum dan praktis. Sebagaimana telah diutarakan, "Fungsi hukum dalam mengatur masyarakat telah diakui sejak tahun 2002, tetapi belum dianggap esensial bagi PKT. Sebaliknya, moralitas dan opini publik secara bersamaan berfungsi sebagai dua alternatif hukum untuk tujuan pemerintahan. Akibatnya, badan administratif dapat mengabaikan hukum atas dasar kebijakan partai, moralitas, opini publik, atau pertimbangan politik lainnya.

Jika dikaitkan kembali dengan kebijakan AI, Tiongkok telah diuntungkan dari melimpahnya data yang secara historis difasilitasi oleh perlindungan privasi yang lemah. Di permukaan, undang-undang privasi Tiongkok tampaknya bertentangan dengan komitmen pembangunan lainnya, seperti Sistem Kredit Sosial, yang mewajibkan data pribadi yang ekstensif. Situasi ini menciptakan ekosistem ganda di mana pemerintah semakin bersedia mengumpulkan data dalam jumlah besar, tanpa menghormati privasi, sekaligus menegur perusahaan teknologi atas langkah-langkah yang mereka terapkan. Perlu diingat bahwa perusahaan swasta, seperti Tim Nasional AI, diandalkan untuk tata kelola di tingkat nasional dan lokal, dan oleh karena itu dapat menerima dukungan diam-diam alih-alih teguran dalam kasus-kasus di mana kepentingan pemerintah dilayani secara langsung. Akibatnya, 'strategi privasi' di Tiongkok tampaknya bertujuan untuk melindungi privasi jenis *konsumen* tertentu, alih-alih privasi *warga negara* secara keseluruhan, yang memungkinkan pemerintah untuk mengumpulkan data pribadi di mana pun dan kapan pun yang mungkin hanya berguna (bahkan tidak sepenuhnya diperlukan) untuk kebijakannya.

Dari perspektif internal, dapat dikatakan bahwa, jika dilihat dengan latar belakang tingkat kepercayaan yang tinggi terhadap pemerintah dan seringnya kebocoran serta penyalahgunaan oleh sektor swasta, kompromi ini tampak lebih mudah dipahami oleh masyarakat Tiongkok. Spesifikasi ini memiliki ruang lingkup yang substansial untuk merevisi dualitas ini, dengan sejumlah celah telah ditutup sejak rilis awal, tetapi tampaknya kecil kemungkinan perlindungan privasi dari intervensi pemerintah akan dikodifikasi dalam waktu dekat. Masalah etika ini masih belum terselesaikan.

Etika Medis

Etika medis merupakan bidang penting lain yang terdampak oleh pendekatan Tiongkok terhadap etika AI. Prinsip Panduan Kesehatan Nasional Tiongkok telah menjadi pusat pengembangan strategis dan tata kelola sistem layanan kesehatan nasionalnya selama 60 tahun terakhir. Prinsip-prinsip tersebut telah ditulis ulang beberapa kali, seiring dengan transisi sistem layanan kesehatan dari sistem satu tingkat sebelum tahun 1978 menjadi sistem dua tingkat yang diperkuat oleh reformasi layanan kesehatan pada tahun 2009.

Prinsip-Prinsip Panduan terakhir kali ditulis ulang pada tahun 1996 dan prinsip-prinsip berikut masih berlaku:

- (a) Masyarakat di daerah pedesaan merupakan prioritas utama
- (b) Pencegahan penyakit harus diutamakan
- (c) Pengobatan tradisional Tiongkok dan pengobatan Barat harus bekerja sama

- (d) Urusan kesehatan harus bergantung pada sains dan pendidikan
- (e) Masyarakat secara keseluruhan harus dimobilisasi untuk berpartisipasi dalam urusan kesehatan, sehingga berkontribusi pada kesehatan masyarakat dan pembangunan negara secara keseluruhan.

Kelima prinsip tersebut relevan untuk memahami sistem layanan kesehatan Tiongkok secara keseluruhan, tetapi dari perspektif analisis etika penggunaan AI di Tiongkok dalam ranah medis, prinsip (a), (b), dan (e) adalah yang paling penting. Prinsip-prinsip tersebut menyoroti bahwa—berbeda dengan Barat, di mana data layanan kesehatan elektronik sebagian besar berfokus pada kesehatan individu, sehingga teknik AI dianggap krusial untuk mewujudkan 'pengobatan yang dipersonalisasi'—di Tiongkok, layanan kesehatan sebagian besar berfokus pada kesehatan penduduk. Dalam konteks ini, ambisi utama AI adalah membebaskan data untuk tujuan kesehatan masyarakat. Hal ini terbukti dari AIDP, yang menguraikan ambisi untuk menggunakan AI guna 'memperkuat pemantauan, pencegahan, dan pengendalian intelijen epidemi', dan untuk 'mencapai terobosan dalam analisis data besar, Internet of Things, dan teknologi kunci lainnya' dengan tujuan memperkuat manajemen kesehatan cerdas komunitas. Aspek yang sama bahkan lebih jelas dalam pemberitahuan resmi Dewan Negara tahun 2016 tentang pengembangan dan penggunaan data besar di sektor perawatan kesehatan, yang secara eksplisit menyatakan bahwa kumpulan data besar kesehatan dan medis adalah sumber daya nasional, dan bahwa pengembangannya harus dilihat sebagai prioritas nasional untuk meningkatkan kesehatan bangsa.

Dari perspektif analisis etika, promosi data perawatan kesehatan sebagai barang publik di seluruh kebijakan publik—termasuk dokumen seperti *Measures on Population Health Information dan Guiding Opinions on Promoting and Regulating the Application of Big Medical and Health Data*—sangat penting. Pendekatan ini, dikombinasikan dengan aturan yang longgar tentang berbagi data di Tiongkok, dan dorongan berbagi data publik secara terbuka antar badan pemerintah, mempromosikan pengumpulan dan agregasi data kesehatan tanpa perlu persetujuan individu, dengan menempatkan *kemurahan hati* kelompok di atas otonomi individu. Hal ini paling baik diilustrasikan dengan sebuah contoh. Sebagai bagian dari rencana 'Made in 2025' Tiongkok, 130 perusahaan, termasuk 'WeDoctor' (didukung oleh Tencent, salah satu juara nasional AI Tiongkok) menandatangani perjanjian kerja sama dengan pemerintah daerah untuk menyediakan pemeriksaan kesehatan yang terdiri dari tekanan darah, elektrokardiogram (EKG), tes urine dan darah, gratis bagi warga pedesaan (Hawkins 2019). Data yang dihasilkan oleh tes-tes ini secara otomatis (yaitu tanpa persetujuan dari individu) ditautkan ke nomor identifikasi pribadi dan kemudian diunggah ke cloud WeDoctor, tempat data tersebut digunakan untuk melatih produk-produk AI WeDoctor.

Produk-produk ini termasuk 'sistem perawatan tambahan untuk praktik umum', yang digunakan oleh dokter desa untuk memberikan diagnosis dan perawatan yang disarankan dari basis data yang berisi lebih dari 5000 gejala dan 2000 penyakit. Dapat dikatakan, sifat sensitif data tersebut dapat membuat "perusahaan—dan regulator—waspada terhadap pencatatan di luar negeri, yang akan memerlukan pengungkapan dan pengawasan yang lebih besar" (Lucas 2019). Meskipun hal ini, dan praktik serupa lainnya, memang melibatkan anonimisasi,

praktik tersebut sangat kontras dengan pendekatan Eropa dan AS terhadap penggunaan data medis, yang memprioritaskan otonomi dan privasi individu, alih-alih kesejahteraan sosial. Keseimbangan yang adil antara kebutuhan individu dan masyarakat sangat penting untuk pendekatan etis terhadap data pribadi, tetapi terdapat asimetri di mana penekanan berlebihan pada pendekatan individualistik dapat dengan mudah diperbaiki dengan konsensus individu, sedangkan pendekatan yang murni bersifat sosial tetap tidak etis sejauh pendekatan tersebut terlalu mudah mengesampingkan hak-hak individu dan tidak dapat diperbaiki dengan mudah.

Kesejahteraan masyarakat mungkin berakhir dengan membenarkan pengorbanan hak individu sebagai sarana. Hal ini tetap tidak etis. Namun, bagaimana hal ini dipersepsikan di Tiongkok masih menjadi pertanyaan yang lebih terbuka. Perlu diingat bahwa Tiongkok memiliki penyediaan layanan kesehatan primer yang sangat buruk, bahwa Tiongkok baru mencapai 95% cakupan kesehatan (melalui skema asuransi nasional) pada tahun 2015, Tiongkok memiliki sekitar 1,8 dokter per 1000 warga negara dibandingkan dengan rata-rata OECD sebesar 3,4, dan didasarkan pada nilai-nilai Konfusianisme yang mempromosikan kesetaraan di tingkat kelompok.

Dalam konteks inilah prinsip etika 'kewajiban penyelamatan yang mudah' dapat ditafsirkan secara lebih mendalam. Prinsip ini menetapkan bahwa, jika suatu tindakan dapat bermanfaat bagi orang lain dan menimbulkan sedikit ancaman bagi individu, maka pilihan etisnya adalah menyelesaikan tindakan tersebut. Dalam hal ini, dari perspektif Tiongkok, seseorang dapat berargumen bahwa berbagi data layanan kesehatan mungkin tidak menimbulkan ancaman langsung bagi individu, terutama karena *Pasal 6 Peraturan tentang Pengelolaan Rekam Medis Institusi Medis, Pasal 8 Peraturan Manajemen tentang Penerapan Rekam Medis Elektronik, Pasal 6 Langkah-Langkah Pengelolaan Informasi Kesehatan, Undang-Undang Keamanan Siber Republik Rakyat Tiongkok, dan Spesifikasi Keamanan Informasi Pribadi yang baru*, semuanya memberikan instruksi spesifik dan terperinci untuk memastikan keamanan dan kerahasiaan data. Namun, hal ini berpotensi memberikan manfaat yang signifikan bagi masyarakat luas.

'Penafsiran dari dalam' sebelumnya tidak menyiratkan bahwa pendekatan Tiongkok terhadap penggunaan AI dalam layanan kesehatan dapat diterima atau tidak menimbulkan masalah etika. Justru sebaliknya. Secara khusus, pendekatan Tiongkok dirusak oleh setidaknya tiga risiko utama. Pertama, terdapat risiko menciptakan pasar untuk perawatan manusia. Sistem medis dua tingkat Tiongkok menyediakan perawatan yang diasuransikan negara untuk semua, dan opsi bagi individu untuk membayar secara pribadi untuk perawatan yang lebih cepat atau berkualitas lebih tinggi. Hal ini sejalan dengan pemikiran Konfusianisme, yang mendorong penggunaan sumber daya pribadi untuk menguntungkan diri sendiri dan keluarga.

Dengan diperkenalkannya 'klinik satu menit' tanpa awak dari Ping An Good Doctor di seluruh Tiongkok (yang sekarang mungkin sudah ada hingga 1000 klinik), pasien dapat datang langsung, memberikan gejala dan riwayat medis, serta menerima diagnosis dan rencana perawatan otomatis (yang hanya ditindaklanjuti dengan saran klinis manusia untuk pelanggan baru), sangat mungkin untuk membayangkan skenario di mana hanya mereka yang mampu

membayar yang dapat mengakses dokter manusia. Dalam bidang di mana perawatan emosional, dan keterlibatan dalam pengambilan keputusan, seringkali sama pentingnya dengan deduksi logis dari sebuah 'diagnosis', hal ini dapat berdampak negatif yang signifikan terhadap tingkat dan kualitas perawatan yang diakses di seluruh populasi dan terhadap integritas diri, setidaknya bagi mereka yang tidak mampu membayar perawatan manusia.

Kedua, dalam konteks populasi yang masih berkembang pesat namun juga menua, Tiongkok berinvestasi secara signifikan dalam informatisasi sosial perawatan kesehatan dan, setidaknya sejak 2015, telah menghubungkan data emosional dan perilaku yang diekstrapolasi dari media sosial dan data perawatan kesehatan harian (yang dihasilkan dari perangkat yang dapat ditelan, yang dapat ditanamkan, perangkat yang dapat dikenakan, carebot, dan perangkat Internet of Things) ke Catatan Kesehatan Elektronik, dengan tujuan memungkinkan perawatan komunitas bagi para lansia.

Hal ini semakin memperkuat budaya pengawasan massal yang dijalankan oleh negara di Tiongkok dan, di era Sistem Kredit Sosial, menunjukkan bahwa teknologi yang sama yang dirancang untuk memungkinkan orang tetap mandiri di masyarakat seiring bertambahnya usia, suatu hari nanti dapat digunakan sebagai alat kontrol sosial (*"Tiongkok Sedang Membangun Surga Kesehatan Digital Terbaik. Atau Bukan?"* 2019), untuk mengurangi kejadian 'penyakit sosial' seperti obesitas dan diabetes tipe II dengan kedok 'meningkatkan kehidupan masyarakat' melalui penggunaan AI untuk meningkatkan tata kelola layanan sosial (sebagaimana dinyatakan dalam AIDP).

Risiko etika ketiga dikaitkan dengan modifikasi gen CRISPR dan AI. CRISPR adalah teknik modifikasi gen kontroversial yang dapat digunakan untuk mengubah presentasi gen dalam organisme hidup, misalnya untuk tujuan menyembuhkan atau mencegah penyakit genetik. Ini terkait erat dengan AI, karena teknik Pembelajaran Mesin dapat digunakan untuk mengidentifikasi gen atau gen mana yang perlu diubah dengan metode CRISPR. Kontroversi, dan potensi masalah etika yang signifikan, yang terkait dengan penelitian di bidang ini terkait dengan fakta bahwa tidak selalu mungkin untuk mengetahui di mana garis antara kebutuhan klinis yang tidak terpenuhi dan peningkatan manusia atau kontrol genetik.

Ini menjadi jelas ketika, pada November 2018, peneliti biofisika He Jiankui mengungkapkan bahwa ia telah berhasil memodifikasi bayi secara genetik menggunakan metode CRISPR untuk membatasi peluang mereka tertular HIV. Pengumuman itu disambut oleh kecaman internasional dan eksperimen He dikutuk oleh pemerintah Tiongkok pada saat itu. Namun, dorongan untuk dilihat sebagai pemimpin dunia dalam perawatan medis, dikombinasikan dengan janji yang ditawarkan pengeditan gen untuk pengobatan penyakit, menunjukkan bahwa respons yang berbeda mungkin terjadi di masa depan.

Perubahan kebijakan pemerintah seperti itu sangat mungkin terjadi karena persaingan global di bidang ini memanas. AS telah mengumumkan bahwa mereka sedang mendaftarkan pasien dalam uji coba untuk menyembuhkan bentuk kebutaan yang diwariskan; dan Inggris telah meluncurkan tantangan Accelerating Detection of Disease untuk membuat kohort pasien beranggotakan lima juta orang yang datanya akan digunakan untuk mengembangkan pendekatan AI baru untuk diagnosis dini dan penemuan biomarker (*"Accelerating Detection of*

Disease" 2019). Pengumuman ini menciptakan insentif yang kuat bagi para peneliti di Tiongkok untuk mendorong batasan regulasi guna mencapai keberhasilan yang cepat. Patut dicatat, Tiongkok juga telah mengajukan jumlah paten terbesar untuk rekayasa genetika pada hewan di dunia. Pemantauan ketat sangat penting untuk mencegah pelanggaran etika lebih lanjut.

4.5 KESIMPULAN

Dalam artikel ini, kami menganalisis sifat kebijakan AI di Tiongkok dan konteks kemunculannya, dengan memetakan inisiatif kebijakan utama tingkat nasional yang menunjukkan niat untuk memanfaatkan AI. Kami mengidentifikasi tiga area yang sangat relevan: *daya saing internasional, pertumbuhan ekonomi, dan tata kelola sosial (konstruksi)*. Pengembangan dan penerapan AI di masing-masing area ini memiliki implikasi bagi Tiongkok dan komunitas internasional. Misalnya, meskipun kebijakan 'kartu truf' untuk mendapatkan keuntungan militer mungkin bukan hal baru, penerapannya pada teknologi AI berisiko memicu perlombaan senjata dan merusak stabilitas internasional.

Upaya untuk melawan tren ini tampaknya sebagian besar sia-sia. Analisis kami menunjukkan bahwa Tiongkok memiliki beberapa peluang terbesar untuk mendapatkan manfaat ekonomi di bidang-bidang seperti otomatisasi, dan bahwa negara tersebut sedang mendorong kemajuan substansial di bidang-bidang terkait AI. Meskipun demikian, upaya untuk meredam gangguan yang muncul dari penggunaan AI di industri saat ini masih kurang. Dengan demikian, AI dapat membantu mendorong peningkatan produktivitas dan tingkat pertumbuhan yang tinggi, tetapi penggunaannya kemungkinan akan memperparah ketimpangan yang ada dalam masyarakat dan bahkan mengurangi dukungan terhadap pemerintah dan kebijakannya. AIDP juga mempromosikan AI sebagai cara untuk membantu mengatasi beberapa masalah sosial utama, mulai dari polusi hingga standar hidup. Namun, dampak positif di bidang ini tampaknya datang dengan peningkatan kendali atas perilaku individu, dengan tata kelola yang meluas ke ranah perilaku moral dan erosi privasi lebih lanjut.

Etika juga memainkan peran sentral dalam upaya kebijakan Tiongkok terkait AI. AIDP menguraikan niat yang jelas untuk mendefinisikan norma dan standar etika, namun upaya untuk melakukannya masih dalam tahap awal, karena secara umum terbatas pada prinsip-prinsip tingkat tinggi, dan kurang implementasi. Analisis terhadap pendekatan Tiongkok yang ada dan perdebatan yang muncul di bidang privasi dan etika medis memberikan wawasan tentang jenis kerangka kerja yang mungkin muncul. Terkait privasi, di permukaan, perlindungan yang baru-baru ini diperkenalkan mungkin tampak kuat, dengan definisi informasi pribadi yang bahkan lebih luas daripada yang digunakan dalam GDPR. Namun, jika ditelusuri lebih lanjut, terungkap banyak celah dan pengecualian yang memungkinkan pemerintah (dan perusahaan-perusahaan yang secara implisit didukung oleh pemerintah) untuk menghindari perlindungan privasi dan isu-isu mendasar terkait kurangnya akuntabilitas dan kekuasaan pengambilan keputusan pemerintah yang tak terbatas terkait pengawasan massal.

Sejalan dengan itu, ketika berfokus pada etika medis, jelas bahwa, meskipun Tiongkok mungkin sependapat dengan Barat mengenai prinsip-prinsip bioetika, fokusnya pada

kesehatan populasi, berbeda dengan fokus Barat pada kesehatan individu, dapat dengan mudah mengarah pada hasil yang tidak etis (pengorbanan yang dibebankan pada satu orang demi kepentingan banyak orang) dan menciptakan sejumlah risiko, seiring AI merambah ranah medis. Risiko-risiko ini kemungkinan akan berkembang seiring waktu, tetapi risiko perawatan yang tidak setara antara mereka yang mampu membayar dokter manusia dan mereka yang tidak mampu, pengendalian penyakit sosial, dan penelitian medis yang tidak etis saat ini merupakan yang paling signifikan.

Tiongkok merupakan aktor sentral dalam debat internasional tentang pengembangan dan tata kelola AI. Penting untuk *memahami* kebutuhan internal Tiongkok, ambisi di arena internasional, dan masalah etika, yang semuanya membentuk pengembangan kebijakan AI Tiongkok. Penting juga untuk memahami semua ini tidak hanya secara eksternal, dari perspektif Barat, tetapi juga secara internal, dari perspektif Tiongkok.

Namun, beberapa perlindungan etika, kendala, dan desiderata bersifat universal dan diterima serta dihargai secara universal, seperti sifat dan ruang lingkup hak asasi manusia. Mereka memungkinkan seseorang untuk *mengevaluasi*, setelah *memahami*, pendekatan Tiongkok terhadap pengembangan AI. Inilah sebabnya mengapa dalam artikel ini kami berusaha berkontribusi pada analisis yang lebih komprehensif dan bernuansa tentang faktor-faktor struktural, budaya, dan politik yang mendasari sikap Tiongkok terhadap AI, serta indikasi kemungkinan lintasannya, sambil juga menyoroti di mana masalah etika tetap ada, muncul, atau kemungkinan akan diperburuk. Mereka harus ditangani sedini mungkin secara kontekstual.

BAB 5

MENERJEMAHKAN PRINSIP KE DALAM PRAKTIK ETIKA DIGITAL

Penerapan prinsip-prinsip etika dalam pengembangan dan penggunaan teknologi digital merupakan tantangan yang kompleks. Meskipun berbagai upaya telah dilakukan untuk mengintegrasikan nilai-nilai etis ke dalam praktik teknologi, kenyataannya, bahkan inisiatif terbaik pun dapat menghadapi risiko ketidakpatuhan terhadap etika. Dalam konteks perdebatan internasional mengenai etika digital, terdapat lima bentuk risiko yang telah diidentifikasi atau dapat diperkirakan terjadi, yaitu: belanja etika (*ethics shopping*), pencucian etika (*ethics washing*), lobi etika (*ethics lobbying*), pembuangan etika (*ethics dumping*), dan penghindaran etika (*ethics shirking*). Bab ini akan membahas secara kritis kelima risiko tersebut serta dampaknya terhadap upaya mewujudkan tata kelola teknologi digital yang etis dan bertanggung jawab.

5.1 PENDAHULUAN

Membutuhkan waktu yang sangat lama, tetapi saat ini, perdebatan tentang dampak dan implikasi etika teknologi digital telah mencapai halaman depan surat kabar. Hal ini dapat dimengerti: teknologi digital dari layanan berbasis web hingga solusi Kecerdasan Buatan (AI) semakin memengaruhi kehidupan sehari-hari miliaran orang, sehingga terdapat banyak harapan sekaligus kekhawatiran tentang desain, pengembangan, dan penerapannya.

Setelah lebih dari setengah abad penelitian akademis, reaksi publik baru-baru ini adalah maraknya inisiatif untuk menetapkan prinsip, pedoman, kode, atau kerangka kerja apa yang dapat memandu inovasi digital secara etis, khususnya dalam AI, untuk memberi manfaat bagi umat manusia dan seluruh lingkungan. Ini merupakan perkembangan positif yang menunjukkan kesadaran akan pentingnya topik ini dan minat untuk menanganinya secara sistematis.

Namun, sudah saatnya perdebatan beralih dari *apa* menjadi *bagaimana*: bukan hanya etika apa yang dibutuhkan, tetapi juga bagaimana etika dapat diterapkan dan diimplementasikan secara efektif dan berhasil untuk menghasilkan dampak positif. Misalnya, *Pedoman Etika untuk AI yang Dapat Dipercaya* menetapkan tolok ukur tentang apa yang mungkin atau mungkin tidak memenuhi syarat sebagai AI yang baik secara etis di Uni Eropa. Publikasinya saat ini sedang diikuti oleh upaya praktis berupa pengujian, penerapan, dan implementasi

Peralihan dari bab pertama yang lebih teoretis tentang *apa*, ke bab kedua yang lebih praktis tentang *bagaimana*, bisa dibilang, adalah wajar dan patut dipuji. Namun, dalam menerjemahkan prinsip ke dalam praktik, upaya terbaik sekalipun dapat dirusak oleh beberapa risiko yang tidak etis.

5.2 BELANJA ETIKA

Sejumlah besar prinsip, kode, pedoman, atau kerangka kerja etika telah diusulkan selama beberapa tahun terakhir. Saat ini terdapat lebih dari 70 rekomendasi, yang diterbitkan dalam 2 tahun terakhir, khususnya tentang etika. Maraknya dokumen ini menimbulkan inkonsistensi dan kebingungan di antara para pemangku kepentingan mengenai dokumen mana yang lebih disukai. Hal ini juga memberi tekanan pada aktor swasta dan publik yang merancang, mengembangkan, atau menerapkan solusi digital untuk membuat deklarasi mereka sendiri karena takut dianggap tertinggal, sehingga semakin berkontribusi pada redundansi informasi.

Dalam kasus ini, risiko utama yang tidak etis adalah bahwa semua hiperaktivitas ini menciptakan "pasar prinsip dan nilai", di mana aktor swasta dan publik dapat mencari jenis etika yang paling sesuai untuk membenarkan perilaku mereka saat ini, alih-alih merevisi perilaku mereka agar konsisten dengan kerangka etika yang diterima secara sosial. Berikut definisi yang lebih ringkas:

Belanja etika digital =_{def.} Malpraktik dalam memilih, mengadaptasi, atau merevisi ("mencampur dan mencocokkan") prinsip, pedoman, kode, kerangka kerja, atau standar etika serupa lainnya (terutama tetapi tidak hanya dalam etika AI), dari berbagai penawaran yang tersedia, untuk menyesuaikan beberapa perilaku yang sudah ada sebelumnya (pilihan, proses, strategi, dll.), dan dengan demikian membenarkannya secara a posteriori, alih-alih menerapkan atau meningkatkan perilaku baru dengan membandingkannya dengan standar etika publik.

Memang, dalam meta-analisis terbaru, kami menunjukkan bahwa sebagian besar keragaman "di pasar etika" tampak jelas dan lebih disebabkan oleh pilihan kata dan kosakata daripada konten yang sebenarnya. Namun, potensi risiko "mencampuradukkan" daftar prinsip etika yang disukai tetap nyata, karena kelonggaran semantik dan redundansi memungkinkan relativisme interpretatif. Ethics shopping kemudian menyebabkan ketidaksesuaian standar (misalnya, sulit untuk memahami apakah dua perusahaan mengikuti prinsip etika yang sama dalam mengembangkan solusi AI), dan dengan demikian mengurangi peluang perbandingan, persaingan, dan akuntabilitas.

Strategi untuk menangani ethics shopping digital adalah dengan menetapkan standar etika yang jelas, bersama, dan diterima publik. Inilah sebabnya saya baru-baru ini berpendapat bahwa penerbitan *Pedoman Etika untuk AI yang Dapat Dipercaya* adalah sebuah peningkatan yang signifikan, mengingat bahwa itu adalah hal yang paling dekat yang tersedia di Uni Eropa (UE) dengan standar yang komprehensif, berwibawa, dan publik tentang apa yang dapat dianggap sebagai AI yang baik secara sosial. Sekarang *Pedoman* tersebut tersedia, malpraktik belanja etika digital seharusnya setidaknya lebih jelas jika tidak lebih sulit untuk dilakukan, karena siapa pun di UE dapat berlangganan kepada mereka, daripada berbelanja untuk (atau bahkan memasak) "etika" mereka sendiri.

5.3 ETIKA BLUEWASHING

Dalam etika lingkungan, *greenwashing* adalah malpraktik yang dilakukan oleh aktor swasta atau publik yang berusaha tampil lebih hijau, lebih berkelanjutan, atau lebih ramah lingkungan daripada yang sebenarnya. Yang saya maksud dengan "etika bluewashing" adalah versi digital dari *greenwashing*. Karena tidak ada warna khusus yang dikaitkan dengan praktik-praktik baik secara etis dalam teknologi digital, "biru" dapat berfungsi untuk mengingatkan kita bahwa kita tidak sedang membicarakan keberlanjutan ekologis, melainkan sekadar kosmetik etika digital:

Etika bluewashing =_{def.} malpraktik dalam membuat klaim yang tidak berdasar atau menyesatkan tentang, atau menerapkan langkah-langkah yang dangkal demi, nilai-nilai dan manfaat etika dari proses, produk, layanan, atau solusi digital lainnya agar tampak lebih etis secara digital daripada yang sebenarnya.

Greenwashing dan bluewashing dalam etika merupakan bentuk misinformasi, yang seringkali dicapai dengan menghabiskan sebagian kecil sumber daya yang dibutuhkan untuk mengatasi masalah etika yang mereka pura-pura atasi. Keduanya berfokus pada pemasaran, periklanan, atau aktivitas hubungan masyarakat lainnya (misalnya, sponsor), termasuk pembentukan kelompok penasihat yang mungkin tidak berdaya atau kurang kritis.

Kedua praktik buruk ini menggoda karena, dalam setiap kasus, tujuannya beragam dan semuanya kompatibel:

- (a) Mengalihkan perhatian penerima pesan biasanya publik, tetapi pemegang saham atau pemangku kepentingan mana pun dapat menjadi target dari segala hal yang salah, yang bisa diperbaiki, atau yang tidak terjadi tetapi seharusnya terjadi;
- (b) Menutupi dan tidak mengubah perilaku apa pun yang seharusnya diperbaiki;
- (c) Mencapai penghematan ekonomi; dan
- (d) Memperoleh keuntungan, misalnya kompetitif atau sosial, misalnya dalam hal "itikad baik".

Namun, berbeda dengan *greenwashing*, *bluewashing* dapat lebih mudah dipadukan dengan belanja etika digital: aktor swasta atau publik mencari prinsip-prinsip yang paling sesuai dengan praktiknya saat ini, mempublikasikannya seluas-luasnya, lalu melanjutkan dengan memoles inovasi teknologinya tanpa perbaikan nyata, biaya yang jauh lebih rendah, dan beberapa potensi manfaat sosial. Saat ini, *bluewashing* etika sangat menggoda dalam konteks AI, di mana isu-isu etika banyak, biaya untuk melakukan hal yang benar mungkin tinggi, dan ketidakpastian normatif atau terkadang kebingungan tersebar luas.

Strategi terbaik melawan *bluewashing* sama dengan yang telah diadopsi untuk melawan *greenwashing*: *transparansi* dan *edukasi*. Transparansi publik, akuntabel, dan berbasis bukti tentang praktik baik dan klaim etika harus menjadi prioritas bagi para aktor yang ingin menghindari kesan terlibat dalam malpraktik *bluewashing*. Pendidikan publik dan faktual, di sisi target *bluewashing* apa pun—tidak hanya masyarakat umum tetapi juga anggota dewan eksekutif dan dewan penasihat, misalnya—tentang apakah dan praktik etika efektif apa yang

benar-benar diterapkan berarti bahwa para aktor mungkin lebih kecil kemungkinannya (tergoda untuk) mengalihkan perhatian publik dari tantangan etika yang mereka hadapi.

Sebagaimana yang kami rekomendasikan dalam Floridi dkk. (2018), pengembangan metrik untuk kepercayaan produk dan layanan AI (dan solusi digital secara umum) akan memungkinkan perbandingan yang digerakkan oleh pengguna dari semua penawaran yang dipasarkan dan memfasilitasi deteksi bluewashing belaka, meningkatkan pemahaman publik, dan menumbuhkan daya saing seputar pengembangan produk dan layanan yang lebih aman, lebih bermanfaat secara sosial dan lingkungan. Dalam jangka panjang, sistem sertifikasi untuk produk dan layanan digital dapat mencapai apa yang telah dicapai oleh solusi serupa lainnya dalam etika lingkungan: menjadikan bluewashing sama terlihat dan memalukannya dengan greenwashing.

5.4 LOBI ETIKA

Terkadang, aktor swasta mencoba menggunakan regulasi mandiri terkait etika AI untuk melobi agar tidak diberlakukannya norma hukum, atau agar norma tersebut diperlemah atau diperlemah penegakannya, atau untuk memberikan dalih bagi kepatuhan yang terbatas. Malapraktik spesifik ini memengaruhi banyak sektor, tetapi tampaknya lebih mungkin terjadi di sektor digital, di mana etika dapat dieksploitasi seolah-olah merupakan alternatif legislasi, dan atas nama inovasi teknologi serta dampak positifnya terhadap pertumbuhan ekonomi, sebuah penalaran yang tidak mudah didukung dalam konteks lingkungan atau biomedis. Berikut definisi yang lebih umum:

Lobi etika digital =_{def.} malpraktik yang mengeksploitasi etika digital untuk menunda, merevisi, mengganti, atau menghindari legislasi yang baik dan diperlukan (atau penegakannya) terkait perancangan, pengembangan, dan penerapan proses, produk, layanan, atau solusi digital lainnya. Seseorang mungkin berpendapat bahwa lobi etika digital adalah strategi yang buruk, kemungkinan besar akan gagal dalam jangka panjang karena paling banter hanya berpandangan pendek: cepat atau lambat legislasi akan menyusul.

Terlepas dari apakah argumen ini meyakinkan atau tidak, lobi etika digital sebagai taktik jangka pendek tetap dapat menyebabkan banyak kerusakan, dengan menunda pengesahan legislasi yang diperlukan, dengan membantu bermanuver menghindari atau menghindari interpretasi yang lebih menuntut dari legislasi yang berlaku, sehingga mempermudah kepatuhan tetapi juga tidak selaras dengan semangat hukum, atau dengan memengaruhi pembuat undang-undang untuk mengesahkan undang-undang yang lebih menguntungkan pelobi daripada yang diharapkan. Lebih jauh lagi, dan yang sangat penting, malpraktik, atau kecurigaan akan malpraktik, berisiko merusak nilai setiap regulasi etika digital secara *menyeluruh*.

Kerusakan tambahan ini sangat disesalkan karena regulasi diri merupakan salah satu alat berharga utama yang tersedia untuk pembuatan kebijakan. Dengan sendirinya, ia tidak

dapat menggantikan hukum, tetapi jika diterapkan dengan benar, ia dapat menjadi pelengkap yang krusial, ketika:

- Legislasi tidak tersedia (misalnya, dalam eksperimen tentang produk realitas ditambah) atau
- Legislasi tersedia, tetapi juga membutuhkan interpretasi etis (misalnya, dalam hal memahami hak atas penjelasan dalam GDPR) atau
- Legislasi tersedia, tetapi juga membutuhkan penyeimbang etis:
- Jika lebih baik tidak melakukan sesuatu, meskipun (belum) ilegal untuk melakukannya (misalnya, untuk mengotomatiskan sepenuhnya suatu prosedur medis tanpa pengawasan manusia) atau
- Jika lebih baik melakukan sesuatu, meskipun (belum) diwajibkan secara hukum (misalnya, untuk menerapkan kondisi pasar tenaga kerja yang lebih baik dalam Ekonomi Gig).

Strategi melawan lobi etika digital ada dua. Di satu sisi, hal itu harus dilawan dengan legislasi yang baik dan penegakan hukum yang efektif. Hal ini lebih mudah jika aktor lobi (swasta maupun publik) kurang berpengaruh terhadap pembuat undang-undang atau ketika opini publik dapat memberikan tekanan etika yang tepat.

Di sisi lain, lobi etika digital harus diungkap kapan pun terjadi dan dibedakan secara jelas dari bentuk-bentuk regulasi diri yang sesungguhnya. Hal ini dapat terjadi dengan lebih kredibel jika proses tersebut juga merupakan bagian dari kode etik regulasi diri dari seluruh sektor industri, dalam kasus kami, industri teknologi digital, yang memiliki kepentingan lebih umum dalam menjaga konteks yang sehat di mana regulasi diri yang sesungguhnya diterima secara sosial dan efektif, sementara lobi etika diungkap sebagai sesuatu yang tidak dapat diterima.

5.5 PEMBUANGAN ETIKA

"*Pembuangan etika*" adalah ungkapan yang dicetuskan pada tahun 2013 oleh Komisi Eropa untuk menggambarkan ekspor praktik penelitian yang tidak etis ke negara-negara dengan kerangka hukum dan etika serta mekanisme penegakan yang lebih lemah (atau lebih longgar, atau mungkin hanya berbeda, dalam hal etika digital). Hal ini berlaku untuk semua jenis penelitian termasuk penelitian dalam ilmu komputer, ilmu data, pembelajaran mesin, robotika, dan jenis AI lainnya tetapi dampaknya paling serius dalam konteks kesehatan dan biologis.

Etika biomedis dan lingkungan dapat dianggap universal dan global; terdapat perjanjian dan kerangka kerja internasional serta lembaga internasional yang memantau penerapan atau penegakannya, sehingga "pembuangan etika" dapat dilawan secara lebih efektif dan koheren ketika penelitian melibatkan konteks biomedis dan ekologi. Namun, dalam konteks digital, keragaman (atau bahkan ketiadaan) rezim hukum dan kerangka etika memfasilitasi ekspor (apa yang dianggap dalam konteks asli tempat "dumper" beroperasi) praktik yang tidak etis (atau bahkan ilegal), dan impor hasil praktik tersebut. Dengan kata lain, masalahnya ada dua, tentang *etika penelitian* dan *etika konsumsi*. Jadi, berikut definisinya:

Dumping etika digital = *def.* malpraktik (a) mengekspor aktivitas penelitian tentang proses, produk, layanan, atau solusi digital lainnya, dalam konteks atau tempat lain (misalnya oleh organisasi Eropa di luar Uni Eropa) dengan cara yang tidak dapat diterima secara etis dalam konteks atau tempat asal dan (b) mengimpor hasil aktivitas penelitian yang tidak etis tersebut. Baik (a) maupun (b) sama-sama penting. Sebagai contoh konkret, meskipun jauh, bukanlah hal yang aneh bagi negara-negara untuk melarang budidaya organisme hasil rekayasa genetika, tetapi mengizinkan impornya.

Asimetri perlakuan etis (dan hukum) antara suatu praktik (penelitian yang tidak etis dan/atau ilegal) dan keluarannya (konsumsi keluaran penelitian yang tidak etis dan dapat diterima secara etis dan hukum) berarti bahwa pembuangan etika dapat memengaruhi etika digital tidak hanya dalam hal ekspor kegiatan penelitian yang tidak etis, tetapi juga dalam hal impor hasil kegiatan tersebut secara tidak etis.

Misalnya, sebuah perusahaan dapat mengekspor penelitiannya dan kemudian merancang, mengembangkan, dan melatih algoritma, misalnya untuk pengenalan wajah, pada data pribadi lokal di negara non-Uni Eropa dengan kerangka etika dan hukum yang berbeda atau lebih lemah untuk perlindungan data pribadi, yang akan dianggap tidak etis dan ilegal di Uni Eropa karena GDPR. Setelah dilatih, algoritma tersebut kemudian dapat diimpor ke Uni Eropa dan diterapkan tanpa dikenakan penalti apa pun atau bahkan tidak disetujui. Sementara langkah pertama (a) mungkin lebih mudah diblokir, setidaknya dalam hal etika penelitian; langkah kedua (b), yang melibatkan konsumsi hasil penelitian yang tidak etis, lebih samar, kurang terlihat bermasalah, dan karenanya lebih sulit untuk dipantau dan dibatasi.

Sayangnya, kemungkinan besar, dalam waktu dekat, masalah pembuangan etika digital akan menjadi semakin serius, karena dampak mendalam teknologi digital terhadap kesehatan dan perawatan sosial serta pertahanan, kepolisian, dan keamanan, kemudahan portabilitas globalnya, kompleksitas proses produksi (beberapa tahapannya mungkin melibatkan pembuangan etika), dan kepentingan ekonomi yang sangat besar. Misalnya, khususnya dalam AI, di mana Uni Eropa merupakan importir bersih solusi dari AS dan Tiongkok, aktor swasta dan publik berisiko tidak hanya mengekspor praktik yang tidak etis tetapi juga (dan secara independen) mengimpor solusi yang mungkin telah dikembangkan dengan cara yang tidak dapat diterima secara etis di Uni Eropa. Dalam hal ini pun, strateginya ada dua. Kita harus berkonsentrasi pada etika penelitian *dan* etika konsumsi. Jika kita ingin koheren, keduanya perlu mendapat perhatian yang setara.

Dalam hal etika penelitian, sedikit lebih mudah untuk melakukan kontrol di sumbernya, melalui manajemen etis pendanaan publik untuk penelitian. Dalam hal ini, Uni Eropa berada di posisi terdepan. Namun, masih ada masalah signifikan yaitu sebagian besar penelitian dan pengembangan tentang solusi digital dilakukan oleh sektor swasta, yang pendanaannya mungkin tidak terlalu dibatasi oleh batas-batas geografis (aktor swasta dapat dengan mudah merelokasi penelitian dan pengembangannya ke tempat yang secara etika kurang menuntut,

variasi geografis dari belanja etika yang terlihat di Bagian 5.2) dan tidak diteliti secara etis dengan cara yang sama seperti penelitian yang didanai publik.

Dalam hal etika konsumsi, khususnya produk dan layanan digital, banyak yang dapat dilakukan, baik melalui pembentukan sistem sertifikasi untuk produk dan layanan yang dapat menginformasikan pengadaan, maupun penggunaan publik dan swasta. Sebagaimana dalam kasus bluewashing, asal-usul sistem dan solusi digital yang andal dan dapat diterima secara etis harus memainkan peran yang semakin penting di tahun-tahun mendatang jika kita ingin menghindari kemunafikan karena berhati-hati dalam etika penelitian dalam konteks digital, namun tetap bersikap hati-hati terhadap penggunaan hasil penelitian yang tidak etis.

5.6 PENGABAIAAN ETIKA

Para ahli etika sangat akrab dengan praktik lama yang menerapkan standar ganda dalam evaluasi moral. Dengan menerapkan pendekatan yang lunak dan ketat, seseorang dapat mengevaluasi dan memperlakukan agen (atau tindakan mereka, atau konsekuensi dari tindakan mereka) secara berbeda dari agen yang serupa (tindakan atau konsekuensi), padahal sebenarnya mereka semua seharusnya diperlakukan sama. Biasanya, risiko standar ganda didasarkan, bahkan secara tidak sengaja, pada bias, ketidakadilan, atau kepentingan pribadi. Risiko yang ingin saya soroti di sini berasal dari keluarga yang sama, tetapi memiliki asal-usul yang berbeda. Untuk menyoroti kekhususannya, saya akan meminjam istilah "pengabaian etika" dari sektor keuangan⁷ dan mendefinisikannya sebagai berikut:

Pengabaian etika = _{def.} praktik yang kurang etis dalam melakukan "pekerjaan yang kurang etis" (seperti memenuhi tugas, menghormati hak, dan menjunjung tinggi komitmen) dalam konteks tertentu, semakin rendah imbal hasil dari pekerjaan etis tersebut dalam konteks tersebut yang secara keliru dipersepsikan.

Pengabaian etika, seperti halnya pembuangan etika, memiliki akar sejarah dan sering kali mengikuti garis geopolitik. Aktor lebih cenderung terlibat dalam pembuangan etika dan pengabaian dalam konteks di mana populasi yang kurang beruntung, lembaga yang lebih lemah, ketidakpastian hukum, rezim yang korup, distribusi kekuasaan yang tidak adil, dan penyakit ekonomi, hukum, politik, atau sosial lainnya berlaku. Bukan hal yang aneh untuk memetakan, dengan benar, kedua malpraktik di sepanjang jurang antara Global Utara dan Global Selatan, atau untuk melihat keduanya terutama memengaruhi Negara Berpenghasilan Rendah dan Menengah. Masa lalu kolonial masih memainkan peran yang memalukan. Penting juga untuk diingat bahwa, dalam konteks digital, malpraktik ini dapat memengaruhi segmen populasi di Global Utara. Ekonomi Gig dapat dilihat sebagai kasus pengabaian etika di negara-negara maju.

Dan pengembangan mobil self-driving dapat diartikan sebagai contoh pembuangan penelitian di beberapa negara bagian AS. Dalam hal ini, Konvensi Wina 1968 tentang Lalu Lintas Jalan, yang menetapkan prinsip-prinsip internasional untuk mengatur undang-undang lalu lintas, mewajibkan pengemudi untuk selalu memegang kendali penuh dan bertanggung

jawab atas perilaku kendaraannya di jalan raya. Namun, Amerika Serikat bukanlah negara penandatangan dan persyaratan tersebut tidak berlaku, yang berarti undang-undang kendaraan negara bagian tidak melarang kendaraan otomatis, dan beberapa negara bagian telah memberlakukan undang-undang untuk kendaraan otomatis. Inilah juga mengapa penelitian tentang mobil self-driving sebagian besar terjadi di Amerika Serikat begitu pula insiden dan penderitaan manusia yang terkait.

Strategi melawan penghindaran etika terdiri dari menangani asal-usulnya, yaitu kurangnya alokasi tanggung jawab yang jelas. Agen mungkin lebih tergoda untuk menghindari pekerjaan etis mereka dalam konteks tertentu semakin mereka (berpikir mereka) dapat memindahkan tanggung jawab ke tempat lain. Hal ini lebih mungkin dan mudah terjadi dalam "konteks D", di mana tanggung jawab seseorang dapat dianggap (secara keliru) lebih rendah karena *jauh, berkurang, didelegasikan, atau didistribusikan*. Dengan demikian, penghindaran etika merupakan biaya tidak etis yang harus ditanggung oleh suatu agensi akibat deresponsabilisasi.

Asal muasal inilah yang menjadikannya kasus khusus dari masalah etika standar ganda. Inilah sebabnya mengapa diperlukan lebih banyak keadilan dan lebih sedikit bias sejauh penghindaran etika merupakan kasus khusus dari masalah standar ganda tetapi keduanya juga tidak cukup untuk menghilangkan insentif untuk melakukan penghindaran etika. Untuk memberantas malpraktik semacam itu, diperlukan pula etika tanggung jawab terdistribusi yang merelokasi tanggung jawab dan dengan demikian, pujian dan celaan, penghargaan dan hukuman, dan pada akhirnya akuntabilitas kausal dan tanggung jawab hukum ke tempat yang semestinya.

5.7 KESIMPULAN

Saya harap artikel singkat ini dapat menjadi peta bagi mereka yang ingin menghindari atau meminimalkan beberapa risiko etika yang paling jelas dan signifikan, ketika bernavigasi dari prinsip ke praktik dalam etika digital. Dari perspektif Sokrates, malpraktik seringkali merupakan akibat dari solusi yang salah perhitungan atau peluang yang keliru. Memahami sedini mungkin bahwa jalan pintas, penundaan, atau perbaikan cepat tidak mengarah pada solusi etika yang lebih baik, melainkan pada masalah yang lebih serius, yang semakin sulit dipecahkan setelah ditangani, tidak menjamin bahwa kelima malpraktik yang dianalisis dalam artikel ini akan hilang, tetapi berarti bahwa malpraktik tersebut akan berkurang sejauh memang didasarkan pada kesalahpahaman dan kesalahan penilaian. Ketidaktahuan akan hal yang lebih baik adalah sumber dari banyak kejahatan. Jadi, solusinya seringkali adalah informasi yang lebih banyak dan lebih baik untuk semua.

BAB 6

AI DAPAT MENJADI KEKUATAN UNTUK KEBAIKAN

Kecerdasan buatan (AI), yang dipahami sebagai bentuk agen otonom yang mampu belajar secara mandiri, membawa potensi besar untuk memberikan manfaat bagi manusia. Namun, karakteristik AI yang unik ini juga menimbulkan tantangan etika yang tidak dapat diabaikan. Penggunaan AI dalam berbagai konteks telah memunculkan sejumlah masalah, seperti diskriminasi yang tidak semestinya, kurangnya transparansi atau penjelasan atas keputusan yang diambil AI, kesenjangan dalam pertanggungjawaban, serta pengaruh AI terhadap kemampuan individu dalam menentukan nasibnya sendiri.

Bab ini mengkaji bagaimana potensi kebaikan AI dapat dimanfaatkan secara etis, sembari meminimalkan dampak negatif yang mungkin ditimbulkan. Fokus utama diberikan pada peran analisis etika dalam mendampingi proses desain, pengembangan, dan implementasi AI. Dibahas pula pentingnya penerjemahan pedoman etika ke dalam praktik yang konkret dan operasional. Dalam konteks inilah etika digital diposisikan sebagai bentuk etika translasional yakni, upaya untuk menjembatani prinsip-prinsip etis dengan praktik inovasi teknologi yang nyata dan bertanggung jawab.

Kecerdasan buatan (AI) bukan sekadar teknologi baru yang membutuhkan regulasi. AI merupakan kekuatan dahsyat yang sedang membentuk kembali praktik sehari-hari, interaksi pribadi dan profesional, serta lingkungan. Demi kesejahteraan umat manusia, sangatlah penting bahwa kekuatan ini digunakan sebagai kekuatan kebaikan. Etika memainkan peran kunci dalam proses ini dengan memastikan bahwa regulasi AI memanfaatkan potensinya sekaligus memitigasi risikonya.

AI dapat didefinisikan dengan berbagai cara. Jika definisinya salah, penilaian apa pun terhadap tantangan etika AI akan menjadi fiksi ilmiah atau bahkan pengalihan perhatian yang tidak bertanggung jawab, seperti dalam kasus perdebatan singularitas. Pendekatan yang ilmiah dan tepat adalah dengan mengacu pada definisi klasiknya sebagai sumber daya yang terus berkembang dari agensi yang interaktif, otonom, dan belajar mandiri, yang memungkinkan artefak komputasi untuk melakukan tugas-tugas yang jika tidak demikian akan membutuhkan kecerdasan manusia untuk dieksekusi dengan sukses. AI kemudian dapat didefinisikan lebih lanjut dalam hal fitur-fitur seperti model komputasi yang diandalkannya atau arsitektur teknologinya.

Namun, ketika menyangkut isu-isu etika dan kebijakan, perbedaan yang terakhir tidaklah diperlukan. Di satu sisi, AI didorong oleh data dan oleh karena itu menghadapi tantangan etika terkait tata kelola data, termasuk persetujuan, kepemilikan, dan privasi. Tantangan-tantangan terkait data ini mungkin diperburuk oleh AI, tetapi akan tetap terjadi bahkan tanpa AI. Di sisi lain, AI merupakan bentuk agensi yang otonom dan belajar mandiri yang unik dan dengan demikian menimbulkan tantangan etika yang unik. Tantangan-tantangan terakhir inilah yang menjadi fokus artikel ini. Perdebatan etika tentang AI sebagai bentuk baru

agensi berawal dari tahun 1960-an. Sejak saat itu, banyak masalah relevan yang berkaitan dengan delegasi dan tanggung jawab. Karena AI digunakan dalam semakin banyak konteks, mulai dari rekrutmen hingga perawatan kesehatan, memahami tugas dan keputusan mana yang akan dipercayakan (didelegasikan) kepada AI dan bagaimana menetapkan tanggung jawab atas kinerjanya merupakan masalah etika yang mendesak.

Pada saat yang sama, ketika AI menjadi tidak terlihat di mana-mana, tantangan etika baru muncul. Perlindungan atas penentuan nasib sendiri manusia adalah salah satu yang paling relevan dan harus ditangani dengan segera. Penerapan AI untuk membuat profil pengguna untuk iklan bertarget, seperti dalam kasus penyedia layanan daring, dan dalam kampanye politik, seperti yang diungkapkan oleh kasus Cambridge Analytica, memberikan contoh yang jelas tentang potensi AI untuk menangkap preferensi dan karakteristik pengguna dan dengan demikian membentuk tujuan mereka dan mendorong perilaku mereka ke tingkat yang dapat merusak penentuan nasib sendiri mereka.

6.1 DELEGASI DAN TANGGUNG JAWAB

Aplikasi AI semakin meluas. Pengguna mengandalkannya untuk menangani berbagai tugas, mulai dari pengiriman barang hingga memastikan pertahanan nasional. Menugaskan tugas-tugas ini kepada AI membawa manfaat besar bagi masyarakat. AI menurunkan biaya, mengurangi risiko, meningkatkan konsistensi dan keandalan, serta memungkinkan solusi baru untuk masalah kompleks. Misalnya, aplikasi AI dapat menurunkan kesalahan diagnostik hingga 85% pada pasien kanker payudara, dan sistem keamanan siber AI dapat mengurangi waktu rata-rata untuk mengidentifikasi dan menetralkan serangan siber dari 101 hari menjadi beberapa jam saja.

Namun, pendelegasian juga dapat menyebabkan konsekuensi yang merugikan dan tidak diinginkan, terutama ketika melibatkan keputusan atau tugas yang sensitif dan mengecualikan atau bahkan menghalangi pengawasan manusia. Kasus COMPAS, sebuah sistem hukum AI yang mendiskriminasi pria Afrika-Amerika dan Hispanik dalam pengambilan keputusan pemberian pembebasan bersyarat, telah menjadi sorotan publik. Prosedur pengawasan manusia yang kuat diperlukan untuk meminimalkan konsekuensi yang tidak diinginkan tersebut dan mengatasi dampak tidak adil dari AI. Namun, pengawasan manusia tidaklah memadai jika hanya menangani masalah setelah terjadi.

Teknik untuk menjelaskan AI dan memprediksi hasilnya juga diperlukan. Program Kecerdasan Buatan yang Dapat Dijelaskan (*Explainable Artificial Intelligence*) dari DARPA (*Defense Advanced Research Project Agency*) adalah contoh yang sangat baik. Tujuan program ini adalah untuk mendefinisikan teknik-teknik baru untuk menjelaskan proses pengambilan keputusan sistem AI. Hal ini akan memungkinkan pengguna untuk memahami cara kerja sistem AI, dan perancang serta pengembang untuk meningkatkan sistem guna menghindari kesalahan dan mengurangi risiko penyalahgunaan. Agar berhasil, proyek serupa harus mencakup analisis dampak etis sejak awal, untuk menilai manfaat dan risiko AI serta mendefinisikan prinsip-prinsip panduan untuk desain dan penggunaan AI yang etis.

Efek dari keputusan atau tindakan berbasis AI seringkali merupakan hasil dari interaksi yang tak terhitung jumlahnya di antara banyak aktor, termasuk perancang, pengembang, pengguna, perangkat lunak, dan perangkat keras. Ini dikenal sebagai agensi terdistribusi (*Distributed Agency*/JO). Dengan agensi terdistribusi, muncul tanggung jawab terdistribusi. Kerangka kerja etika yang ada membahas tanggung jawab individu dan manusia, dengan tujuan mengalokasikan hukuman atau penghargaan berdasarkan tindakan dan niat individu. Kerangka kerja tersebut tidak dikembangkan untuk menangani tanggung jawab terdistribusi.

Baru-baru ini, teori-teori etika baru telah didefinisikan untuk memperhitungkan agensi terdistribusi. Teori-teori yang diusulkan bergantung pada tanggung jawab kontraktual dan gugatan atau tanggung jawab ketat dan mengadopsi model tanggung jawab tanpa kesalahan. Model ini memisahkan tanggung jawab agen dari niat mereka untuk melakukan tindakan tertentu atau kemampuan mereka untuk mengendalikan hasilnya, dan menjadikan semua agen sistem terdistribusi, seperti perusahaan, bertanggung jawab. Ini adalah kunci ketika mempertimbangkan kasus AI, karena model ini mendistribusikan tanggung jawab moral di antara perancang, regulator, dan pengguna.

Dengan demikian, model ini memainkan peran sentral dalam mencegah kejahatan dan mendorong kebaikan, karena mendorong semua agen yang terlibat untuk mengadopsi perilaku yang bertanggung jawab. Menetapkan praktik yang baik untuk pendelegasian dan mendefinisikan model-model baru untuk menetapkan tanggung jawab moral sangat penting untuk memanfaatkan peluang yang diciptakan oleh AI dan mengatasi tantangan terkait, tetapi itu masih belum cukup. Analisis etika harus diperluas untuk memperhitungkan pengaruh tak terlihat yang dilakukan oleh AI terhadap perilaku manusia.

6.2 KETIDAKTAMPAKAN DAN PENGARUH

AI mendukung layanan, platform, dan perangkat yang ada di mana-mana dan digunakan setiap hari. Pada tahun 2017, Federasi Robotika Internasional memperkirakan bahwa pada tahun 2020, lebih dari 1,7 juta robot baru bertenaga AI akan dipasang di pabrik-pabrik di seluruh dunia. Pada tahun yang sama, perusahaan Juniper Networks mengeluarkan laporan yang memperkirakan bahwa, pada tahun 2022, 55% rumah tangga di seluruh dunia akan memiliki asisten suara, seperti Amazon Alexa.

Seiring perkembangannya dan penyebarannya, AI menyatu dengan kehidupan, pengalaman, dan lingkungan kita, serta menjadi fasilitator tak kasat mata yang memediasi interaksi kita dengan cara yang praktis dan nyaris tak terlihat. Meskipun menciptakan peluang baru, integrasi AI yang tak kasat mata ini ke dalam lingkungan kita menimbulkan masalah etika lebih lanjut. Beberapa di antaranya bergantung pada domain. Misalnya, kepercayaan dan transparansi sangat penting ketika menanamkan solusi AI di rumah, sekolah, atau rumah sakit, sedangkan kesetaraan, keadilan, dan perlindungan kreativitas dan hak-hak karyawan sangat penting dalam integrasi AI di tempat kerja.

Namun, integrasi AI juga menimbulkan risiko mendasar lainnya: erosi penentuan nasib sendiri manusia karena ketidaktampakan dan kekuatan pengaruh AI. Ketidaktampakan ini meningkatkan kekuatan pengaruh AI. Dengan kemampuan prediktif dan dorongan tanpa henti,

ada di mana-mana tetapi tak terlihat, sistem AI dapat membentuk pilihan dan tindakan kita dengan mudah dan diam-diam. Ini tidak selalu merugikan. Misalnya, AI dapat mendorong interaksi dan kerja sama sosial.

Namun, AI juga dapat menggunakan kekuatan pengaruhnya melampaui keinginan atau pemahaman kita, melemahkan kendali kita terhadap lingkungan, masyarakat, dan pada akhirnya terhadap pilihan, proyek, identitas, dan kehidupan kita. Perancangan dan penggunaan AI tak kasat mata yang tidak tepat dapat mengancam kemampuan kita yang rapuh, namun konstitutif, untuk menentukan kehidupan dan identitas kita sendiri serta menjaga pilihan kita tetap terbuka.

6.3 ETIKA TRANSLASIONAL

Untuk menghadapi risiko yang ditimbulkan oleh AI, sangat penting untuk mengidentifikasi seperangkat prinsip etika fundamental yang tepat untuk menginformasikan desain, regulasi, dan penggunaan AI serta memanfaatkannya untuk memberi manfaat sekaligus menghormati individu dan masyarakat. Hal ini bukanlah tugas yang mudah, karena prinsip-prinsip etika dapat bervariasi tergantung pada konteks budaya dan domain analisis. Ini adalah masalah yang ditangani oleh Inisiatif Global IEEE tentang Etika Sistem Otonom dan Cerdas (*IEEE Standards Association*, t.t.) dengan tujuan memajukan debat publik tentang nilai-nilai dan prinsip-prinsip yang seharusnya mendasari penggunaan AI yang etis.

Yang lebih penting, beberapa kesepakatan tentang prinsip-prinsip fundamental mulai muncul. Sebuah analisis komparatif terbaru dari inisiatif-inisiatif internasional utama yang berfokus pada etika AI menyoroti tumpang tindih substantif antara prinsip-prinsip yang didukung oleh inisiatif-inisiatif ini dan beberapa prinsip utama bioetika, yaitu kebaikan, nonmaleficence, otonomi, dan keadilan. Ada alasan untuk optimistis tentang konvergensi lebih lanjut, karena prinsip-prinsip lain dapat diekstraksi dari Deklarasi Universal Hak Asasi Manusia. Konvergensi ini akan mendorong koherensi, dan dengan demikian kompatibilitas, berbagai kerangka etika untuk AI dan memberikan panduan etika menyeluruh untuk desain, regulasi, dan penggunaan teknologi ini.

Setelah diidentifikasi, prinsip-prinsip etika harus diterjemahkan menjadi pedoman yang layak untuk membentuk inovasi berbasis AI. Penerjemahan semacam itu memiliki preseden, terutama dalam bidang kedokteran, di mana penelitian translasi berpindah "dari bangku ke samping tempat tidur;" membangun kemajuan penelitian dalam biologi untuk mengembangkan terapi dan perawatan baru. Demikian pula, etika translasi dibangun di atas kemajuan akademis untuk membentuk pendekatan regulasi dan tata kelola. Pendekatan ini mendukung rekomendasi yang akan datang untuk desain dan regulasi AI yang etis yang akan dikeluarkan oleh proyek AI4People.

Diluncurkan di Parlemen Eropa pada Februari 2018, AI4People dibentuk untuk membantu mengarahkan AI menuju kebaikan masyarakat dan semua orang di dalamnya. Inisiatif ini menggabungkan upaya komite ilmiah para ahli internasional dan forum para pemangku kepentingan, dengan berkonsultasi dengan Kelompok Ahli Tingkat Tinggi tentang

Kecerdasan Buatan dari Komisi Eropa, untuk mengusulkan serangkaian rekomendasi konkret dan dapat ditindaklanjuti untuk pengembangan AI yang etis dan lebih disukai secara sosial.

Etika translasional AI perlu merumuskan metodologi pandangan ke depan untuk menunjukkan risiko dan peluang etis serta mencegah konsekuensi yang tidak diinginkan. Analisis penilaian dampak adalah contoh dari metodologi ini. Analisis ini memberikan evaluasi langkah demi langkah tentang dampak praktik atau teknologi yang diterapkan dalam organisasi tertentu terhadap aspek-aspek seperti privasi, transparansi, atau liabilitas.

Metodologi pandangan ke depan tidak akan pernah dapat memetakan seluruh spektrum peluang, risiko, dan konsekuensi yang tidak diinginkan dari sistem AI, tetapi dapat mengidentifikasi alternatif yang lebih baik, tindakan yang berharga, risiko yang mungkin terjadi, dan strategi mitigasi. Hal ini memiliki keuntungan ganda. Sebagai strategi peluang, metodologi pandangan ke depan dapat membantu memanfaatkan solusi etis. Sebagai bentuk manajemen risiko, metodologi ini dapat membantu mencegah atau memitigasi kesalahan yang merugikan, dengan menghindari keputusan atau tindakan yang secara etis tidak dapat diterima. Hal ini akan menurunkan biaya peluang dari pilihan yang tidak dibuat atau opsi yang tidak diambil karena kurangnya kejelasan atau ketakutan akan reaksi negatif.

Regulasi etis atas desain dan penggunaan AI merupakan tugas yang kompleks namun penting. Alternatifnya dapat menyebabkan devaluasi hak-hak individu dan nilai-nilai sosial, penolakan inovasi berbasis AI, dan pada akhirnya hilangnya kesempatan untuk menggunakan AI guna meningkatkan kesejahteraan individu dan sosial. Umat manusia mempelajari pelajaran ini dengan cara yang sulit ketika tidak mengatur dampak revolusi industri terhadap angkatan kerja, dan juga ketika terlambat menyadari dampak lingkungan dari industrialisasi besar-besaran dan konsumerisme global. Butuh waktu yang sangat lama, kerusakan sosial, dan bahkan revolusi untuk melindungi hak-hak pekerja dan membangun kerangka kerja keberlanjutan.

Revolusi AI sama pentingnya, dan umat manusia tidak boleh mengulangi kesalahan yang sama. Sangat penting untuk menjawab pertanyaan-pertanyaan baru tentang hakikat masyarakat pasca-AI dan nilai-nilai yang seharusnya mendasari desain, regulasi, dan penggunaan AI dalam masyarakat tersebut. Inilah mengapa inisiatif seperti proyek AI4People dan IEEE yang disebutkan di atas, strategi Uni Eropa (UE) untuk AI, Deklarasi Kerja Sama UE tentang Kecerdasan Buatan, dan Kemitraan Kecerdasan Buatan untuk Memberi Manfaat bagi Manusia dan Masyarakat sangat penting (lihat materi tambahan untuk bacaan lebih lanjut yang disarankan). Upaya terkoordinasi oleh masyarakat sipil, politik, bisnis, dan akademisi akan membantu mengidentifikasi dan mengupayakan strategi terbaik untuk menjadikan AI sebagai kekuatan kebaikan dan membuka potensinya untuk mendorong kemajuan manusia sekaligus menghormati martabat manusia.

BAB 7

MASALAH DAN SOLUSI ETIKA ALGORITMA

Selama satu dekade terakhir, penelitian mengenai etika algoritma mengalami perkembangan yang sangat pesat. Peningkatan signifikan dalam penggunaan dan penerapan algoritma pembelajaran mesin di berbagai bidang kehidupan telah menimbulkan berbagai persoalan etis yang mendesak untuk ditelaah secara kritis. Munculnya tantangan baru ini juga mendorong munculnya beragam solusi etika yang bertujuan untuk memastikan bahwa teknologi digunakan secara adil, transparan, dan bertanggung jawab.

Bab ini disusun sebagai kelanjutan dari tinjauan awal mengenai etika algoritma yang diterbitkan pada tahun 2016. Tujuannya adalah untuk memperkaya diskusi tentang bagaimana mengidentifikasi dan menganalisis implikasi etis dari algoritma dalam konteks sosial yang terus berkembang. Analisis yang disajikan mencakup isu-isu epistemik dan normatif terkini, serta menawarkan panduan praktis yang dapat diterapkan dalam proses perancangan, pengembangan, dan implementasi algoritma. Dengan demikian, bab ini berkontribusi dalam membangun kerangka tata kelola algoritma yang berorientasi pada prinsip-prinsip etika.

7.1 PENDAHULUAN

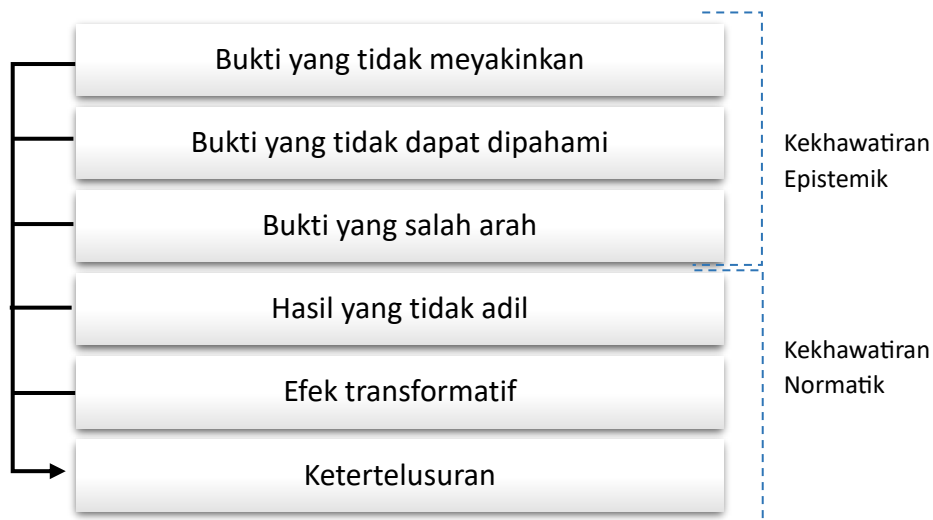
Algoritma telah menjadi elemen kunci yang mendasari layanan dan infrastruktur krusial masyarakat informasi. Setiap hari, individu berinteraksi dengan sistem rekomendasi sistem algoritmik yang memberikan saran tentang apa yang mungkin disukai pengguna baik untuk memilih lagu, film, produk, atau bahkan teman. Pada saat yang sama, sekolah dan rumah sakit, lembaga keuangan, pengadilan, badan pemerintah daerah, dan pemerintah nasional, semuanya semakin bergantung pada algoritma untuk membuat keputusan penting.

Potensi algoritma untuk meningkatkan kesejahteraan individu dan sosial disertai dengan risiko etika yang signifikan. Algoritma tidak netral secara etika. Pertimbangkan, misalnya, bagaimana keluaran algoritma penerjemahan dan mesin pencari sebagian besar dipersepsikan objektif, namun seringkali mengodekan bahasa dengan cara yang bergender. Bias juga telah dilaporkan dalam iklan algoritmik, dengan peluang untuk pekerjaan bergaji tinggi dan pekerjaan di bidang sains dan teknologi lebih sering diiklankan kepada pria daripada wanita. Demikian pula, algoritma prediksi yang digunakan untuk mengelola data kesehatan jutaan pasien di Amerika Serikat memperburuk masalah yang sudah ada, dengan pasien kulit putih diberikan perawatan yang jauh lebih baik daripada pasien kulit hitam yang memiliki tingkat kesulitan serupa. Sementara solusi untuk masalah ini sedang dibahas dan dirancang, jumlah sistem algoritmik yang menunjukkan masalah etika terus bertambah.

Sejak 2012, kecerdasan buatan (AI) telah mengalami 'musim panas' baru, baik dari segi kemajuan teknis yang telah dicapai maupun perhatian yang diterima bidang ini dari akademisi, pembuat kebijakan, teknolog, dan investor. Dalam hal ini, terdapat semakin banyak penelitian tentang implikasi etis algoritma, khususnya yang berkaitan dengan *keadilan*, *akuntabilitas*, dan *transparansi*. Pada tahun 2016, kelompok riset kami di Digital Ethics Lab menerbitkan sebuah

Peran AI dalam Etika Tata Kelola dan Kebijakan
Dr. Ir. Agus Wibowo M.Kom., M.Si., M.M.

studi komprehensif yang berupaya memetakan isu-isu etis ini. Namun, bidang ini berubah dengan cepat dan masalah-masalah etika baru serta cara-cara untuk mengatasinya telah bermunculan, sehingga studi tersebut perlu ditingkatkan dan diperbarui. Khususnya, penelitian tentang etika algoritma telah meningkat secara signifikan sejak 2016, ketika pemerintah nasional, organisasi non-pemerintah, dan perusahaan swasta mulai mengambil peran penting dalam diskusi tentang AI dan algoritma yang "adil" dan "beretika".



Gambar 7.1 Enam jenis masalah etika yang ditimbulkan oleh algoritma.

Baik kuantitas maupun kualitas penelitian yang tersedia tentang topik ini telah berkembang pesat. Mengingat perubahan ini, artikel ini memperbarui penelitian kami sebelumnya dengan mempertimbangkan wawasan baru tentang etika algoritma, memperbarui analisis awal, menyertakan referensi literatur yang terlewatkan dalam tinjauan awal, dan memperluas topik yang dianalisis, termasuk misalnya penelitian tentang AI untuk kebaikan sosial (lihat Kesimpulan).

Pada saat yang sama, peta konseptual yang diusulkan pada tahun 2016 (lihat Gambar 7.1) tetap menjadi kerangka kerja yang bermanfaat untuk meninjau perdebatan terkini tentang etika algoritma, mengidentifikasi permasalahan etika yang ditimbulkan oleh algoritma, dan solusi yang telah diusulkan dalam literatur relevan terkini. Secara khusus, pada Bagian 2, kami merangkum peta konseptual tersebut. Pada Bagian 3, 4, 5, 6, 7, dan 8, kami menawarkan meta-analisis dari perdebatan terkini tentang etika algoritma dan menghubungkannya dengan jenis-jenis permasalahan etika yang telah diidentifikasi sebelumnya. Bagian 9 menyimpulkan artikel ini dengan ikhtisar.

7.2 PETA ETIKA ALGORITMA

Terdapat sedikit kesepakatan dalam literatur terkait mengenai definisi algoritma. Istilah ini sering digunakan untuk menunjukkan definisi formal algoritma sebagai konstruk matematika, dengan "struktur kendali gabungan yang terbatas, abstrak, efektif, dan imperatif, yang diberikan secara imperatif, mencapai tujuan tertentu dengan ketentuan tertentu", serta

pemahaman spesifik domain yang berfokus pada implementasi konstruk matematika ini ke dalam teknologi yang dikonfigurasi untuk tugas tertentu.

Dalam artikel ini, kami memutuskan untuk mempertahankan pendekatan yang sama yang diadopsi dalam artikel tahun 2016 dan berfokus pada isu-isu etika yang ditimbulkan oleh algoritma sebagai konstruk matematika, implementasinya sebagai program dan konfigurasi (aplikasi), serta cara-cara untuk mengatasinya. Kami mempertimbangkan algoritma yang digunakan untuk (1) mengubah data menjadi bukti untuk hasil tertentu, yang digunakan untuk (2) memicu dan memotivasi tindakan yang mungkin memiliki konsekuensi etis.

Tindakan (1) dan (2) dapat dilakukan oleh algoritma (semi-)otonom seperti algoritma Pembelajaran Mesin (ML) dan hal ini mempersulit (3) atribusi tanggung jawab atas dampak tindakan yang dapat dipicu oleh suatu algoritma. Di sini, ML menjadi perhatian khusus, sebagai bidang yang mencakup arsitektur pembelajaran mendalam. Sistem komputer yang menerapkan algoritma ML dapat digambarkan sebagai "otonom" atau "semi-otonom", sejauh keluarannya diinduksi dari data dan dengan demikian bersifat non-deterministik.

Berdasarkan pendekatan ini, kami menggunakan peta konseptual yang ditunjukkan pada Gambar 8.1 untuk mengidentifikasi isu-isu etika yang ditimbulkan oleh algoritma. Peta tersebut mengidentifikasi enam isu etika, yang mendefinisikan ruang konseptual etika algoritma sebagai bidang penelitian. Tiga dari isu etika tersebut merujuk pada faktor-faktor epistemik, khususnya: bukti yang tidak meyakinkan, tidak dapat dipahami, dan salah arah. Dua di antaranya secara eksplisit normatif: hasil yang tidak adil dan dampak transformatif; sementara satu ketertelusuran relevan baik untuk tujuan epistemik maupun normatif.

Faktor-faktor epistemik dalam peta menyoroti relevansi kualitas dan akurasi data terhadap justifikasi kesimpulan yang dicapai algoritma dan yang, pada gilirannya, dapat membentuk keputusan berbobot moral yang memengaruhi individu, masyarakat, dan lingkungan. Kekhawatiran normatif yang diidentifikasi dalam peta secara eksplisit merujuk pada dampak etis dari tindakan dan keputusan yang digerakkan oleh algoritma, termasuk kurangnya transparansi (opasitas) proses algoritmik, hasil yang tidak adil, dan konsekuensi yang tidak diinginkan. Kekhawatiran epistemik dan normatif, bersama dengan distribusi desain, pengembangan, dan penerapan algoritma, menyulitkan penelusuran rantai peristiwa dan faktor yang mengarah pada suatu hasil tertentu, sehingga menghambat kemungkinan mengidentifikasi penyebabnya, dan mengaitkan tanggung jawab moral atasnya. Inilah yang dimaksud dengan kekhawatiran etis keenam, yaitu ketertelusuran.

Penting untuk ditegaskan bahwa peta konseptual ini dapat diinterpretasikan baik pada tingkat mikro-etika maupun makro-etika. Pada tingkat mikro-etika, peta ini menyoroti permasalahan etika yang mungkin ditimbulkan oleh algoritma tertentu. Dengan menyoroti bagaimana isu-isu ini tak terpisahkan dari isu-isu terkait data dan tanggung jawab, hal ini menunjukkan perlunya pendekatan makro-etika untuk membahas etika algoritma sebagai bagian dari ruang konseptual yang lebih luas, yaitu etika digital (Floridi dan Taddeo 2016). Sebagaimana dikemukakan Floridi dan Taddeo:

Meskipun merupakan lini penelitian yang berbeda, etika data, algoritma, dan praktik jelas saling terkait... Etika harus membahas keseluruhan ruang konseptual dan dengan demikian ketiga sumbu penelitian tersebut secara bersamaan, meskipun dengan prioritas dan fokus yang berbeda.

Dalam sisa artikel ini, kami membahas masing-masing dari enam isu etika ini secara bergantian, menawarkan analisis terbaru tentang literatur etika algoritma (pada tingkat mikro), dengan tujuan berkontribusi pada perdebatan tentang etika digital (pada tingkat makro).

7.3 BUKTI YANG TIDAK KONKLUSIF

Penelitian yang berfokus pada *bukti yang tidak konklusif* mengacu pada cara algoritma pembelajaran mesin non-deterministik menghasilkan keluaran yang dinyatakan dalam bentuk probabilistik. Jenis algoritma ini umumnya mengidentifikasi asosiasi dan korelasi antar variabel dalam data yang mendasarinya, tetapi bukan hubungan sebab akibat. Dengan demikian, algoritma ini mendorong praktik apophenia: "melihat pola yang sebenarnya tidak ada, hanya karena data dalam jumlah besar dapat menawarkan koneksi yang menyebar ke segala arah".

Hal ini sangat bermasalah, karena pola yang diidentifikasi oleh algoritma mungkin merupakan hasil dari sifat inheren sistem yang dimodelkan oleh data, dari kumpulan data (yaitu, dari model itu sendiri, alih-alih sistem yang mendasarinya), atau dari manipulasi kumpulan data yang terampil (sifat yang bukan milik model maupun sistem). Hal ini, misalnya, terjadi pada paradoks Simpson, ketika tren yang diamati dalam kelompok data yang berbeda berbalik ketika data diagregasi. Dalam dua kasus terakhir, kualitas data yang buruk menyebabkan bukti yang tidak meyakinkan untuk mendukung keputusan manusia.

Penelitian terbaru telah menggarisbawahi kekhawatiran bahwa bukti yang tidak meyakinkan dapat menimbulkan risiko etika yang serius. Misalnya, berfokus pada indikator non-kausal dapat mengalihkan perhatian dari akar penyebab suatu masalah. Bahkan dengan penggunaan metode kausal, data yang tersedia mungkin tidak selalu mengandung informasi yang cukup untuk membenarkan suatu tindakan atau membuat keputusan yang adil. Kualitas data ketepatan waktu, kelengkapan, dan kebenaran suatu kumpulan data membatasi pertanyaan yang dapat dijawab menggunakan kumpulan data tertentu.

Selain itu, wawasan yang dapat diekstraksi dari kumpulan data pada dasarnya bergantung pada asumsi yang memandu proses pengumpulan data itu sendiri. Misalnya, algoritma yang dirancang untuk memprediksi luaran pasien dalam pengaturan klinis sepenuhnya bergantung pada masukan data yang dapat dikuantifikasi (misalnya, tanda-tanda vital dan tingkat keberhasilan perawatan komparatif sebelumnya), sementara mengabaikan fakta emosional lainnya (misalnya, keinginan untuk hidup) yang dapat berdampak signifikan pada luaran pasien, dan dengan demikian melemahkan akurasi prediksi algoritmik. Contoh ini menyoroti bagaimana wawasan yang berasal dari pemrosesan data algoritmik dapat bersifat tidak pasti, tidak lengkap, dan sensitif terhadap waktu.

Seseorang mungkin menganut pendekatan induktif yang naif dan berasumsi bahwa bukti yang tidak meyakinkan dapat dihindari jika algoritma diberi data yang cukup, bahkan jika

penjelasan kausal untuk hasil ini tidak dapat ditetapkan. Namun, penelitian terbaru menolak pandangan ini. Secara khusus, literatur yang berfokus pada risiko etika pembuatan profil ras menggunakan sistem algoritmik telah menunjukkan batasan pendekatan ini, antara lain menyoroti bahwa ketimpangan struktural yang telah berlangsung lama seringkali tertanam kuat dalam kumpulan data algoritma dan jarang, bahkan mungkin tidak pernah, dikoreksi.

Lebih banyak data saja tidak akan menghasilkan akurasi atau representasi yang lebih baik. Sebaliknya, data tersebut justru dapat memperburuk masalah data yang tidak meyakinkan dengan memungkinkan ditemukannya korelasi yang sebenarnya tidak ada. Seperti yang dikatakan Ruha Benjamin (2020), "kedalaman komputasi tanpa kedalaman historis atau sosiologis hanyalah pembelajaran yang dangkal. Keterbatasan ini menimbulkan kendala serius terhadap justifikasi keluaran algoritmik, yang dapat berdampak negatif pada individu atau seluruh populasi karena inferensi yang kurang optimal atau, dalam kasus ilmu fisika, bahkan memengaruhi bukti yang mendukung atau menentang "teori ilmiah tertentu".

Inilah sebabnya mengapa sangat penting untuk memastikan bahwa data yang dimasukkan ke dalam algoritma divalidasi secara independen, dan langkah-langkah retensi dan reproduktifitas data tersedia untuk memitigasi bukti yang tidak meyakinkan yang mengarah pada tindakan yang tidak dapat dibenarkan, serta proses audit untuk mengidentifikasi hasil yang tidak adil dan konsekuensi yang tidak diinginkan.

Bahaya yang timbul dari bukti yang tidak meyakinkan dan wawasan yang dapat ditindaklanjuti yang keliru juga berasal dari objektivitas mekanistik yang dirasakan terkait dengan analitik yang dihasilkan komputer. Hal ini dapat menyebabkan para pengambil keputusan mengabaikan penilaian pengalaman mereka sendiri yang disebut 'bias otomatisasi' atau bahkan mengabaikan sebagian tanggung jawab mereka atas keputusan. Sebagaimana akan kita lihat di Bagian 4 dan 8, kurangnya pemahaman tentang bagaimana algoritma menghasilkan keluaran memperburuk masalah ini.

7.4 BUKTI YANG TAK TERPAHAMI YANG MENYEBABKAN KETIDAKJELASAN

Bukti yang tak terpahami berfokus pada permasalahan terkait kurangnya transparansi yang seringkali menjadi ciri algoritma (terutama algoritma dan model pembelajaran mesin); infrastruktur sosio-teknis tempat algoritma dan model tersebut berada; dan keputusan yang didukungnya. Kurangnya transparansi baik yang melekat karena keterbatasan teknologi maupun yang diperoleh melalui keputusan desain dan pengaburan data yang mendasarinya sering kali berujung pada kurangnya pengawasan dan/atau akuntabilitas, dan menyebabkan kurangnya "kepercayaan".

Menurut literatur terkini, faktor-faktor yang berkontribusi terhadap kurangnya transparansi algoritmik secara keseluruhan meliputi: ketidakmungkinan kognitif manusia untuk menginterpretasikan model dan kumpulan data algoritmik yang masif; kurangnya alat yang tepat untuk memvisualisasikan dan melacak sejumlah besar kode dan data; kode dan data yang sangat buruk terstruktur sehingga tidak mungkin dibaca; dan pembaruan yang berkelanjutan dan pengaruh manusia atas suatu model.

Kurangnya transparansi juga merupakan karakteristik inheren dari algoritma pembelajaran mandiri, yang mengubah logika keputusan mereka (menghasilkan serangkaian aturan baru) selama proses pembelajaran, sehingga menyulitkan pengembang untuk mempertahankan pemahaman terperinci tentang mengapa perubahan tertentu dibuat. Namun, ini tidak selalu menghasilkan hasil yang buruk, karena bahkan tanpa memahami setiap langkah logis, pengembang dapat menyesuaikan hiperparameter, parameter yang mengatur proses pelatihan, untuk menguji berbagai keluaran. Dalam hal ini, Martin (2019) menekankan bahwa, meskipun kesulitan menjelaskan keluaran algoritma ML memang nyata, penting untuk tidak membiarkan kesulitan ini mendorong organisasi untuk mengembangkan sistem yang kompleks demi menghindari tanggung jawab.

Kurangnya transparansi juga dapat diakibatkan oleh sifat lunak algoritma, yang memungkinkan algoritma diprogram ulang secara berkelanjutan, terdistribusi, dan dinamis. Sifat lunak algoritma memungkinkan pengembang untuk memantau dan meningkatkan algoritma yang telah diterapkan, tetapi juga dapat disalahgunakan untuk mengaburkan riwayat evolusinya dan membuat pengguna akhir bingung tentang *affordance* suatu algoritma. Pertimbangkan misalnya algoritma pencarian utama Google. Sifat lunaknya memungkinkan perusahaan untuk melakukan revisi berkelanjutan, yang menunjukkan ketidakstabilan permanen. Hal ini mengharuskan mereka yang terpengaruh oleh algoritma untuk memantaunya secara konstan dan memperbarui pemahaman mereka sebuah tugas yang mustahil bagi sebagian besar orang.

Sebagaimana dicatat oleh Floridi dan Turilli (2009, 105), transparansi bukanlah "prinsip etika itu sendiri, melainkan kondisi pro-etika yang memungkinkan atau menghambat praktik atau prinsip etika lainnya". Bahkan, transparansi penuh dapat menimbulkan masalah etika tersendiri: transparansi dapat memberikan informasi penting kepada pengguna tentang fitur dan keterbatasan suatu algoritma, tetapi juga dapat membanjiri pengguna dengan informasi sehingga membuat algoritma menjadi lebih buruk.

Penelitian lain menekankan bahwa fokus yang berlebihan pada transparansi dapat menghambat inovasi dan mengalihkan sumber daya yang seharusnya dapat digunakan untuk meningkatkan keselamatan, kinerja, dan akurasi secara tidak perlu. Misalnya, perdebatan tentang prioritas transparansi (dan kemudahan dijelaskan) khususnya kontroversial dalam konteks algoritma medis.

Transparansi dapat memungkinkan individu untuk memanipulasi sistem. Pengetahuan tentang sumber dataset, asumsi yang mendasari pengambilan sampel, atau metrik yang digunakan algoritma untuk mengurutkan masukan baru, dapat digunakan untuk menemukan cara memanfaatkan algoritma. Namun, kemampuan untuk memanipulasi algoritma hanya dapat dicapai oleh beberapa kelompok populasi misalnya mereka yang memiliki literasi digital lebih tinggi sehingga menciptakan bentuk ketimpangan sosial lainnya. Oleh karena itu, mengaburkan transparansi sebagai tujuan itu sendiri, alih-alih faktor pro-etika yang memungkinkan praktik etika krusial, mungkin tidak menyelesaikan masalah etika yang ada terkait penggunaan algoritma dan, bahkan, menimbulkan masalah baru. Inilah sebabnya mengapa penting untuk membedakan berbagai faktor yang dapat menghambat transparansi

algoritma, mengidentifikasi penyebabnya, dan menjelaskan seruan untuk transparansi dengan menentukan faktor mana yang diperlukan dan pada lapisan sistem algoritmik mana faktor-faktor tersebut harus ditangani.

Ada berbagai cara untuk mengatasi masalah yang berkaitan dengan kurangnya transparansi. Misalnya, Gebru dkk. mengusulkan bahwa kendala transparansi yang ditimbulkan oleh kelenturan algoritma dapat diatasi, sebagian, dengan menggunakan prosedur dokumentasi standar yang serupa dengan yang digunakan dalam industri elektronik, di mana setiap komponen, betapa pun sederhana atau rumitnya, disertai dengan lembar data yang menjelaskan karakteristik operasinya, hasil pengujian, penggunaan yang direkomendasikan, dan informasi lainnya. Sayangnya, dokumentasi yang tersedia untuk publik saat ini jarang ditemukan dalam pengembangan sistem algoritmik, dan belum ada format yang disepakati mengenai apa yang harus disertakan saat mendokumentasikan asal suatu set data.

Meskipun relatif baru, pendekatan lain yang berpotensi menjanjikan untuk menegakkan transparansi algoritmik adalah penggunaan perangkat teknis untuk menguji dan mengaudit sistem algoritmik dan pengambilan keputusan. Menguji apakah algoritma menunjukkan kecenderungan negatif, seperti diskriminasi yang tidak adil, dan mengaudit prediksi atau jejak keputusan secara detail, dapat membantu menjaga tingkat transparansi yang tinggi. Untuk tujuan ini, kerangka kerja diskursif telah dikembangkan untuk membantu bisnis dan organisasi sektor publik memahami potensi dampak algoritma yang tidak transparan, sehingga mendorong praktik yang baik.

Misalnya, AI Now Institute di New York University telah menghasilkan panduan penilaian dampak algoritmik, yang bertujuan untuk meningkatkan kesadaran dan memperbaiki dialog mengenai potensi bahaya algoritma ML. Hal ini mencakup dua tujuan, yaitu memungkinkan pengembang merancang algoritma ML yang lebih transparan, dan karenanya lebih tepercaya, serta meningkatkan pemahaman dan kendali publik terhadap algoritma. Senada dengan itu, Diakopoulos dan Koliska telah menyediakan daftar lengkap "faktor transparansi" di empat lapisan sistem algoritmik: data, model, inferensi, dan antarmuka. Faktor-faktor tersebut antara lain meliputi

ketidakpastian (misalnya, margin kesalahan), ketepatan waktu (misalnya, kapan data dikumpulkan), kelengkapan atau elemen yang hilang, metode pengambilan sampel, asal (misalnya, sumber), dan volume (misalnya, data pelatihan yang digunakan dalam pembelajaran mesin).

Prosedur transparansi yang efektif kemungkinan besar, dan memang seharusnya, melibatkan penjelasan yang dapat diinterpretasikan tentang proses internal sistem ini. berpendapat bahwa meskipun kurangnya transparansi merupakan fitur inheren dari banyak algoritma ML, hal ini tidak berarti bahwa perbaikan tidak dapat dilakukan. Misalnya, perusahaan seperti Google dan IBM telah meningkatkan upaya mereka untuk membuat algoritma ML lebih mudah diinterpretasikan dan inklusif dengan menyediakan alat-alat seperti Explainable AI, AI Explainability 360, dan What-If Tool yang tersedia untuk umum. Alat-alat ini menyediakan

antarmuka visual interaktif bagi pengembang dan juga masyarakat umum yang meningkatkan keterbacaan manusia, mengeksplorasi berbagai hasil model, menyediakan penalaran berbasis kasus, aturan yang dapat diinterpretasikan secara langsung, dan bahkan mengidentifikasi dan memitigasi bias yang tidak diinginkan dalam kumpulan data dan model algoritmik.

Namun, penjelasan untuk algoritma ML dibatasi oleh jenis penjelasan yang dicari, fakta bahwa keputusan sering kali bersifat multidimensi, dan bahwa pengguna yang berbeda mungkin memerlukan penjelasan yang berbeda. Mengidentifikasi metode yang tepat untuk memberikan penjelasan telah menjadi masalah sejak akhir 1990-an, tetapi upaya kontemporer dapat dikategorikan ke dalam dua pendekatan utama: penjelasan yang berpusat pada subjek dan penjelasan yang berpusat pada model. Dalam yang pertama, akurasi dan panjang penjelasan disesuaikan dengan pengguna dan interaksi spesifik mereka dengan algoritma tertentu dan model seperti permainan yang diusulkan oleh; pada yang terakhir, penjelasannya menyangkut model secara keseluruhan dan tidak bergantung pada audiensnya.

Keterjelasan khususnya penting ketika mempertimbangkan semakin banyaknya model dan set data sumber terbuka dan mudah digunakan. Semakin banyak orang awam yang bereksperimen dengan model algoritmik mutakhir yang tersedia secara luas melalui pustaka atau platform daring, seperti GitHub, tanpa selalu memahami sepenuhnya batasan dan propertinya (Hutson 2019). Hal ini mendorong para akademisi untuk menyarankan bahwa, untuk mengatasi masalah kompleksitas teknis, perlu berinvestasi lebih besar dalam pendidikan publik untuk meningkatkan literasi komputasi dan data. Melakukan hal itu tampaknya menjadi solusi jangka panjang yang tepat untuk masalah berlapis-lapis yang diperkenalkan oleh algoritma yang ada di mana-mana, dan perangkat lunak sumber terbuka sering disebut sebagai hal yang penting untuk solusi tersebut.

7.5 BUKTI YANG SALAH ARAH MENYEBABKAN BIAS YANG TIDAK DIINGINKAN

Para pengembang terutama berfokus untuk memastikan bahwa algoritma mereka menjalankan tugas yang dirancang untuknya. Dengan demikian, jenis pemikiran yang memandu para pengembang sangat penting untuk memahami munculnya bias dalam algoritma dan pengambilan keputusan algoritmik. Beberapa akademisi menyebut pemikiran dominan dalam bidang pengembangan algoritma didefinisikan oleh "formalisme algoritmik" suatu kepatuhan terhadap aturan dan bentuk yang ditentukan.

Meskipun pendekatan ini berguna untuk mengabstraksi dan mendefinisikan proses analitis, pendekatan ini cenderung mengabaikan kompleksitas sosial dunia nyata. Memang, pendekatan ini mengarah pada intervensi algoritmik yang berusaha untuk bersikap 'netral' tetapi dengan demikian, pendekatan ini berisiko memperkuat kondisi sosial yang ada, sekaligus menciptakan ilusi presisi. Karena alasan-alasan ini, penggunaan algoritma dalam beberapa situasi dipertanyakan sepenuhnya. Misalnya, semakin banyak akademisi yang mengkritik penggunaan alat penilaian risiko berbasis algoritma di lingkungan pengadilan.

Beberapa akademisi menegaskan batasan abstraksi terkait bias yang tidak diinginkan dalam algoritma dan berpendapat perlunya mengembangkan kerangka sosioteknis untuk menangani dan meningkatkan kewajaran algoritma. Dalam hal ini, menunjukkan lima

“jebakan” abstraksi, atau kegagalan memperhitungkan konteks sosial tempat algoritma beroperasi, yang terus ada dalam desain algoritmik karena tidak adanya kerangka sosioteknis, yaitu:

- (i) kegagalan memodelkan keseluruhan sistem yang di atasnya kriteria sosial, seperti keadilan, akan ditegakkan;
- (ii) kegagalan untuk memahami bagaimana penggunaan kembali solusi algoritmik yang dirancang untuk satu konteks sosial dapat menyesatkan, tidak akurat, atau justru merugikan ketika diterapkan pada konteks yang berbeda;
- (iii) kegagalan untuk memperhitungkan makna utuh dari konsep-konsep sosial seperti keadilan, yang dapat bersifat prosedural, kontekstual, dan dapat diperdebatkan, serta tidak dapat diselesaikan melalui formalisme matematika;
- (iv) kegagalan untuk memahami bagaimana penyisipan teknologi ke dalam sistem sosial yang ada mengubah perilaku dan nilai-nilai yang tertanam dalam sistem yang telah ada sebelumnya; dan
- (v) kegagalan untuk mengenali kemungkinan bahwa solusi terbaik untuk suatu masalah mungkin tidak melibatkan teknologi.

Istilah 'bias' sering kali berkonotasi negatif, tetapi digunakan di sini untuk menunjukkan "penyimpangan dari standar", yang dapat terjadi pada setiap tahap proses desain, pengembangan, dan penerapan. Data yang digunakan untuk melatih algoritma merupakan salah satu sumber utama munculnya bias, baik melalui data yang dipilih secara selektif maupun dari data yang mencerminkan bias sosial yang ada.

Misalnya, ketimpangan struktural yang bermasalah secara moral dan merugikan etnis tertentu mungkin tidak terlihat dalam data sehingga tidak dikoreksi. Selain itu, data yang digunakan untuk melatih algoritma jarang diperoleh "berdasarkan rancangan eksperimen tertentu" dan digunakan meskipun mungkin tidak akurat, bias, atau bias sistemik, sehingga menghasilkan representasi yang buruk dari populasi yang diteliti.

Salah satu pendekatan yang mungkin untuk memitigasi masalah ini adalah dengan sengaja mengecualikan beberapa variabel data tertentu dari informasi pengambilan keputusan algoritmik. Memang, pemrosesan variabel sensitif atau "dilindungi" yang relevan secara statistik seperti jenis kelamin atau ras biasanya dibatasi atau dilarang berdasarkan undang-undang antidiskriminasi dan perlindungan data, untuk membatasi risiko diskriminasi yang tidak adil.

Sayangnya, bahkan jika perlindungan untuk kelas tertentu dapat dikodekan dalam suatu algoritma, selalu ada bias yang tidak dipertimbangkan *sebelumnya*, seperti dalam kasus, misalnya, model bahasa yang mereproduksi teks yang sangat berfokus pada laki-laki. Meskipun bias dapat diantisipasi dan variabel yang dilindungi dikecualikan dari data, proksi yang tidak diantisipasi untuk variabel-variabel ini masih dapat digunakan untuk merekonstruksi bias, yang mengarah ke "bias berdasarkan proksi" yang sulit dideteksi dan dihindari.

Pada saat yang sama, mungkin ada alasan kuat untuk mengandalkan estimator yang bias secara statistik dalam pemrosesan algoritmik, karena estimator tersebut dapat digunakan untuk memitigasi bias data pelatihan. Dengan cara ini, satu jenis bias algoritmik yang

bermasalah diimbangi oleh jenis bias algoritmik lain atau dengan memperkenalkan bias kompensasi saat menginterpretasikan keluaran algoritmik. Pendekatan yang lebih sederhana untuk memitigasi bias dalam data melibatkan uji coba algoritma dalam konteks yang berbeda dan dengan berbagai kumpulan data.

Mempublikasikan model, kumpulan data, dan metadata (berdasarkan asal-usul) untuk memungkinkan pengawasan eksternal juga dapat membantu mengoreksi bias yang tidak terlihat atau tidak diinginkan. Perlu juga dicatat bahwa apa yang disebut 'data sintesis', atau data yang dihasilkan AI, yang dihasilkan melalui pembelajaran penguatan atau jaringan adversarial generatif (GAN) menawarkan peluang untuk mengatasi masalah bias data.

Pembuatan data yang adil dengan GAN dapat membantu mendiversifikasi kumpulan data yang digunakan dalam algoritma visi komputer. Misalnya, StyleGAN2 mampu menghasilkan citra wajah manusia yang tidak ada dengan kualitas tinggi, dan telah terbukti sangat berguna dalam menciptakan kumpulan data wajah manusia yang beragam, sesuatu yang saat ini tidak dimiliki oleh banyak sistem algoritmik untuk pengenalan wajah.

Bias yang tidak diinginkan juga terjadi akibat penerapan algoritma yang tidak tepat. Pertimbangkan bias konteks transfer: bias bermasalah yang muncul ketika algoritma yang berfungsi digunakan di lingkungan baru. Misalnya, jika algoritma layanan kesehatan rumah sakit riset digunakan di klinik pedesaan dan mengasumsikan bahwa tingkat sumber daya yang sama tersedia untuk klinik pedesaan tersebut seperti rumah sakit riset, keputusan alokasi sumber daya layanan kesehatan yang dihasilkan oleh algoritma tersebut akan tidak akurat dan cacat.

Senada dengan itu, Grgić-Hlača dkk. (2018) memperingatkan adanya lingkaran setan ketika algoritma membuat penilaian berantai yang keliru. Misalnya, dalam konteks algoritma penilaian risiko COMPAS, salah satu kriteria penilaian untuk memprediksi residivisme adalah riwayat kriminal teman-teman terdakwa. Oleh karena itu, memiliki teman dengan riwayat kriminal akan menciptakan lingkaran setan di mana terdakwa dengan teman yang telah dihukum akan dianggap lebih mungkin melakukan pelanggaran, dan oleh karena itu dijatuhi hukuman penjara. Hal ini meningkatkan jumlah orang dengan catatan kriminal dalam suatu kelompok berdasarkan korelasi semata.

Contoh-contoh bias algoritmik yang menonjol dalam beberapa tahun terakhir terutama pelaporan investigasi seputar sistem COMPAS telah menyebabkan meningkatnya fokus pada isu-isu keadilan algoritmik. Definisi dan operasionalisasi keadilan algoritmik telah menjadi "tugas mendesak di dunia akademis dan industri", sebagaimana disoroti oleh peningkatan signifikan jumlah makalah, lokakarya, dan konferensi yang didedikasikan untuk 'keadilan, akuntabilitas, dan transparansi' (FAT). Kami menganalisis topik-topik kunci dan kontribusi di bidang ini di bagian selanjutnya.

7.6 HASIL YANG TIDAK ADIL YANG MENYEBABKAN DISKRIMINASI

Terdapat kesepakatan luas mengenai perlunya keadilan algoritmik, terutama untuk memitigasi risiko diskriminasi langsung dan tidak langsung (berdasarkan hukum AS, masing-masing disebut 'perlakuan berbeda' dan 'dampak berbeda') akibat keputusan algoritmik.

Namun, masih terdapat ketidaksepakatan di antara para peneliti mengenai definisi, pengukuran, dan standar keadilan algoritmik. Wong (2019) mengidentifikasi hingga 21 definisi keadilan di berbagai literatur, dan definisi-definisi tersebut seringkali tidak konsisten.

Terdapat banyak nuansa dalam definisi, pengukuran, dan penerapan berbagai standar keadilan algoritmik. Misalnya, keadilan algoritmik dapat didefinisikan baik dalam kaitannya dengan kelompok maupun individu. Empat definisi utama keadilan algoritmik telah menjadi sorotan dalam literatur terkini:

- (i) anti-klasifikasi, yang mengacu pada kategori yang dilindungi, seperti ras dan gender, dan proksinya yang tidak digunakan secara eksplisit dalam pengambilan keputusan;
- (ii) paritas klasifikasi, yang menganggap suatu model adil jika ukuran umum kinerja prediktif, termasuk rasio positif dan negatif palsu, sama di seluruh kelompok yang dilindungi;
- (iii) kalibrasi, yang menganggap keadilan sebagai ukuran seberapa baik algoritma terkalibrasi di antara kelompok yang dilindungi;
- (iv) paritas statistik, yang mendefinisikan keadilan sebagai estimasi probabilitas rata-rata yang sama untuk semua anggota kelompok yang dilindungi.

Namun, masing-masing definisi keadilan yang umum digunakan ini memiliki kekurangan dan umumnya saling bertentangan. Mengambil contoh anti-klasifikasi, karakteristik yang dilindungi, seperti ras, gender, dan agama, tidak dapat begitu saja dihapus dari data pelatihan untuk mencegah diskriminasi, sebagaimana disebutkan di atas. Ketimpangan struktural berarti bahwa titik data yang secara formal non-diskriminatif seperti kode pos dapat bertindak sebagai proksi untuk, dan digunakan, baik secara sengaja maupun tidak sengaja, untuk menyimpulkan karakteristik yang dilindungi, seperti ras.

Ada kasus-kasus penting di mana mempertimbangkan karakteristik yang dilindungi adalah tepat untuk membuat keputusan yang adil. Misalnya, tingkat residivisme perempuan yang lebih rendah berarti bahwa mengecualikan gender sebagai input dalam algoritma residivisme akan membuat perempuan memiliki peringkat risiko yang sangat tinggi. Oleh karena itu, Binns (2018a) menekankan pentingnya mempertimbangkan konteks historis dan sosiologis yang tidak dapat ditangkap dalam data yang disajikan kepada algoritma, tetapi dapat menginformasikan pendekatan yang sesuai konteks terhadap keadilan dalam algoritma. Penting juga untuk dicatat bahwa model algoritmik seringkali menghasilkan hasil yang tidak terduga, bertentangan dengan intuisi manusia, dan mengganggu pemahaman mereka. Misalnya, seperti yang disoroti oleh Grgić-Hlača dkk. (2018), penggunaan fitur yang diyakini orang adil dalam beberapa kasus dapat meningkatkan rasisme yang ditunjukkan oleh algoritma dan menurunkan akurasi.

Mengenai metode untuk meningkatkan keadilan algoritmik, Veale dan Binns dan Katell. menawarkan dua pendekatan. Pendekatan pertama membayangkan intervensi pihak ketiga, di mana entitas di luar penyedia algoritma akan menyimpan data mengenai karakteristik sensitif atau terlindungi dan berupaya mengidentifikasi serta mengurangi diskriminasi yang disebabkan oleh data dan model tersebut. Pendekatan kedua mengusulkan metode berbasis pengetahuan kolaboratif yang berfokus pada sumber daya data berbasis komunitas yang berisi

pengalaman praktis dalam pembelajaran mesin dan pemodelan. Kedua pendekatan ini tidak saling eksklusif, keduanya dapat memberikan manfaat yang berbeda tergantung pada konteks penerapannya, dan kombinasi keduanya juga dapat bermanfaat.

Mengingat dampak signifikan keputusan algoritmik terhadap kehidupan masyarakat dan pentingnya konteks untuk memilih ukuran keadilan yang tepat, sungguh mengejutkan bahwa hanya ada sedikit upaya untuk menangkap pandangan publik tentang keadilan algoritmik. Meneliti persepsi publik tentang berbagai definisi keadilan algoritmik, mencatat bahwa dalam konteks keputusan pinjaman, orang menunjukkan preferensi untuk "definisi keadilan yang terkalibrasi", atau pemilihan berbasis prestasi, dibandingkan dengan "memperlakukan orang yang serupa secara serupa" dan berargumen mendukung prinsip tindakan afirmatif. Dalam studi serupa, Lee (2018) menawarkan bukti yang menunjukkan bahwa, ketika mempertimbangkan tugas-tugas yang membutuhkan keterampilan unik manusia, orang menganggap keputusan algoritmik kurang adil dan algoritma kurang dapat dipercaya.

Melaporkan pekerjaan empiris yang dilakukan pada interpretabilitas dan transparansi algoritmik, Webb et al. (2019) mengungkapkan bahwa referensi moral, khususnya tentang keadilan, konsisten di antara partisipan yang mendiskusikan preferensi mereka terhadap algoritma. Studi ini mencatat bahwa orang cenderung melampaui preferensi pribadi untuk berfokus pada "perilaku benar dan salah", sebagai cara untuk menunjukkan perlunya memahami konteks penerapan algoritma dan kesulitan memahami algoritma beserta konsekuensinya. Dalam konteks sistem rekomendasi, Burke (2017) mengusulkan pendekatan multi-pemangku kepentingan dan multi-pihak untuk mendefinisikan keadilan, melampaui definisi yang berpusat pada pengguna untuk mencakup kepentingan pemangku kepentingan sistem lainnya.

Telah menjadi jelas bahwa memahami pandangan publik tentang keadilan algoritmik akan membantu para teknolog dalam mengembangkan algoritma dengan prinsip-prinsip keadilan yang selaras dengan sentimen masyarakat umum terhadap gagasan keadilan yang berlaku. Mendasarkan keputusan desain penyedia algoritma "dengan alasan yang dapat diterima oleh mereka yang paling terdampak" serta "terbuka terhadap penyesuaian berdasarkan alasan baru" sangat penting untuk meningkatkan dampak sosial algoritma. Namun, penting untuk dipahami bahwa ukuran keadilan seringkali sama sekali tidak memadai ketika digunakan untuk memvalidasi model yang diterapkan pada kelompok masyarakat yang sudah dirugikan dalam masyarakat karena asal usul, tingkat pendapatan, atau orientasi seksual mereka. Kita tidak dapat "mengoptimalkan" dinamika kekuatan ekonomi, sosial, dan politik yang ada.

7.7 DAMPAK TRANSFORMATIF BAGI OTONOMI DAN PRIVASI INFORMASI

Dampak kolektif algoritma telah memicu diskusi tentang otonomi yang diberikan kepada pengguna akhir. Layanan berbasis algoritma semakin banyak ditampilkan "dalam ekosistem permasalahan sosio-teknis yang kompleks", yang dapat menghambat otonomi pengguna. Batasan otonomi pengguna berasal dari tiga sumber:

- (i) distribusi yang luas dan proaktivitas algoritma (pembelajaran) untuk menginformasikan pilihan pengguna
- (ii) pemahaman pengguna yang terbatas tentang algoritma;
- (iii) kurangnya kekuatan orde kedua (atau banding) atas hasil algoritmi.

Dalam mempertimbangkan tantangan etika AI, berfokus pada dampak algoritma otonom dan belajar mandiri terhadap penentuan nasib sendiri manusia dan menekankan bahwa "kekuatan prediktif dan dorongan AI yang tak henti-hentinya, meskipun tidak disengaja, seharusnya mendorong dan bukan melemahkan martabat dan penentuan nasib sendiri manusia".

Risiko bahwa sistem algoritmik dapat menghambat otonomi manusia dengan membentuk pilihan pengguna telah banyak dilaporkan dalam literatur dan telah menjadi pusat perhatian dalam sebagian besar prinsip etika tingkat tinggi untuk AI, termasuk, *antara lain*, prinsip-prinsip dari Kelompok Etika Eropa dalam Sains dan Teknologi Komisi Eropa, dan Komite Kecerdasan Buatan Dewan Perwakilan Rakyat Inggris. Dalam analisis mereka terhadap prinsip-prinsip tingkat tinggi ini, Floridi dan Cowls (2019) mencatat bahwa tidaklah cukup jika algoritma hanya mempromosikan otonomi manusia: otonomi algoritma harus dibatasi dan dapat dibalik. Melihat lebih jauh dari Barat, Prinsip-Prinsip AI Beijing yang dikembangkan oleh konsorsium perusahaan dan universitas terkemuka Tiongkok untuk memandu penelitian dan pengembangan AI juga menekankan bahwa otonomi manusia harus dihormati.

Otonomi manusia juga dapat dibatasi oleh ketidakmampuan individu untuk memahami informasi atau membuat keputusan yang tepat. Seperti yang dikemukakan Shin dan Park, algoritma "tidak memiliki affordance yang memungkinkan pengguna untuk memahaminya atau cara terbaik untuk memanfaatkannya guna mencapai tujuan mereka". Dengan demikian, isu utama yang diidentifikasi dalam perdebatan mengenai otonomi pengguna adalah kesulitan untuk mencapai keseimbangan yang tepat antara pengambilan keputusan pengguna sendiri dan keputusan yang mereka delegasikan kepada algoritma. Hal ini semakin rumit karena kurangnya transparansi atas proses pengambilan keputusan yang mendelegasikan keputusan-keputusan tertentu kepada algoritma. Ananny dan Crawford (2018) mencatat bahwa seringkali proses ini tidak memperhitungkan semua pemangku kepentingan, dan bukannya tanpa ketimpangan struktural.

Sebagai sebuah metode Penelitian dan Inovasi yang Bertanggung Jawab (RRI), 'desain partisipatif' seringkali disebut-sebut karena fokusnya pada desain algoritma untuk mempromosikan nilai-nilai pengguna akhir dan melindungi otonomi mereka. Desain partisipatif bertujuan untuk "membawa pengetahuan tacit dan pengalaman yang diwujudkan peserta ke dalam proses desain". Misalnya, kerangka kerja konseptual 'Society-in-the-Loop' (2018) karya Rahwan berupaya untuk memungkinkan berbagai pemangku kepentingan dalam masyarakat merancang sistem algoritmik sebelum penerapan, dan untuk mengubah serta membalikkan keputusan sistem algoritmik yang sudah mendasari aktivitas sosial.

Kerangka kerja ini bertujuan untuk menjaga "kontrak sosial algoritmik" yang berfungsi dengan baik, yang didefinisikan sebagai "sebuah perjanjian antara berbagai pemangku kepentingan manusia, yang dimediasi oleh mesin". Kerangka kerja ini mencapai hal ini dengan mengidentifikasi dan menegosiasikan nilai-nilai dari berbagai pemangku kepentingan yang

terdampak oleh sistem algoritmik sebagai dasar untuk memantau kepatuhan terhadap kontrak sosial.

Privasi informasi berkaitan erat dengan otonomi pengguna. Privasi informasi menjamin kebebasan berpikir, berkomunikasi, dan menjalin hubungan, di antara aktivitas esensial manusia lainnya. Namun, meningkatnya interaksi manusia dengan sistem algoritmik telah secara efektif mengurangi kemampuan mereka untuk mengontrol siapa yang memiliki akses ke informasi yang menjadi perhatian mereka dan apa yang dilakukan dengannya. Banyaknya data sensitif yang diperlukan dalam pembuatan profil dan prediksi algoritmik, yang merupakan inti dari sistem rekomendasi, menimbulkan berbagai masalah terkait privasi informasi individu.

Pembuatan profil algoritmik berlangsung dalam jangka waktu yang tidak terbatas, di mana individu dikategorikan berdasarkan logika internal sistem, dan profil mereka diperbarui seiring diperolehnya informasi baru tentang mereka. Informasi ini biasanya diperoleh secara langsung, dari saat seseorang berinteraksi dengan sistem tertentu, atau secara tidak langsung, disimpulkan dari kelompok individu yang disusun secara algoritmik. Memang, pembuatan profil algoritmik juga akan bergantung pada informasi yang dikumpulkan tentang individu dan kelompok orang lain yang telah dikategorikan dengan cara yang serupa dengan orang yang ditargetkan. Ini mencakup informasi mulai dari karakteristik seperti lokasi geografis dan usia, hingga informasi tentang perilaku dan preferensi tertentu, termasuk jenis konten apa yang kemungkinan besar paling dicari seseorang di platform tertentu. Meskipun hal ini menimbulkan masalah *bukti yang tidak meyakinkan*, hal ini juga menunjukkan bahwa jika privasi kelompok tidak terjamin, mungkin mustahil bagi individu untuk melepaskan diri dari proses pembuatan profil algoritmik dan prediksi. Dengan kata lain, privasi informasi individu tidak dapat diamankan tanpa mengamankan privasi kelompok.

Pengguna mungkin tidak selalu menyadari, atau mungkin tidak memiliki kemampuan untuk memperoleh kesadaran tentang, jenis informasi yang disimpan tentang mereka dan untuk apa informasi tersebut digunakan. Mengingat sistem rekomendasi berkontribusi pada konstruksi dinamis identitas individu dengan mengintervensi pilihan mereka, kurangnya kendali atas informasi seseorang mengakibatkan hilangnya otonomi.

Memberikan individu kemampuan untuk berkontribusi pada desain sistem rekomendasi dapat membantu menciptakan profil yang lebih akurat yang memperhitungkan atribut dan kategori sosial yang mungkin tidak akan dimasukkan dalam pelabelan yang digunakan sistem untuk mengkategorikan pengguna. Meskipun keinginan untuk meningkatkan pembuatan profil algoritmik akan bervariasi tergantung konteksnya, peningkatan desain algoritmik dengan memasukkan umpan balik dari berbagai pemangku kepentingan algoritma sejalan dengan kajian yang telah disebutkan sebelumnya tentang RRI dan meningkatkan kemampuan pengguna untuk menentukan nasib sendiri.

Pengetahuan tentang siapa pemilik data seseorang dan apa yang dilakukan dengannya juga dapat membantu menginformasikan pertimbangan antara privasi informasi dan manfaat pemrosesan informasi. Misalnya, dalam konteks medis, individu cenderung lebih bersedia berbagi informasi yang dapat membantu menginformasikan diagnostik mereka sendiri, atau orang lain, namun kurang demikian dalam konteks rekrutmen kerja.

Norma koordinasi informasi, sebagaimana dikemukakan Sloan dan Warner (2018), dapat berfungsi untuk memastikan bahwa trade-off ini beradaptasi dengan tepat pada konteks yang berbeda dan tidak memberikan tanggung jawab dan upaya yang berlebihan pada individu tertentu. Misalnya, informasi pribadi seharusnya mengalir secara berbeda dalam konteks prosedur penegakan hukum dibandingkan dengan proses rekrutmen kerja. Peraturan Perlindungan Data Umum Uni Eropa telah memainkan peran penting dalam melembagakan dasar norma-norma tersebut.

Terakhir, semakin banyaknya kajian tentang privasi diferensial menyediakan metode perlindungan privasi baru bagi organisasi yang ingin melindungi privasi pengguna mereka sekaligus menjaga kualitas model yang baik, serta biaya dan kompleksitas perangkat lunak yang dapat dikelola, sehingga mencapai keseimbangan antara utilitas dan privasi. Kemajuan teknis semacam ini, yang memungkinkan organisasi untuk membagikan kumpulan data secara publik sambil menjaga kerahasiaan informasi tentang individu (mencegah identifikasi ulang), dan dapat memastikan perlindungan privasi yang dapat dibuktikan pada data sensitif, seperti data genomik. Bahkan, privasi diferensial baru-baru ini digunakan oleh Social Science One dan Facebook untuk merilis salah satu kumpulan data terbesar (38 juta URL yang dibagikan secara publik di Facebook) secara aman untuk penelitian akademis tentang dampak sosial media.

7.8 KETERTELUSSURAN YANG MENGARAH PADA TANGGUNG JAWAB MORAL

Keterbatasan teknis berbagai algoritma pembelajaran mesin, seperti kurangnya transparansi dan kurangnya kemampuan menjelaskan, melemahkan kemampuan pengawasannya dan menyoroti perlunya pendekatan baru untuk menelusuri tanggung jawab moral dan akuntabilitas atas tindakan yang dilakukan oleh algoritma pembelajaran mesin. Mengenai tanggung jawab moral, mencatat adanya ketidakjelasan umum antara keterbatasan teknis algoritma dan batasan hukum, etika, dan kelembagaan yang lebih luas di mana mereka beroperasi. Bahkan untuk algoritma non-pembelajaran, konsepsi tanggung jawab linear tradisional terbukti menawarkan panduan yang terbatas dalam konteks sosioteknis kontemporer. Struktur sosioteknis yang lebih luas menyulitkan penelusuran kembali tanggung jawab atas tindakan yang dilakukan oleh sistem hibrida terdistribusi yang terdiri dari agen manusia dan buatan.

Selain itu, karena struktur dan operasi pasar pialang data, dalam banyak kasus mustahil untuk "melacak data apa pun ke sumber aslinya" setelah diperkenalkan ke pasar. Alasannya antara lain perlindungan rahasia dagang; pasar kompleks yang "memisahkan" proses pengumpulan data dari proses jual beli; dan campuran informasi yang dihasilkan secara komputasional dalam jumlah besar tanpa sumber empiris yang 'nyata' yang dikombinasikan dengan data asli.

Kompleksitas teknis dan dinamisme algoritma pembelajaran mesin (ML) membuatnya rentan terhadap kekhawatiran "pencucian agensi": kesalahan moral yang terdiri dari menjauhkan diri dari tindakan yang secara moral mencurigakan, terlepas dari apakah tindakan tersebut disengaja atau tidak, dengan menyalahkan algoritma. Hal ini dipraktikkan oleh

organisasi maupun individu. memberikan contoh yang lugas dan mengerikan tentang pencucian uang agensi oleh Facebook:

Dengan menggunakan sistem otomatis Facebook, tim ProPublica menemukan kategori buatan pengguna yang disebut "pembenci Yahudi" dengan lebih dari 2.200 anggota. Untuk membantu ProPublica menemukan audiens yang lebih luas (dan dengan demikian mendapatkan pembelian iklan yang lebih baik), Facebook menyarankan sejumlah kategori tambahan. ProPublica menggunakan platform tersebut untuk memilih profil lain yang menampilkan kategori anti-Semit, dan Facebook menyetujui iklan ProPublica dengan sedikit perubahan. Ketika ProPublica mengungkapkan kategori anti-Semit dan outlet berita lain melaporkan kategori serupa yang menjijikkan, Facebook menanggapi dengan menjelaskan bahwa algoritma telah membuat kategori berdasarkan respons pengguna terhadap bidang target dan bahwa "kami tidak pernah bermaksud atau mengantisipasi fungsi ini digunakan dengan cara ini".

Saat ini, kegagalan memahami dampak tak terduga dari pemrosesan dan komersialisasi data pribadi massal, sebuah masalah yang lazim dalam sejarah teknologi, dibarengi dengan terbatasnya penjelasan yang diberikan sebagian besar algoritma pembelajaran mesin (ML). Pendekatan ini berisiko mengutamakan penghindaran tanggung jawab melalui penyangkalan "kata komputer begitu". Hal ini dapat menyebabkan para pakar lapangan, seperti dokter, menghindari mempertanyakan saran suatu algoritma meskipun mungkin tampak aneh bagi mereka. Interaksi antara pakar lapangan dan algoritma ML dapat memicu "keburukan epistemik", seperti dogmatisme atau mudah tertipu, dan menghambat atribusi tanggung jawab dalam sistem terdistribusi.

Untuk mengatasi masalah ini, analisis Shah (2018) menekankan bahwa risiko bahwa beberapa pemangku kepentingan dapat melanggar tanggung jawab mereka dapat diatasi, misalnya, dengan membentuk badan terpisah untuk pengawasan etika algoritma (misalnya, DeepMind Health membentuk Panel Peninjau Independen dengan akses tak terbatas ke perusahaan tersebut hingga Google menghentikannya pada tahun 2019).

Namun, mengharapkan badan pengawas tunggal, seperti komite etika penelitian atau dewan peninjau institusional, untuk "bertanggung jawab penuh dalam memastikan ketelitian, kegunaan, dan integritas data besar" tidaklah realistis. Bahkan, beberapa pihak berpendapat bahwa inisiatif-inisiatif ini kurang konsisten dan justru dapat mengarah pada "ethics bluwashing", yang dipahami sebagai

penerapan langkah-langkah dangkal yang mengutamakan nilai-nilai etika dan manfaat dari proses, produk, layanan, atau solusi digital lainnya agar tampak lebih etis secara digital daripada yang sebenarnya.

Ketika dihadapkan dengan rezim hukum yang ketat, para pelaku yang memiliki sumber daya juga dapat melakukan apa yang disebut "ethics dumping", yaitu mengeksport "proses, produk,

atau layanan” yang tidak etis ke negara-negara dengan kerangka kerja dan mekanisme penegakan hukum yang lebih lemah, dan kemudian hasil dari aktivitas tidak etis tersebut “diimpor kembali”.

Terdapat sejumlah pendekatan terperinci untuk membangun akuntabilitas algoritmik dalam literatur yang diulas. Meskipun algoritma pembelajaran mesin (ML) memang memerlukan intervensi teknis untuk meningkatkan keterjelasannya, sebagian besar pendekatan berfokus pada intervensi normatif. Misalnya, Ananny dan Crawford berpendapat bahwa, setidaknya, penyedia algoritma seharusnya memfasilitasi wacana publik tentang teknologi mereka. Demikian pula, untuk mengatasi masalah tindakan etis *ad hoc*, beberapa pihak berpendapat bahwa akuntabilitas pertama-tama harus ditangani sebagai masalah konvensi.

Dalam upaya mengisi "kesenjangan" konvensi tersebut, meminjam prinsip tujuh prinsip algoritma yang ditetapkan oleh *Association for Computing Machinery* (ACM), dengan menyatakan bahwa melalui, *antara lain*, kesadaran akan algoritma, validasi, dan pengujian, suatu organisasi harus bertanggung jawab atas algoritma mereka terlepas dari seberapa buruknya algoritma tersebut. Keputusan mengenai penerapan algoritma harus mempertimbangkan faktor-faktor seperti keinginan dan konteks yang lebih luas di mana algoritma akan beroperasi, yang kemudian akan mengarah pada "budaya algoritma" yang lebih akuntabel. Untuk menangkap pertimbangan tersebut, "forum dan proses interaktif dan diskursif" dengan para pemangku kepentingan yang relevan, seperti yang disarankan oleh Buhmann dkk., dapat terbukti menjadi cara yang berguna.

Senada dengan itu, Binns (2018b) berfokus pada konsep "akal publik" yang bersifat politis-filosofis. Menimbang bahwa proses untuk menetapkan tanggung jawab atas tindakan suatu algoritma berbeda, baik dalam sifat maupun cakupannya, di sektor publik versus swasta, Binns menyerukan pembentukan kerangka kerja yang dibagikan secara publik, yang menurutnya keputusan algoritmik harus mampu bertahan pada tingkat pengawasan publik yang sama dengan yang akan diterima oleh pengambilan keputusan manusia. Pendekatan ini telah digaungkan oleh banyak pihak lain dalam literatur yang diulas.

Permasalahan terkait 'pencucian agensi' dan 'pengabaian etika' muncul dari ketidakcukupan kerangka kerja konseptual yang ada untuk melacak dan menetapkan tanggung jawab moral. Sebagaimana ditunjukkan Floridi, ketika mempertimbangkan sistem algoritmik dan dampak tindakannya,

kita berhadapan dengan DMA [tindakan moral terdistribusi] yang muncul dari interaksi netral moral dari jaringan agen (yang berpotensi hibrida)? Dengan kata lain, siapa yang bertanggung jawab (*tanggung jawab moral terdistribusi, DMR*) atas DMA?

Analisis Floridi menyarankan untuk menetapkan tanggung jawab moral penuh "secara default dan mutlak" kepada *semua* agen dalam jaringan yang secara kausal relevan dengan tindakan jaringan tersebut. Pendekatan yang diusulkan didasarkan pada konsep back-propagation dari

teori jaringan, tanggung jawab mutlak dari yurisprudensi, dan pengetahuan umum dari logika epistemik. Khususnya, pendekatan ini memisahkan tanggung jawab moral dari intensionalitas para aktor dan dari gagasan hukuman dan imbalan atas tindakan tertentu, untuk berfokus pada kebutuhan untuk memperbaiki kesalahan (back-propagation) dan meningkatkan etika kerja semua agen dalam jaringan.

7.9 KESIMPULAN

Artikel ini mengembangkan dan memperbarui penelitian sebelumnya yang dilakukan oleh kelompok kami untuk meninjau literatur relevan yang diterbitkan sejak 2016 tentang etika algoritma. Meskipun artikel tersebut kini tak terelakkan lagi ketinggalan zaman dalam hal referensi spesifik dan informasi detail tentang literatur yang diulas, peta tersebut, dan enam kategori yang disediakan, telah teruji oleh waktu dan tetap menjadi alat yang berharga untuk mengkaji etika algoritma sebagai suatu bidang penelitian, dengan semakin banyaknya literatur yang berfokus pada masing-masing dari enam kategori tersebut yang berkontribusi untuk menyempurnakan pemahaman kita tentang permasalahan yang ada atau memberikan solusi untuk mengatasinya.

Sejak 2016, etika algoritma telah menjadi topik diskusi utama di antara para akademisi, penyedia teknologi, dan pembuat kebijakan. Perdebatan ini semakin menarik perhatian juga karena apa yang disebut "musim panas AI", dan seiring dengan itu meluasnya penggunaan algoritma ML. Banyak pertanyaan etika yang dianalisis dalam artikel ini dan tinjauan pustakanya telah dibahas dalam pedoman dan prinsip etika nasional dan internasional, seperti Kelompok Etika Eropa dalam Sains dan Teknologi dari Komisi Eropa yang telah disebutkan sebelumnya, Komite Kecerdasan Buatan Dewan Perwakilan Rakyat Inggris, dan prinsip-prinsip OECD tentang AI.

Salah satu aspek yang tidak secara eksplisit tercakup dalam peta asli, dan yang menjadi poin utama diskusi dalam literatur terkait, adalah meningkatnya fokus pada penggunaan algoritma, AI, dan teknologi digital secara lebih luas, untuk menghasilkan hasil yang baik secara sosial. Meskipun benar, setidaknya secara prinsip, bahwa setiap inisiatif yang bertujuan menggunakan algoritma untuk kebaikan sosial harus mengatasi risiko yang diidentifikasi oleh masing-masing dari enam kategori dalam peta secara memuaskan, terdapat pula perdebatan yang berkembang mengenai prinsip dan kriteria yang harus menginformasikan desain dan tata kelola algoritma, dan teknologi digital secara lebih luas, untuk tujuan kebaikan sosial yang eksplisit.

Analisis etis diperlukan untuk memitigasi risiko sekaligus memanfaatkan potensi kebaikan dari teknologi-teknologi ini, sejauh analisis tersebut melayani dua tujuan, yaitu mengklarifikasi sifat risiko etis dan potensi kebaikan algoritma dan teknologi digital, serta menerjemahkan pemahaman ini menjadi panduan yang baik dan dapat ditindaklanjuti untuk tata kelola desain dan penggunaan artefak digital.

BAB 8

FAKTOR ESENSIAL MERANCANG AI UNTUK KEBAIKAN SOSIAL

Gagasan mengenai Artificial Intelligence for Social Good (AI4SG), atau Kecerdasan Buatan untuk Kebaikan Sosial, semakin mendapatkan perhatian luas, baik dalam masyarakat informasi secara umum maupun di kalangan komunitas pengembang AI secara khusus. Konsep ini menawarkan potensi besar dalam menyelesaikan berbagai permasalahan sosial melalui penerapan solusi berbasis kecerdasan buatan. Meskipun demikian, masih terdapat sejumlah tantangan mendasar yang perlu diatasi, terutama dalam hal pemahaman teoritis mengenai karakteristik AI yang benar-benar memberikan manfaat sosial, definisi operasional AI4SG dalam praktik, serta strategi untuk mereplikasi keberhasilan awalnya dalam bentuk kebijakan yang efektif.

Bab ini bertujuan untuk membahas kesenjangan tersebut dengan mengidentifikasi tujuh faktor etika utama yang dianggap penting dalam pengembangan dan implementasi inisiatif AI4SG di masa depan. Analisis ini didasarkan pada studi terhadap 27 proyek AI4SG yang telah dilaksanakan, yang mencerminkan keberagaman pendekatan dan dampak yang dihasilkan. Beberapa dari faktor etika ini merupakan hal baru dalam konteks pengembangan AI, sementara yang lainnya semakin relevan seiring dengan kemajuan teknologi ini. Melalui identifikasi faktor-faktor tersebut, dirumuskan pula praktik-praktik terbaik yang dapat dijadikan pedoman awal dalam merancang sistem AI yang tidak hanya efektif, tetapi juga memberikan kontribusi nyata bagi kebaikan sosial. Dengan mempertimbangkan konteks dan keseimbangan etis yang tepat, pedoman ini diharapkan dapat meningkatkan kemungkinan tercapainya manfaat maksimal dari penggunaan AI dalam ranah sosial.

8.1 PENDAHULUAN

Ide "Kecerdasan Buatan (AI) untuk Kebaikan Sosial" (selanjutnya disebut AI4SG) semakin populer di banyak masyarakat informasi dan mendapatkan daya tarik dalam komunitas AI. Proyek-proyek yang berupaya menggunakan AI untuk kebaikan sosial sangat beragam. Proyek-proyek tersebut berkisar dari model untuk memprediksi syok septik hingga model teori permainan untuk mencegah perburuan liar; dari pembelajaran penguatan daring untuk menargetkan pendidikan HIV pada remaja tunawisma hingga model probabilistik untuk mencegah praktik kepolisian yang merugikan dan mendukung retensi siswa. Memang, aplikasi baru AI4SG muncul hampir setiap hari, memungkinkan hasil yang baik secara sosial yang dulunya sulit dicapai, tidak layak, atau tidak terjangkau.

Beberapa kerangka kerja untuk perancangan, pengembangan, dan penerapan AI etis secara umum telah muncul baru-baru ini. Namun, pemahaman tentang apa yang dimaksud dengan AI "untuk kebaikan sosial" masih terbatas. Pendekatan ad hoc AI4SG, dengan menganalisis area aplikasi spesifik seperti penanggulangan bencana atau penanggulangan bencana seperti yang telah dilakukan dalam pertemuan tahunan industri dan pemerintah AI,

menunjukkan adanya fenomena tersebut, tetapi tidak menjelaskannya, juga tidak menyarankan bagaimana solusi AI4SG lainnya dapat dan seharusnya dirancang untuk memanfaatkan potensi AI secara maksimal. Lebih lanjut, banyak proyek yang menghasilkan dampak positif secara sosial menggunakan AI tidak (dijelaskan sendiri) sebagai proyek tersebut.

Kurangnya pemahaman yang jelas tentang apa yang membuat AI baik secara sosial secara teori, apa yang dapat digambarkan sebagai AI4SG dalam praktik, dan bagaimana mereproduksi keberhasilan awalnya dalam hal kebijakan merupakan masalah karena para perancang AI4SG menghadapi setidaknya dua tantangan utama: kegagalan yang tidak perlu dan hilangnya peluang. Perangkat lunak AI dibentuk oleh nilai-nilai kemanusiaan yang, jika tidak dipilih dengan cermat, dapat mengarah pada skenario "AI baik yang berakhir buruk".

Misalnya, pertimbangkan kegagalan perangkat lunak pendukung onkologi IBM, yang mencoba menggunakan pembelajaran mesin untuk mengidentifikasi tumor kanker, tetapi ditolak oleh praktisi medis "di lapangan". Sistem ini dilatih menggunakan data sintetis dan tidak cukup disempurnakan untuk menafsirkan rekam medis pasien yang ambigu, bernuansa, atau "berantakan". Sistem ini juga bergantung pada protokol medis AS, yang tidak berlaku di seluruh dunia. Penerapan yang ceroboh dan desain perangkat lunak yang buruk menyebabkan kesalahan diagnosis dan saran pengobatan yang keliru, sehingga merusak kepercayaan para dokter. Desain dan penerapan yang spesifik konteks dapat membantu mencegah ketidakselarasan nilai tersebut dan menghasilkan proyek AI4SG yang sukses secara lebih konsisten.

Di saat yang sama, hasil AI yang benar-benar baik secara sosial mungkin muncul hanya secara kebetulan, misalnya melalui penerapan solusi AI secara tidak sengaja dalam konteks yang berbeda. Hal ini terjadi dengan penggunaan versi sistem kognitif IBM yang berbeda. Dalam hal ini, sistem Watson awalnya dirancang untuk mengidentifikasi mekanisme biologis, tetapi ketika digunakan di ruang kelas, sistem ini menginspirasi mahasiswa teknik untuk memecahkan masalah desain. Dalam hal ini, AI menyediakan mode pendidikan yang unik. Namun, kurangnya pemahaman yang jelas tentang AI4SG berarti bahwa keberhasilan ini bersifat kebetulan dan tidak dapat diulang secara sistematis, sementara untuk setiap "keberhasilan yang tidak disengaja" mungkin terdapat banyak contoh peluang yang terlewatkan untuk memanfaatkan manfaat AI dalam memajukan hasil yang baik secara sosial dalam berbagai pengaturan.

Untuk menghindari kegagalan yang tidak perlu dan peluang yang terlewatkan, AI4SG akan mendapat manfaat dari analisis faktor-faktor penting yang mendukung dan menjamin keberhasilan perancangan dan penerapan AI4SG. Dalam artikel ini, kami menyajikan analisis pertama yang terperinci mengenai faktor-faktor ini. Tujuan kami di sini bukanlah untuk mendokumentasikan setiap pertimbangan etis dalam sebuah proyek AI4SG. Misalnya, sangatlah penting, dan semoga terbukti dengan sendirinya, bahwa sebuah proyek AI4SG tidak boleh memajukan proliferasi senjata pemusnah massal, sebuah keharusan yang tidak akan kami bahas di sini. Demikian pula, penting untuk mengakui sejak awal bahwa terdapat banyak keadaan di mana AI bukanlah cara yang paling efektif untuk mengatasi masalah sosial tertentu.

Hal ini dapat disebabkan oleh adanya pendekatan alternatif yang lebih efektif (yaitu, "Bukan AI untuk Kebaikan Sosial") atau karena risiko yang tidak dapat diterima yang akan ditimbulkan oleh penerapan AI (yaitu, "AI untuk Kebaikan Sosial yang Tidak Memadai" jika dibandingkan dengan risikonya). Kami juga tidak memperkirakan banyak (atau mungkin ada) kasus di mana AI merupakan "solusi ajaib" solusi tunggal untuk masalah sosial yang mengakar (yaitu, "Hanya AI untuk Kebaikan Sosial").

Oleh karena itu, yang esensial tentang faktor-faktor dan praktik terbaik yang sesuai bukanlah penggabungannya dalam setiap keadaan; kami mencatat beberapa contoh di mana secara moral dapat dipertahankan untuk tidak menggabungkan faktor tertentu. Sebaliknya, yang esensial adalah bahwa setiap praktik terbaik (i) dipertimbangkan secara proaktif, dan (ii) tidak digabungkan jika dan hanya jika ada alasan yang jelas, dapat dibuktikan, dan dapat dipertahankan secara moral mengapa hal itu tidak boleh dilakukan.

Dalam artikel ini, kami berfokus pada identifikasi faktor-faktor yang khususnya relevan dengan AI sebagai infrastruktur teknologi, sejauh ia dirancang dan digunakan untuk kemajuan kebaikan sosial. Untuk mengantisipasinya, ketujuh faktor ini adalah: (1) falsifikasi dan penerapan inkremental; (2) perlindungan terhadap manipulasi prediktor; (3) intervensi yang dikontekstualisasikan oleh penerima; (4) penjelasan kontekstual penerima dan tujuan yang transparan; (5) perlindungan privasi dan persetujuan subjek data; (6) keadilan situasional; dan (7) semantikasi yang ramah manusia. Dengan faktor-faktor ini teridentifikasi, pertanyaan yang mungkin muncul adalah: bagaimana faktor-faktor ini seharusnya dievaluasi dan diselesaikan, oleh siapa, dan dengan mekanisme pendukung apa (misalnya regulasi atau kode etik).

Pertanyaan-pertanyaan ini, yang tidak berada dalam cakupan artikel ini dan akan dibahas pada tahap selanjutnya dari penelitian ini, terkait erat dengan isu-isu etika dan politik yang lebih luas mengenai legitimasi pengambilan keputusan dengan, dan tentang, AI. Sisa artikel ini disusun sebagai berikut. Pada Bagian 8.2, kami menjelaskan bagaimana kami mengidentifikasi ketujuh faktor tersebut. Pada Bagian 8.3, kami menganalisis ketujuh faktor tersebut secara individual. Kami menjelaskan masing-masing faktor dengan mengacu pada satu atau lebih studi kasus, dan kami menyimpulkan dari setiap faktor praktik terbaik yang sesuai untuk diikuti oleh para kreator AI4SG. Pada bagian penutup, kami membahas faktor-faktor tersebut dan menyarankan bagaimana ketegangan di antara faktor-faktor tersebut dapat diselesaikan.

8.2 METODOLOGI

Inisiatif AI4SG berhasil sejauh membantu mengurangi, memitigasi, atau memberantas suatu masalah moral tertentu. Dengan demikian, analisis kami tentang faktor-faktor penting keberhasilan AI4SG didasarkan pada definisi kerja berikut:

AI4SG $\frac{1}{4}$ def. perancangan, pengembangan, dan penerapan sistem AI dengan cara yang (i) mencegah, memitigasi, atau menyelesaikan masalah yang berdampak buruk terhadap kehidupan manusia dan/atau kesejahteraan alam, dan/atau (ii)

memungkinkan pembangunan yang lebih disukai secara sosial dan/atau berkelanjutan secara lingkungan.

Berdasarkan definisi ini, kami menganalisis 27 proyek, yang diperoleh melalui tinjauan sistematis literatur relevan yang dilakukan oleh para penulis, untuk mengidentifikasi kasus-kasus yang jelas dan signifikan dari contoh-contoh AI4SG yang berhasil dan tidak berhasil. Analisis literatur yang mendasari artikel ini melibatkan pencarian di lima basis data (Google Scholar, PhilPapers, Scopus, SSRN, dan Web of Science) antara Oktober 2018 dan Mei 2019.

Awalnya, kami melakukan pencarian luas untuk AI for Social Good di masing-masing mesin pencari ini. Pencarian umum ini menghasilkan banyak hasil tentang penerapan AI untuk kebaikan. Oleh karena itu, kami mencari penggunaan AI di bidang-bidang yang berkaitan dengan kehidupan manusia dan kesejahteraan alam, seperti 'perawatan kesehatan', 'pendidikan', 'kesetaraan', 'perubahan iklim', dan 'perlindungan lingkungan'. Pencarian ini menghasilkan kata kunci yang terpisah, yang darinya kami memperoleh sinonim terpilih untuk melakukan pencarian spesifik area. Setiap pencarian di area penelitian menggunakan kueri: < area dan sinonim> DAN ("Kecerdasan Buatan" ATAU "Pembelajaran Mesin" ATAU "AI") DAN "Kebajikan Sosial". Dari 27 kasus yang diteliti, kami mengidentifikasi 7 kasus (lihat Lampiran untuk daftarnya) yang paling representatif dalam hal cakupan, variasi, dampak, dan potensinya untuk menguatkan faktor-faktor esensial yang kami anggap seharusnya menjadi ciri desain proyek AI4SG.

Faktor-faktor yang dijelaskan dalam artikel ini telah diidentifikasi secara koheren dengan karya yang lebih umum di bidang etika AI. Setiap faktor berkaitan dengan setidaknya satu dari lima prinsip etika AI beneficence, nonmaleficence, justice, autonomy, dan explicability yang diidentifikasi dalam analisis komparatif yang disebutkan di atas. Koherensi ini krusial: AI4SG tidak boleh bertentangan dengan kerangka etika yang memandu desain dan evaluasi AI secara umum. Prinsip beneficence memiliki relevansi khusus.

Dinyatakan bahwa penggunaan AI harus memberikan manfaat bagi manusia dan alam, dan memang proyek AI4SG tidak hanya harus mematuhi tetapi juga mewujudkan prinsip ini, sehingga manfaat AI4SG harus lebih baik dan berkelanjutan, sejalan dengan definisi di atas. Dengan demikian, kemurahan hati merupakan syarat yang diperlukan dari AI4SG, namun itu tidak cukup, paling tidak karena dampak baik dari proyek AI4SG dapat "diimbangi" oleh penciptaan atau amplifikasi risiko atau bahaya lainnya.

Selain itu, sementara prinsip-prinsip etika lainnya, seperti otonomi dan eksplikabilitas, memang berulang di seluruh diskusi kita, faktor-faktor yang kita tunjukkan di bawah ini lebih erat kaitannya dengan pertimbangan desain yang khusus untuk AI4SG, dan dapat dioperasionalkan dalam bentuk praktik terbaik yang sesuai yang disediakan untuk masing-masing. Dengan cara ini, analisis etika yang menginformasikan desain dan penerapan inisiatif AI4SG memiliki peran sentral dalam mengurangi risiko yang dapat diperkirakan dari konsekuensi yang tidak diinginkan dan kemungkinan penyalahgunaan teknologi.

Sebelum membahas faktor-faktor tersebut, penting untuk mengklarifikasi tiga fitur umum dari keseluruhan rangkaian: ketergantungan, keteraturan, dan koherensi. Ketujuh

faktor tersebut seringkali saling terkait dan saling bergantung, tetapi demi kesederhanaan, kami membahasnya secara terpisah. Tidak ada yang boleh disimpulkan dari pilihan ini. Dengan cara yang sama, semua faktor tersebut esensial, tidak ada yang "lebih penting" daripada yang lain, jadi kami akan memperkenalkannya bukan berdasarkan prioritas, tetapi agak historis, dimulai dengan faktor-faktor yang mendahului AI, namun menjadi lebih penting ketika teknologi AI digunakan, karena kemampuan dan risiko khusus AI. Ini termasuk falsifiabilitas dan penerapan inkremental serta perlindungan terhadap manipulasi data.

Ada juga faktor-faktor yang lebih terkait secara intrinsik dengan karakteristik sosioteknis AI seperti yang ada saat ini, seperti keadilan situasional dan semantikasi yang ramah manusia. Faktor-faktor tersebut kokoh secara etis dan dapat diterapkan secara pragmatis, dalam artian faktor-faktor tersebut memunculkan pertimbangan desain dalam bentuk praktik terbaik yang harus didukung secara etis. Penting untuk ditegaskan di sini bahwa ketujuh faktor yang kami identifikasi tidak cukup untuk AI yang baik secara sosial, tetapi pertimbangan yang cermat terhadap masing-masing faktor, menurut kami, diperlukan.

Oleh karena itu, serangkaian faktor yang kami identifikasi tidak boleh dianggap sebagai "daftar periksa" yang, jika dipatuhi begitu saja, menjamin hasil yang baik secara sosial dari penggunaan AI dalam domain tertentu. Sejalan dengan itu, kami menyoroti perlunya mencapai keseimbangan antara berbagai faktor, dan bahkan antara ketegangan yang mungkin muncul bahkan dalam satu faktor. Oleh karena itu, upaya untuk membingkai suatu proyek sebagai "untuk kebaikan sosial" atau "bukan untuk kebaikan sosial" secara biner tampaknya reduktif dan tidak perlu, apalagi subjektif. Tujuan artikel ini bukanlah untuk mengidentifikasi, atau menawarkan cara untuk mengidentifikasi proyek-proyek AI4SG; tujuan kami adalah untuk mengidentifikasi karakteristik-karakteristik proyek yang penting secara etis yang secara layak dapat digambarkan sebagai AI4SG.

8.3 TUJUH FAKTOR ESENSIAL UNTUK KEBERHASILAN AI4SG

Sebagaimana telah diantisipasi, faktor-faktor tersebut adalah (1) falsifikasi dan penerapan inkremental; (2) perlindungan terhadap manipulasi prediktor; (3) intervensi yang dikontekstualisasikan oleh penerima; (4) penjelasan yang dikontekstualisasikan oleh penerima dan tujuan yang transparan; (5) perlindungan privasi dan persetujuan subjek data; (6) keadilan situasional; dan (7) semantikasi yang ramah manusia. Kami akan menjelaskan setiap faktor dengan satu atau lebih contoh proyek dalam contoh ini, dan menawarkan praktik terbaik yang sesuai.

Falsifikasi dan Penerapan Inkremental

Kepercayaan sangat penting bagi teknologi secara umum, dan khususnya bagi aplikasi AI4SG, agar dapat diadopsi dan memberikan dampak positif yang berarti bagi kehidupan manusia dan kesejahteraan lingkungan. Kepercayaan suatu aplikasi AI mensyaratkan probabilitas tinggi bahwa aplikasi tersebut akan menghormati prinsip kebaikan, atau setidaknya prinsip nonmaleficence. Meskipun tidak ada aturan atau pedoman universal yang dapat memastikan atau menjamin kepercayaan, falsifiabilitas merupakan faktor penting untuk meningkatkan kepercayaan aplikasi teknologi secara umum, dan aplikasi AI4SG secara khusus.

Falsifiabilitas mensyaratkan spesifikasi, dan kemungkinan pengujian empiris, dari satu atau lebih persyaratan kritis, yaitu, kondisi, sumber daya, atau sarana esensial agar suatu kapabilitas dapat beroperasi penuh, sehingga sesuatu dapat atau tidak seharusnya berfungsi tanpanya. Keamanan merupakan persyaratan kritis yang jelas. Oleh karena itu, agar sistem AI4SG dapat dipercaya, keamanannya harus dapat difalsifikasi. Jika falsifiabilitas tidak memungkinkan, maka persyaratan kritis tidak dapat diperiksa, dan sistem tersebut tidak boleh dianggap dapat dipercaya. Inilah sebabnya mengapa falsifiabilitas merupakan faktor penting untuk semua proyek AI4SG yang dapat dibayangkan.

Sayangnya, kita tidak dapat memastikan keamanan suatu aplikasi AI4SG kecuali kita dapat mengujinya dalam semua konteks yang memungkinkan. Dalam hal ini, peta pengujian hanya akan setara dengan wilayah penerapannya. Sebagaimana yang diperjelas oleh *reductio ad absurdum* ini, kepastian yang utuh tidak dapat dicapai. Yang dapat dicapai, dalam dunia yang tidak pasti dan tidak jelas dengan banyak situasi yang tak terduga, adalah kemungkinan untuk mengetahui kapan suatu persyaratan kritis tidak diimplementasikan atau mungkin gagal berfungsi dengan baik. Oleh karena itu, jika persyaratan kritis tersebut dapat difalsifikasi, kita dapat mengetahui kapan aplikasi AI4SG tidak dapat dipercaya, tetapi tidak dapat mengetahui apakah aplikasi tersebut dapat dipercaya.

Persyaratan kritis harus diuji dengan siklus penerapan yang bertahap. Efek berbahaya yang tidak diinginkan mungkin baru terlihat setelah pengujian. Di saat yang sama, perangkat lunak hanya boleh diuji di dunia nyata jika aman untuk dilakukan. Hal ini memerlukan penerapan siklus penerapan di mana pengembang: (a) memastikan bahwa persyaratan atau asumsi paling kritis aplikasi dapat difalsifikasi, (b) melakukan pengujian hipotesis terhadap persyaratan dan asumsi paling kritis tersebut dalam konteks yang aman dan terlindungi, dan, jika hipotesis ini tidak terbukti salah dalam konteks yang sesuai, maka (c) melakukan pengujian di konteks yang semakin luas, dan/atau menguji serangkaian persyaratan yang kurang kritis yang lebih besar, dan semua ini sambil (d) bersiap untuk menghentikan atau memodifikasi penerapan segera setelah efek berbahaya atau efek yang tidak diinginkan lainnya muncul.

Aplikasi AI4SG dapat menggunakan pendekatan formal untuk mencoba menguji persyaratan kritis. Misalnya, pendekatan tersebut dapat mencakup penggunaan verifikasi formal untuk memastikan bahwa kendaraan otonom, dan sistem AI dalam konteks kritis keselamatan lainnya, akan menjadi pilihan yang lebih baik secara etis. Metode semacam itu menawarkan pemeriksaan keamanan yang, dalam hal falsifikasi, dapat dibuktikan kebenarannya. Simulasi dapat menawarkan jaminan yang kurang lebih serupa. Simulasi memungkinkan seseorang untuk menguji apakah persyaratan kritis (sekali lagi, pertimbangkan keselamatan) terpenuhi berdasarkan serangkaian asumsi formal. Tidak seperti pembuktian formal, simulasi tidak selalu dapat menunjukkan bahwa properti yang dipersyaratkan selalu terpenuhi. Namun, simulasi sering kali memungkinkan seseorang untuk menguji serangkaian kasus yang jauh lebih luas yang tidak dapat ditangani secara formal, misalnya, karena kompleksitas pembuktiannya.

Akan keliru jika hanya mengandalkan properti formal atau simulasi untuk memalsukan aplikasi AI4SG. Asumsi model-model ini membatasi penerapan kesimpulan apa pun yang

mungkin dibuat di dunia nyata. Dan asumsi-asumsi tersebut mungkin salah dalam kenyataan. Apa yang mungkin terbukti benar melalui pembuktian formal, atau kemungkinan benar melalui pengujian dalam simulasi, dapat dibantah kemudian dengan penerapan sistem di dunia nyata. Misalnya, pengembang model teori permainan untuk keamanan satwa liar mengasumsikan topografi yang relatif datar tanpa hambatan serius.

Oleh karena itu, perangkat lunak yang mereka kembangkan awalnya memiliki definisi yang salah tentang rute patroli yang optimal. Pengujian aplikasi secara bertahap memungkinkan penyempurnaan rute patroli yang optimal dengan membuktikan salahnya asumsi topografi datar. Jika dilema baru dalam konteks dunia nyata memerlukan perubahan asumsi sebelumnya yang dibuat di lab, salah satu solusinya adalah memperbaiki asumsi apriori setelah penerapan. Alternatifnya, seseorang dapat mengadopsi sistem "*on-the-fly*" atau runtime untuk pembaruan konstan pemrosesan program ("pemahaman") terhadap masukannya. Namun, pendekatan ini juga memiliki banyak masalah.

Misalnya, bot Twitter Microsoft yang terkenal, Tay, memperoleh makna, dalam arti yang sangat longgar, saat runtime, saat ia belajar dari pengguna Twitter bagaimana ia harus menanggapi tweet. Namun, setelah diterapkan di dunia media sosial yang nyata dan seringkali kejam kemampuan bot untuk terus-menerus menyesuaikan "pemahaman konseptualnya" menjadi bug yang disayangkan, karena Tay "mempelajari" dan memuntahkan bahasa yang menyinggung dan asosiasi yang tidak etis antara konsep dari pengguna lain.

Penggunaan pendekatan retrodiktif yaitu, upaya untuk memahami beberapa aspek realitas melalui informasi apriori untuk menangani falsifiabilitas persyaratan menghadirkan masalah serupa. Hal ini patut dicatat, karena retrodiksi merupakan metode utama pendekatan pembelajaran mesin terawasi yang belajar dari data (misalnya, pembelajaran fungsi transformasi berkelanjutan dalam kasus jaringan saraf).

Dari analisis sebelumnya, dapat disimpulkan bahwa faktor esensial falsifikasi dan pengembangan inkremental terdiri dari sebuah siklus: persyaratan rekayasa yang dapat difalsifikasi (sehingga setidaknya memungkinkan untuk mengetahui apakah persyaratan tersebut tidak terpenuhi); pengujian falsifikasi untuk meningkatkan tingkat kepercayaan secara bertahap; penyesuaian asumsi apriori; dan baru kemudian penerapan dalam konteks yang semakin luas dan kritis. Pendekatan Jerman dalam mengatur kendaraan otonom merupakan contoh yang baik dari pendekatan inkremental ini.

Zona yang dideregulasi memungkinkan eksperimen otonomi terbatas dan, setelah meningkatkan tingkat kepercayaan, produsen dapat menguji kendaraan dengan tingkat otonomi yang lebih tinggi. Memang, pembentukan zona deregulasi semacam itu, atau teststrecken, merupakan salah satu rekomendasi untuk mendukung kebijakan AI yang lebih etis di tingkat Eropa. Identifikasi faktor penting ini menghasilkan praktik terbaik berikut:

- 1) Perancang AI4SG harus mengidentifikasi persyaratan yang dapat dipalsukan dan mengujinya secara bertahap dari laboratorium ke "dunia luar".

Perlindungan terhadap Manipulasi Prediktor

Penggunaan AI untuk memprediksi tren atau pola masa depan sangat populer dalam konteks AI4SG, mulai dari penerapan prediksi otomatis untuk memperbaiki kegagalan

akademis, hingga pencegahan pengawasan ilegal, dan pendeteksian penipuan perusahaan. Kekuatan prediktif AI4SG menghadapi dua risiko: manipulasi data masukan, dan ketergantungan berlebihan pada indikator non-kausal.

Manipulasi data bukanlah masalah baru, dan tidak terbatas pada sistem AI saja. Temuan-temuan yang telah mapan seperti Hukum Goodhart, yang sering diringkas sebagai "ketika suatu ukuran menjadi target, ia tidak lagi menjadi ukuran yang baik", telah ada jauh sebelum adopsi sistem AI secara luas. Namun dalam kasus AI, masalah manipulasi data dapat diperburuk dan menyebabkan hasil yang tidak adil yang melanggar prinsip keadilan. Oleh karena itu, hal ini merupakan risiko yang patut diperhatikan bagi setiap inisiatif AI4SG, karena dapat melemahkan daya prediktif AI dan menyebabkan penghindaran intervensi yang baik secara sosial di tingkat individu. Pertimbangkan kekhawatiran yang dikemukakan oleh Ghani mengenai guru yang menghadapi evaluasi terkait:

persentase siswa di kelas mereka yang berada di atas ambang batas risiko tertentu. Jika modelnya transparan misalnya, sangat bergantung pada IPK matematika guru dapat meningkatkan nilai matematika dan mengurangi skor risiko menengah siswa mereka.

Sebagaimana Ghani selanjutnya mengemukakan, kekhawatiran yang sama berlaku untuk prediktor interaksi yang merugikan dengan petugas polisi:

Sistem-sistem ini sangat mudah dipahami dan ditafsirkan, tetapi hal itu juga membuatnya mudah untuk dimanipulasi. Seorang petugas yang telah menggunakan kekerasan dua kali dalam 80 hari terakhir mungkin memilih untuk sedikit lebih berhati-hati selama 10 hari berikutnya, hingga hitungannya kembali ke nol.

Contoh-contoh hipotetis ini memperjelas bahwa, ketika model yang digunakan mudah dipahami "di lapangan", model tersebut sudah terbuka untuk disalahgunakan atau "dimanipulasi", terlepas dari apakah AI digunakan atau tidak. Pengenalan AI memperumit masalah, karena skala penerapan AI biasanya. Seperti yang telah kita lihat, jika informasi yang digunakan untuk memprediksi hasil tertentu diketahui, agen dengan informasi tersebut (yang diprediksi akan mengambil tindakan tertentu) dapat mengubah nilai setiap variabel prediktif untuk menghindari intervensi. Dengan cara ini, daya prediktif keseluruhan model berkurang, seperti yang telah ditunjukkan oleh penelitian empiris dalam domain penipuan perusahaan. Fenomena semacam itu dapat berlanjut, mulai dari deteksi penipuan hingga ke domain yang ingin ditangani oleh inisiatif AI4SG.

Pada saat yang sama, terdapat risiko bahwa ketergantungan yang berlebihan pada indikator non-kausal yaitu, data yang berkorelasi dengan, tetapi bukan kausal, suatu fenomena dapat mengalihkan perhatian dari konteks di mana perancang AI4SG ingin melakukan intervensi. Agar efektif, intervensi semacam itu harus mengubah akar penyebab suatu masalah, seperti masalah rumah tangga mahasiswa atau tata kelola perusahaan yang tidak

memadai, alih-alih prediktor non-kausal. Melakukan sebaliknya berisiko hanya mengatasi gejala, alih-alih akar penyebab suatu masalah.

Risiko-risiko ini menunjukkan perlunya mempertimbangkan penggunaan perlindungan sebagai faktor desain untuk proyek AI4SG. Perlindungan semacam itu dapat membatasi pemilihan indikator yang akan digunakan dalam desain proyek AI4SG; sejauh mana indikator-indikator ini seharusnya membentuk intervensi; dan/atau tingkat transparansi yang seharusnya diterapkan terkait bagaimana indikator memengaruhi keputusan. Hal ini menghasilkan praktik terbaik berikut:

Perancang AI4SG harus mengadopsi perlindungan yang (i) memastikan bahwa indikator non-kausal tidak mendistorsi intervensi secara tidak tepat, dan (ii) membatasi, jika sesuai, pengetahuan tentang bagaimana masukan memengaruhi keluaran dari sistem AI4SG, untuk mencegah manipulasi.

Intervensi yang Dikontekstualisasikan Penerima

Sangat penting bahwa perangkat lunak mengintervensi kehidupan pengguna hanya dengan cara yang menghormati otonomi mereka. Sekali lagi, ini bukan masalah yang hanya muncul dengan intervensi yang digerakkan oleh AI, tetapi penggunaan AI memperkenalkan pertimbangan baru. Secara khusus, tantangan inti untuk proyek AI4SG adalah merancang intervensi yang menyeimbangkan manfaat saat ini dan masa depan.

Masalah keseimbangan, yang lazim dalam penelitian penggalian preferensi, bermuara pada saling ketergantungan pilihan temporal. Intervensi saat ini dapat memunculkan preferensi pengguna yang kemudian memungkinkan perangkat lunak untuk mengontekstualisasikan intervensi di masa mendatang kepada pengguna tertentu. Akibatnya, strategi intervensi yang tidak berdampak pada otonomi pengguna (misalnya, strategi yang tidak memiliki intervensi apa pun) mungkin tidak efektif dalam mengekstraksi informasi yang diperlukan untuk intervensi di masa mendatang yang dikontekstualisasikan dengan benar. Sebaliknya, intervensi yang terlalu melanggar otonomi pengguna dapat menyebabkan pengguna menolak teknologi, sehingga intervensi di masa mendatang menjadi mustahil.

Pertimbangan keseimbangan ini umum untuk inisiatif AI4SG. Ambil contoh, perangkat lunak pengenalan aktivitas interaktif untuk penyandang disabilitas kognitif. Perangkat lunak ini dirancang untuk mendorong pasien agar mempertahankan jadwal aktivitas harian (misalnya, minum obat), sekaligus meminimalkan gangguan terhadap tujuan mereka yang lebih luas. Setiap intervensi dikontekstualisasikan sedemikian rupa sehingga perangkat lunak mempelajari waktu intervensi di masa mendatang dari respons terhadap intervensi sebelumnya. Lebih lanjut, hanya intervensi penting yang dilakukan, namun semua intervensi sebagian bersifat opsional karena menolak satu permintaan akan menyebabkan permintaan yang sama di kemudian hari.

Di sini, kekhawatirannya adalah pasien akan menolak teknologi yang terlalu mengganggu; oleh karena itu dicari keseimbangan. Keseimbangan ini tidak ada dalam contoh kedua kami. Sebuah aplikasi teori permainan mengintervensi patroli petugas keamanan satwa

liar dengan menawarkan rute yang disarankan. Namun, jika suatu rute menimbulkan hambatan fisik, perangkat lunak tidak memiliki kemungkinan untuk memberikan saran alternatif. Petugas mungkin mengabaikan saran tersebut dengan mengambil rute yang berbeda, tetapi tidak tanpa keluar dari aplikasi.

Sangat penting untuk melonggarkan batasan tersebut, sehingga pengguna dapat mengabaikan intervensi, tetapi menerima intervensi selanjutnya yang lebih tepat (dalam bentuk saran) di kemudian hari. Contoh-contoh ini menunjukkan pentingnya memandang pengguna sebagai mitra yang setara dalam perancangan dan penerapan sistem pengambilan keputusan otonom. Penerapan pola pikir ini mungkin telah membantu mencegah hilangnya dua pesawat Boeing 737 Max secara tragis. Tampaknya para pilot penerbangan ini kesulitan memulihkan kerusakan perangkat lunak yang disebabkan oleh sensor yang rusak, sebagian karena tidak adanya "fitur keselamatan opsional" yang dijual terpisah oleh Boeing.

Risiko positif palsu (intervensi yang tidak perlu, menciptakan kekecewaan) seringkali sama bermasalahnya dengan negatif palsu (tidak ada intervensi padahal diperlukan, sehingga membatasi efektivitas). Oleh karena itu, intervensi kontekstual-penerima yang tepat adalah intervensi yang mencapai tingkat disrupti yang tepat dengan tetap menghormati otonomi melalui pilihan. Kontekstualisasi ini didasarkan pada informasi tentang kapasitas, preferensi, dan tujuan pengguna, serta keadaan di mana intervensi akan berlaku.

Ada lima dimensi yang relevan dengan intervensi kontekstual-penerima. Empat dimensi ini muncul dari taksonomi penelitian interdisipliner McFarlane tentang interupsi disruptif antara komputer dan manusia. Dimensi-dimensi tersebut adalah: karakteristik individu penerima intervensi; metode koordinasi antara penerima dan sistem; makna atau tujuan intervensi; dan dampak keseluruhan intervensi. Dimensi relevansi kelima adalah opsionalitas: pengguna dapat memilih untuk mengabaikan semua saran yang ditawarkan atau menjalankan proses dan meminta intervensi lain yang lebih sesuai dengan kebutuhan mereka.

Kelima dimensi ini dapat diringkas dalam bentuk praktik terbaik berikut untuk intervensi kontekstual penerima:

- 2) Perancang AI4SG harus membangun sistem pengambilan keputusan dengan berkonsultasi dengan pengguna yang berinteraksi dengan, dan terdampak, oleh sistem ini; dengan memahami karakteristik pengguna, metode koordinasi, serta tujuan dan dampak intervensi; dan dengan menghormati hak pengguna untuk mengabaikan atau memodifikasi intervensi.

Penjelasan Kontekstual Penerima dan Tujuan Transparan

Aplikasi AI4SG harus dirancang untuk membuat operasi dan hasil sistem ini dapat dijelaskan dan untuk membuat transparan tujuannya. Kedua persyaratan ini tentu saja terkait erat, karena operasi dan hasil sistem AI mencerminkan tujuan yang lebih luas dari perancang manusia; di bagian ini, kami membahas keduanya secara bergantian. Menjadikan sistem AI dapat dijelaskan merupakan prinsip etika yang penting. Prinsip ini telah menjadi fokus penelitian setidaknya sejak tahun 1975. Dan hal ini telah mendapatkan lebih banyak perhatian baru-baru ini mengingat semakin meluasnya distribusi sistem AI. Seperti yang telah kita lihat di atas, proyek AI4SG harus menawarkan intervensi yang dikontekstualisasikan kepada

penerima. Selain itu, penjelasan untuk intervensi harus dikontekstualisasikan agar memadai, dan melindungi otonomi penerima.

Perancang proyek AI4SG telah mencoba meningkatkan keterjelasan sistem pengambilan keputusan dengan berbagai cara. Misalnya, para peneliti telah menggunakan pembelajaran mesin untuk memprediksi kesulitan akademik. Prediktor ini menggunakan konsep-konsep yang dianggap familiar dan menonjol oleh pejabat sekolah yang menafsirkan sistem tersebut, seperti skor IPK dan kategorisasi sosial-ekonomi. Para peneliti juga telah menggunakan pembelajaran penguatan untuk membantu pejabat di tempat penampungan tunawisma mendidik remaja tunawisma tentang HIV.

Sistem ini mempelajari cara memaksimalkan pengaruh pendidikan HIV dengan memilih remaja tunawisma mana yang akan dididik, dengan asumsi bahwa remaja tunawisma tersebut dapat mewariskan pengetahuan mereka. Salah satu versi sistem menjelaskan remaja mana yang dipilih dengan menampilkan grafik jaringan sosial mereka. Namun, petugas penampungan tunawisma menemukan bahwa penjelasan ini kontra-intuitif, yang berpotensi memengaruhi pemahaman tentang cara kerja sistem dan, oleh karena itu, kepercayaan pengguna terhadap sistem. Kedua kasus ini menunjukkan pentingnya konseptualisasi yang tepat ketika menjelaskan keputusan berbasis AI.

Konseptualisasi yang tepat kemungkinan akan bervariasi antar proyek AI4SG, karena mereka sangat berbeda dalam tujuan, pokok bahasan, konteks, dan pemangku kepentingan. Kerangka konseptual, yaitu Tingkat Abstraksi, bergantung pada apa yang dijelaskan dan kepada siapa. LoA merupakan komponen kunci dari sebuah teori, dan karenanya dari setiap penjelasan. Sebuah teori terdiri dari lima komponen:

1. Sistem, yang merupakan referensi atau objek yang dianalisis oleh sebuah teori;
2. Tujuan, yang merupakan "untuk apa" yang memotivasi analisis suatu sistem (perhatikan bahwa ini menjawab pertanyaan "untuk apa analisis?" dan jangan disamakan dengan tujuan sistem, yang menjawab pertanyaan "untuk apa sistem ini?". Di bawah ini, kami menggunakan istilah "tujuan" untuk tujuan sistem jika terdapat risiko kebingungan);
3. Tingkat Abstraksi, yang menyediakan sudut pandang untuk menganalisis sistem, dan menghasilkan;
4. Model, yaitu, beberapa informasi relevan dan andal tentang sistem yang dianalisis, yang mengidentifikasi;
5. Struktur sistem, yang terdiri dari fitur-fitur yang dimiliki oleh sistem yang dianalisis.

Terdapat saling ketergantungan antara pilihan tujuan spesifik, LoA relevan yang dapat memenuhi tujuan tersebut, sistem yang dianalisis, dan model yang diperoleh dengan menganalisis sistem pada LoA tertentu untuk tujuan tertentu. LoA menyediakan konseptualisasi sistem (misalnya, skor IPK, dan latar belakang sosial ekonomi). Namun, tujuan tersebut membatasi konstruksi LoA. Misalnya, jika kita memilih untuk menjelaskan sistem pengambilan keputusan itu sendiri (misalnya, penggunaan teknik pembelajaran mesin tertentu), maka LoA hanya dapat mengonseptualisasikan teknik AI tersebut. Selanjutnya, LoA menghasilkan model yang menjelaskan sistem tersebut. Model tersebut mengidentifikasi

struktur sistem, seperti skor IPK siswa tertentu, tingkat kehadiran yang buruk, dan latar belakang sosial ekonomi mereka yang menjadi prediktor kegagalan akademik mereka. Akibatnya, perancang harus memilih tujuan dan LoA yang sesuai dengan cermat, sehingga model penjelasan dapat memberikan penjelasan yang tepat tentang sistem yang dimaksud untuk penerima tertentu.

LoA dipilih untuk tujuan tertentu: misalnya, LoA yang dipilih untuk menjelaskan keputusan yang diambil berdasarkan hasil yang diperoleh melalui prosedur algoritmik bervariasi, tergantung pada apakah penjelasan tersebut ditujukan untuk penerima keputusan tersebut atau untuk insinyur yang bertanggung jawab atas desain prosedur algoritmik. Hal ini karena, tergantung pada tujuan dan tingkat ketelitiannya (misalnya, penjelasan yang ramah pelanggan vs. ramah insinyur), tidak semua LoA sesuai untuk penerima tertentu. Terkadang, pandangan konseptual penerima tentang dunia mungkin berbeda dari pandangan yang menjadi dasar penjelasan tersebut. Dalam kasus lain, penerima dan penjelasan mungkin selaras secara konseptual, tetapi penerima mungkin tidak sepakat tentang tingkat ketelitian (LoA) informasi (yang lebih tepatnya kami sebut model) yang diberikan. Ketidakselarasan konseptual berarti bahwa penerima mungkin menganggap penjelasan tersebut tidak relevan, tidak dapat dipahami, atau, seperti yang akan kita lihat di bawah, dipertanyakan.

Sehubungan dengan (ketidak)terpahaman, LoA dapat menggunakan label yang tidak diketahui (disebut *observabel*), atau label yang memiliki arti berbeda bagi pengguna yang berbeda. Studi empiris menunjukkan bahwa kesesuaian suatu penjelasan berbeda-beda di antara penerima sesuai dengan keahlian mereka. Penerima mungkin memerlukan penjelasan tentang bagaimana perangkat lunak AI sampai pada suatu keputusan, terutama ketika mereka harus mengambil tindakan berdasarkan keputusan tersebut. Bagaimana sistem AI sampai pada suatu kesimpulan bisa sama pentingnya dengan justifikasi atas kesimpulan tersebut. Akibatnya, perancang juga harus mengontekstualisasikan metode penjelasan kepada penerima.

Kasus perangkat lunak yang menggunakan algoritma maksimisasi pengaruh untuk menysar remaja tunawisma dalam pendidikan HIV memberikan contoh yang baik tentang relevansi kontekstualisasi konsep oleh penerima. Para peneliti yang terlibat dalam proyek ini mempertimbangkan tiga kemungkinan LoA ketika merancang model penjelasan: LoA pertama mencakup kalkulasi utilitas; LoA kedua berfokus pada konektivitas grafik sosial; dan LoA ketiga berfokus pada tujuan pedagogis. LoA pertama menyoroti manfaat menargetkan satu remaja tunawisma daripada yang lain. Menurut para peneliti, dalam kasus ini, petugas penampungan tunawisma (penerima) mungkin salah memahami perhitungan utilitas atau menganggapnya tidak relevan.

Perhitungan utilitas hanya menawarkan sedikit daya penjelasan di luar keputusan itu sendiri, karena seringkali hanya menunjukkan bahwa pilihan "terbaik" telah dibuat, dan seberapa baik pilihan tersebut. Penjelasan berdasarkan LoA kedua menghadapi masalah yang berbeda: penerima berasumsi bahwa simpul paling sentral dalam jaringan adalah yang terbaik untuk memaksimalkan pengaruh pendidikan, sementara pilihan optimal seringkali merupakan serangkaian simpul yang kurang terhubung dengan baik. Kesenjangan ini mungkin muncul dari

sifat konektivitas antara anggota jaringan remaja tunawisma, yang mencerminkan ketidakpastian dalam kehidupan nyata tentang persahabatan. Karena siapa yang dianggap sebagai "teman" seringkali tidak jelas dan dapat berubah seiring waktu, para peneliti mengklasifikasikan batas-batas dalam jaringan sebagai "pasti" atau "tidak pasti" berdasarkan pengetahuan domain. Untuk hubungan yang "tidak pasti", probabilitas terjalinnya persahabatan antara dua remaja ditentukan oleh para ahli domain.

LoA ketiga akhirnya dipilih, setelah pengujian pengguna selanjutnya terhadap berbagai kerangka kerja penjelasan. Mengingat tujuan mereka untuk membenarkan keputusan dengan cara yang intuitif bagi petugas penampungan tunawisma, para peneliti mempertimbangkan untuk menghilangkan referensi ke perhitungan Utilitas Harapan Maksimum (MEU), meskipun hal inilah yang sebenarnya mendasari keputusan yang dibuat oleh sistem. Sebaliknya, para peneliti mempertimbangkan untuk membenarkan keputusan menggunakan konsep-konsep yang lebih nyaman dan familier bagi para petugas, seperti sentralitas node (yaitu, remaja) yang direkomendasikan sistem untuk diprioritaskan oleh petugas untuk diintervensi. Dengan cara ini, para peneliti berusaha memberikan informasi yang paling relevan dan sesuai konteks kepada penerima.

Seperti yang ditunjukkan contoh ini, dalam suatu sistem tertentu, tujuan yang dipilih seseorang untuk dicari penjelasannya, pada LoA berapa, dan model penerbitan yang diperoleh merupakan variabel krusial yang memengaruhi efektivitas penjelasan. Keterjelasan menumbuhkan kepercayaan dan mendorong adopsi solusi AI4SG. Inilah mengapa penting bagi perangkat lunak untuk menggunakan argumentasi persuasif bagi audiens target. Hal ini kemungkinan mencakup informasi tentang fungsionalitas umum dan logika yang digunakan oleh suatu sistem serta alasan di balik keputusan spesifik yang dibuat. Transparansi dalam tujuan (yaitu, tujuan sistem) juga krusial, karena hal ini mengikuti langsung prinsip otonomi. Pertimbangan, misalnya, pengembangan solusi AI untuk mendorong penyandang disabilitas kognitif minum obat mereka. Sekilas, aplikasi ini mungkin tampak invasif, melibatkan pengguna yang rentan, dan membatasi efektivitas penjelasan yang dikonseptualisasikan oleh penerima.

Namun, sistem ini tidak dirancang untuk memaksa pasien melakukan perilaku tertentu, juga tidak dirancang untuk menyerupai manusia. Pasien memiliki otonomi untuk tidak berinteraksi dengan sistem AI yang dimaksud. Kasus ini menyoroti pentingnya transparansi dalam tujuan, terutama dalam konteks di mana operasi dan hasil yang dapat dijelaskan tidak dapat dijalankan atau tidak diinginkan. Dengan demikian, transparansi dalam tujuan mendukung perlindungan lain seputar perlindungan populasi sasaran dan dapat membantu memastikan kepatuhan terhadap undang-undang dan preseden yang relevan.

Sebaliknya, tujuan yang tidak transparan dapat memicu kesalahpahaman dan potensi bahaya. Misalnya, ketika pengguna sistem AI tidak yakin tentang jenis agen yang mereka hadapi manusia, buatan, atau kombinasi hibrida keduanya mereka mungkin keliru berasumsi bahwa norma-norma diam-diam interaksi sosial antarmanusia ditegakkan (misalnya, tidak merekam setiap detail percakapan).

Seperti biasa, konteks sosial tempat penerapan AI4SG berlangsung memengaruhi sejauh mana sistem AI harus transparan dalam operasionalnya. Karena transparansi merupakan posisi default, tetapi bukan posisi absolut, mungkin ada alasan yang sah bagi perancang untuk menghindari pemberian informasi kepada pengguna tentang tujuan perangkat lunak. Misalnya, nilai ilmiah suatu proyek atau kondisi kesehatan dan keselamatan ruang publik dapat membenarkan tujuan yang sementara tidak transparan. Pertimbangkan sebuah studi yang menipu mahasiswa agar percaya bahwa mereka berinteraksi dengan asisten mata kuliah manusia yang ternyata, seiring waktu, ternyata adalah bot.

Penipuan bot tersebut, sebagaimana yang dikemukakan oleh penulis, adalah untuk memainkan "permainan imitasi" tanpa membuat mahasiswa memilih kueri bahasa alami yang lebih sederhana dan kurang manusiawi berdasarkan prakonsepsi tentang kemampuan AI. Dalam kasus seperti itu, pilihan antara opasitas dan transparansi dapat diinformasikan oleh gagasan yang sudah ada sebelumnya tentang persetujuan berdasarkan informasi untuk eksperimen subjek manusia yang tertanam dalam Kode Nuremberg, Deklarasi Helsinki, dan Laporan Belmont.

Lebih luas lagi, kemampuan untuk menghindari penggunaan sistem AI menjadi lebih mungkin ketika perangkat lunak AI mengungkapkan tujuan endogeniknya, seperti mengklasifikasikan data tentang seseorang. Misalnya, perangkat lunak AI dapat memberi tahu staf di bangsal rumah sakit bahwa tujuannya adalah mengklasifikasikan tingkat kebersihan mereka. Dalam hal ini, staf dapat memutuskan untuk menghindari klasifikasi tersebut jika ada tindakan alternatif yang wajar yang dapat mereka ambil. Dalam kasus lain, mengungkapkan suatu tujuan membuatnya lebih kecil kemungkinannya untuk tercapai.

Transparansi tujuan dan motivasi pengembang AI4SG sendiri merupakan faktor penting bagi keberhasilan setiap proyek, tetapi faktor ini mungkin bertentangan dengan tujuan sistem itu sendiri. Inilah sebabnya mengapa penting untuk menilai, pada tahap desain, tingkat transparansi (yaitu seberapa besar transparansi, jenis transparansi apa, untuk siapa, dan tentang apa?) yang akan dianut proyek mengingat tujuan keseluruhannya dan konteks implementasinya. Digabungkan dengan kebutuhan akan penjelasan yang dikonseptualisasikan oleh penerima, pertimbangan ini menghasilkan serangkaian praktik terbaik berikut:

Perancang AI4SG harus memilih Tingkat Abstraksi untuk penjelasan AI yang memenuhi tujuan penjelasan yang diinginkan dan sesuai dengan sistem dan penerima; kemudian menerapkan argumen yang rasional dan cukup persuasif bagi penerima untuk menyampaikan penjelasan; dan memastikan bahwa tujuan (tujuan sistem) yang menjadi tujuan pengembangan dan penerapan sistem AI4SG dapat diketahui oleh penerima keluarannya secara default.

Perlindungan Privasi dan Persetujuan Subjek Data

Dari ketujuh faktor yang kami bahas, privasi mungkin memiliki literatur yang paling banyak. Hal ini tidak mengherankan, karena privasi dianggap sebagai syarat penting bagi keselamatan, martabat manusia, dan kohesi sosial, antara lain, dan karena gelombang

teknologi digital sebelumnya telah berdampak besar pada privasi. Keselamatan masyarakat dapat terancam ketika aktor jahat atau negara menguasai individu melalui pelanggaran privasi.

Menghormati privasi juga merupakan syarat penting bagi martabat manusia, karena kita dapat memandang informasi pribadi sebagai sesuatu yang membentuk individu, dan mendeprivatisasi catatan tanpa persetujuan kemungkinan besar merupakan pelanggaran martabat manusia. Konsepsi privasi individu sebagai hak fundamental mendasari tindakan legislatif baru-baru ini, misalnya, di Eropa (melalui Peraturan Perlindungan Data Umum) dan Jepang (melalui Undang-Undang tentang Perlindungan Informasi Pribadi), serta keputusan pengadilan di yurisdiksi seperti India. Privasi mendukung orang-orang untuk menyimpang dari norma sosial tanpa menyebabkan pelanggaran, dan komunitas dalam mempertahankan struktur sosial mereka, sehingga privasi juga menopang kohesi sosial.

Dalam kasus AI4SG, sangat penting untuk menekankan relevansi persetujuan pengguna terhadap penggunaan data pribadi. Ketegangan dapat muncul antara ambang batas persetujuan yang berbeda. Ketegangan ini seringkali paling menegangkan dalam situasi "hidup atau mati" seperti keadaan darurat nasional dan pandemi. Pertimbangan wabah Ebola di Afrika Barat pada tahun 2014, yang menimbulkan dilema etika yang kompleks (*The Economist* 2014). Dalam kasus ini, rilis dan analisis cepat data panggilan dari pengguna ponsel di wilayah tersebut mungkin memungkinkan para ahli epidemiologi untuk melacak penyebaran penyakit mematikan tersebut. Namun, rilis data tersebut terhambat oleh kekhawatiran yang valid seputar privasi pengguna, serta nilai data tersebut bagi para pesaing industri.

Dalam keadaan di mana kecepatan tidak begitu krusial, dimungkinkan untuk mendapatkan persetujuan subjek untuk dan sebelum data tersebut digunakan. Tingkat atau jenis persetujuan yang diminta dapat bervariasi tergantung konteksnya. Dalam pelayanan kesehatan, seseorang dapat mengadopsi ambang batas persetujuan yang diasumsikan, di mana melaporkan masalah medis kepada dokter merupakan persetujuan yang diasumsikan dari pihak pasien. Dalam keadaan lain, ambang batas persetujuan berdasarkan informasi akan lebih tepat. Namun, karena persetujuan berdasarkan informasi mengharuskan peneliti untuk mendapatkan persetujuan spesifik dari pasien sebelum menggunakan data mereka untuk tujuan yang tidak disetujui, praktisi dapat memilih ambang batas persetujuan yang eksplisit untuk pemrosesan data umum, yaitu, untuk penggunaan medis apa pun. Ambang batas ini tidak mengharuskan pasien untuk diberitahu tentang semua kemungkinan cara peneliti dapat menggunakan data mereka. Alternatif lain adalah gagasan "persetujuan dinamis" yang terus berkembang, di mana individu dapat memantau dan menyesuaikan preferensi privasi mereka secara terperinci.

Dalam kasus lain, persetujuan berdasarkan informasi dapat diabaikan sepenuhnya. Hal ini terjadi dengan penciptaan perangkat lunak pembelajaran mesin baru-baru ini untuk memprediksi prognosis penderita kanker ovarium dengan memanfaatkan analisis retrospektif gambar anonim. Penggunaan data kesehatan pasien dalam pengembangan solusi AI tanpa persetujuan pasien juga telah menarik perhatian regulator perlindungan data. Pada tahun 2017, Komisioner Informasi Inggris memutuskan bahwa Royal Free NHS Foundation Trust melanggar Undang-Undang Perlindungan Data ketika memberikan detail pasien kepada

Google DeepMind, untuk tujuan melatih sistem AI guna mendiagnosis cedera ginjal akut. Komisioner mencatat sebagai sebuah “kekurangan” bahwa “pasien tidak diberi informasi yang memadai bahwa data mereka akan digunakan sebagai bagian dari pengujian” (*“Royal Free—Uji Coba Google DeepMind Gagal Mematuhi Hukum Perlindungan Data” 2017*).

Namun, mencapai keseimbangan antara menghormati privasi pasien dan menciptakan AI4SG yang efektif masih mungkin dilakukan. Inilah tantangan yang dihadapi para peneliti di Haque .yang ingin menciptakan sistem untuk melacak kepatuhan terhadap aturan seputar kebersihan tangan di rumah sakit, untuk mencegah penyebaran infeksi. Meskipun pendekatan berbasis visi komputer memiliki keuntungan teknis yang jelas untuk masalah ini, penggunaan rekaman video berbenturan dengan peraturan privasi yang membatasinya.

Bahkan dalam kasus di mana rekaman video diizinkan, akses ke rekaman (untuk melatih algoritma) seringkali ketat. Sebagai gantinya, para peneliti menggunakan "gambar kedalaman", yang menghilangkan identitas subjek, menjaga privasi mereka. Meskipun pilihan desain ini berarti "kehilangan isyarat tampilan visual yang penting dalam prosesnya", hal ini memenuhi aturan privasi, dan sistem non-intrusif para peneliti masih berhasil mengungguli solusi yang ada.

Terakhir, persetujuan di ruang daring juga bermasalah; Pengguna sering kali tidak memiliki pilihan dan disajikan dengan opsi 'ambil atau tinggalkan' saat mengakses layanan daring. Kurangnya perlindungan atau persetujuan untuk penggunaan data pribadi bekas yang dibagikan secara publik secara daring memungkinkan pengembangan perangkat lunak AI yang bermasalah secara etika. Misalnya, sebuah makalah baru-baru ini menggunakan gambar wajah yang tersedia untuk umum yang diunggah ke situs web kencan sebagai cara untuk melatih perangkat lunak AI untuk mendeteksi seksualitas seseorang berdasarkan sejumlah kecil foto. Meskipun penelitian tersebut menerima persetujuan komite etik, hal itu menimbulkan pertanyaan lebih lanjut seputar persetujuan, karena tidak masuk akal bahwa pengguna situs web kencan dapat atau tentu saja akan menyetujui penggunaan data mereka untuk tujuan khusus ini. Privasi bukanlah masalah baru, tetapi sentralitas data pribadi bagi banyak aplikasi AI (dan AI4SG) meningkatkan signifikansi etisnya dan menciptakan masalah seputar persetujuan. Dari sini, kita dapat menyimpulkan praktik terbaik berikut:

Perancang AI4SG harus menghormati ambang batas persetujuan yang ditetapkan untuk pemrosesan kumpulan data pribadi.

Keadilan Situasional

Pengembang AI biasanya mengandalkan data, yang mungkin bias dalam hal-hal yang signifikan secara sosial. Bias ini dapat berlanjut ke pengambilan keputusan algoritmik yang mendasari banyak sistem AI, dengan cara yang tidak adil bagi subjek dalam proses pengambilan keputusan dan, dengan demikian, dapat melanggar prinsip keadilan. Keputusan-keputusan ini mungkin didasarkan pada faktor-faktor yang penting secara etis (misalnya, etnis, gender, atau agama) dan tidak relevan dengan pengambilan keputusan yang sedang dilakukan, atau mungkin relevan tetapi dilindungi secara hukum sebagai karakteristik non-diskriminatif.

Lebih lanjut, keputusan yang didorong oleh AI dapat berasal dari faktor-faktor yang tidak memiliki kepentingan etis yang jelas, namun secara kolektif merupakan pengambilan keputusan yang bias secara tidak adil.

Inisiatif AI4SG yang mengandalkan data bias dapat menyebarkan bias ini melalui lingkaran setan. Siklus seperti itu akan dimulai dengan kumpulan data bias yang menginformasikan fase pertama pengambilan keputusan AI, yang mengakibatkan tindakan diskriminatif, yang pada gilirannya mengarah pada pengumpulan dan penggunaan data bias. Pertimbangkan penggunaan AI untuk memprediksi kelahiran prematur di Amerika Serikat, di mana hasil kesehatan ibu hamil telah lama dipengaruhi oleh etnis mereka. Bias yang telah lama ada terhadap perempuan Afrika-Amerika yang mencari pengobatan, karena stereotip historis yang merugikan, berkontribusi pada tingkat morbiditas ibu yang lebih dari tiga kali lebih tinggi daripada perempuan kulit putih.

Di sini, AI mungkin menawarkan potensi besar untuk mengurangi kesenjangan rasial yang tajam ini, tetapi hanya jika diskriminasi historis yang sama tidak terulang dalam sistem AI. Atau pertimbangkan penggunaan perangkat lunak kepolisian prediktif. Pengembang dapat melatih perangkat lunak kepolisian prediktif pada data kepolisian yang mengandung prasangka yang mengakar kuat. Ketika diskriminasi memengaruhi tingkat penangkapan, hal itu akan tertanam dalam data penuntutan. Bias semacam itu dapat menyebabkan keputusan diskriminatif (misalnya, peringatan atau penangkapan) yang berdampak pada kumpulan data yang semakin bias, sehingga menciptakan lingkaran setan.

Contoh-contoh ini melibatkan penggunaan AI untuk meningkatkan hasil di domain-domain yang datanya telah dikumpulkan. Namun, dalam banyak konteks lain, proyek-proyek AI4SG (atau inisiatif serupa) pada dasarnya membuat warga negara "terlihat" dengan cara yang sebelumnya tidak terlihat, termasuk dalam konteks negara-negara berkembang. Peningkatan visibilitas ini menekankan pentingnya perlindungan terhadap potensi amplifikasi bias berbahaya oleh teknologi AI.

Jelas, para perancang harus membersihkan kumpulan data yang digunakan untuk melatih AI. Namun, ada pula risiko menerapkan disinfektan yang terlalu kuat, misalnya, dengan menghilangkan nuansa kontekstual penting yang dapat meningkatkan pengambilan keputusan etis. Jadi, para perancang juga harus memastikan bahwa pengambilan keputusan AI tetap sensitif terhadap faktor-faktor yang penting bagi inklusivitas. Misalnya, kita harus memastikan bahwa pengolah kata berinteraksi secara identik dengan pengguna manusia tanpa memandang jenis kelamin dan etnis pengguna tersebut, tetapi juga mengantisipasi bahwa pengolah kata tersebut dapat beroperasi dengan cara yang tidak setara namun adil dengan membantu penyandang disabilitas visual.

Harapan seperti itu tidak selalu terpenuhi dalam konteks penalaran yang digerakkan oleh AI. Dibandingkan dengan pengolah kata, AI memungkinkan rentang pengambilan keputusan dan modalitas interaksi yang jauh lebih luas, yang banyak di antaranya didorong oleh data yang berpotensi bias. Set data pelatihan mungkin berisi bahasa alami yang membawa asosiasi yang tidak adil antara gender dan kata-kata yang, pada gilirannya, membawa kekuatan normatif. Dalam konteks dan kasus penggunaan lain, pendekatan yang

adil mungkin memerlukan perbedaan dalam komunikasi, berdasarkan faktor-faktor seperti gender.

Pertimbangkan kasus asisten pengajar virtual yang gagal membedakan dengan cukup baik antara pria dan wanita dalam tanggapannya ketika diberi tahu bahwa seorang pengguna sedang mengandung bayi, memberi selamat kepada para pria dan mengabaikan para wanita. Investigasi BBC News menyoroti contoh yang lebih mengerikan: chatbot kesehatan mental yang dianggap cocok untuk digunakan oleh anak-anak tidak dapat memahami seorang anak yang secara eksplisit melaporkan pelecehan seksual di bawah umur. Sebagaimana yang diperjelas oleh kasus-kasus ini, penggunaan AI dalam interaksi manusia-komputer, seperti chatbot, membutuhkan pemahaman yang tepat tentang kelompok-kelompok penting tempat pengguna berada dan karakteristik yang mereka wujudkan ketika berinteraksi dengan perangkat lunak.

Menghormati keadilan situasional sangat penting bagi keberhasilan implementasi AI4SG. Untuk mencapainya, proyek AI4SG perlu menghilangkan faktor-faktor (dan proksinya) yang penting secara etis tetapi tidak relevan dengan hasil, dan memasukkan faktor-faktor yang sama ketika diperlukan, baik demi inklusivitas, keamanan, maupun pertimbangan etis lainnya. Masalah bias historis yang memengaruhi pengambilan keputusan di masa mendatang sudah lama ada. Yang baru adalah potensi bahwa bias-bias ini akan tertanam, diperkuat, dan diabadikan kembali oleh mekanisme pembelajaran penguatan yang keliru.

Risiko ini khususnya terasa jelas ketika dipertimbangkan bersamaan dengan risiko ketidakjelasan dalam sistem pengambilan keputusan AI dan hasilnya. Kita akan kembali membahas topik ini di bagian selanjutnya. Dari identifikasi kami terhadap keadilan situasional sebagai faktor penting, kami dapat menyimpulkan praktik terbaik berikut:

Perancang AI4SG harus menghilangkan variabel dan proksi yang tidak relevan dengan hasil dari kumpulan data yang relevan, kecuali jika inklusi mereka mendukung inklusivitas, keamanan, atau keharusan etis lainnya.

Semantisasi yang Ramah Manusia

AI4SG harus memungkinkan manusia untuk mengkurasi dan mengembangkan "modal semantik" mereka, yaitu,

konten apa pun yang dapat meningkatkan kekuatan seseorang untuk memberi makna dan memaknai (semantisasi) sesuatu.

Hal ini krusial untuk mempertahankan dan mengembangkan otonomi manusia. Dengan AI, kita mungkin seringkali memiliki kapasitas teknis untuk mengotomatiskan penciptaan makna dan rasa (semantisasi), tetapi ketidakpercayaan atau ketidakadilan juga dapat muncul jika kita melakukannya secara ceroboh. Dua masalah muncul. Masalah pertama adalah bahwa perangkat lunak AI dapat mendefinisikan semantikasi dengan cara yang berbeda dari pilihan kita sendiri.

Hal ini terjadi jika suatu prosedur mendefinisikan makna secara sembarangan (misalnya, berdasarkan lemparan koin). Masalah yang sama dapat muncul jika perangkat lunak AI mendukung suatu bentuk semantikasi berdasarkan penggunaan yang sudah ada sebelumnya. Misalnya, para peneliti telah mengembangkan aplikasi yang memprediksi makna hukum 'pelanggaran' berdasarkan kasus-kasus masa lalu.

Jika seseorang menggunakan perangkat lunak untuk mendefinisikan makna 'pelanggaran', maka orang tersebut akan membatasi peran hakim dan hakim agung. Mereka tidak akan lagi dapat melakukan semantikasi (memperbaiki dan mendefinisikan ulang makna, dan kemungkinan untuk memahami) "pelanggaran", ketika mereka menafsirkan hukum. Ini menjadi masalah, karena penggunaan masa lalu tidak selalu memprediksi bagaimana kita akan melakukan semantikasi konsep atau fenomena yang sama di masa depan.

Masalah kedua adalah, dalam lingkungan sosial, tidak praktis bagi perangkat lunak AI untuk mendefinisikan semua makna dan pengertian. Beberapa semantikasi bersifat subjektif, karena siapa atau apa yang terlibat dalam semantikasi juga sebagian merupakan bagian dari proses dan hasilnya. Misalnya, hanya agen yang berwenang secara hukum yang dapat mendefinisikan makna hukum 'pelanggaran'. Demikian pula, makna dan arti simbol afektif, seperti ekspresi wajah, juga bergantung pada jenis agen yang menunjukkan ekspresi tertentu. AI afektif dapat mendeteksi emosi, agen buatan dapat menyatakan secara akurat bahwa seseorang tampak sedih, tetapi tidak dapat mengubah makna kesedihan.

Solusi untuk kedua masalah ini terletak pada perbedaan antara tugas yang harus dan tidak boleh didelegasikan ke sistem buatan. AI harus digunakan untuk memfasilitasi semantikasi yang ramah manusia, tetapi bukan untuk menyediakannya sendiri. Hal ini berlaku, misalnya, ketika mempertimbangkan pasien dengan penyakit Alzheimer. Penelitian tentang hubungan pengasuh-pasien menyoroti tiga poin.

Pertama, pengasuh memainkan peran penting, tetapi memberatkan, dalam mengingatkan pasien tentang aktivitas yang mereka ikuti, misalnya, minum obat. Kedua, pengasuh juga memainkan peran penting dalam menyediakan interaksi yang bermakna bagi pasien. Dan ketiga, ketika perawat mengingatkan pasien untuk minum obat, hubungan pasien-perawat dapat melemah karena pasien merasa terganggu, sehingga perawat kehilangan sebagian kapasitasnya untuk memberikan empati dan dukungan yang bermakna. Akibatnya, para peneliti telah mengembangkan perangkat lunak AI yang menyeimbangkan antara mengingatkan pasien dan mengganggu pasien.

Keseimbangan ini dipelajari dan dioptimalkan menggunakan pembelajaran penguatan. Para peneliti merancang sistem agar perawat dapat menghabiskan sebagian besar waktu mereka untuk memberikan dukungan empati dan menjaga hubungan yang bermakna dengan pasien. Seperti yang ditunjukkan contoh ini, AI dapat digunakan untuk menyingkirkan tugas-tugas formulais sekaligus mempertahankan semantikasi yang ramah manusia. Semantikasi yang berpusat pada manusia, sebagai faktor penting untuk AI4SG, mendasari praktik terbaik terakhir kami:

- 3) Perancang AI4SG tidak boleh menghalangi kemampuan orang untuk melakukan semantifikasi (yaitu, memberi makna pada, dan memaknai) sesuatu.

8.4 KESIMPULAN

Faktor-Faktor Penyeimbang untuk AI demi Kebaikan Sosial

Tujuh faktor yang dianalisis pada halaman-halaman sebelumnya dirangkum dalam Tabel 8.1, beserta praktik terbaiknya, dan lima prinsip etika AI yang diidentifikasi dalam yang paling erat kaitannya dengan setiap faktor. Sebagai pengulangan, prinsip kebaikan diasumsikan sebagai prasyarat untuk AI4SG, sehingga faktor-faktor tersebut berkaitan dengan satu atau lebih dari empat prinsip lainnya: non-maleficence, otonomi, keadilan, dan eksplikabilitas. Tujuh faktor tersebut menunjukkan bahwa menciptakan AI4SG yang sukses membutuhkan dua jenis keseimbangan: intra dan inter.

Di satu sisi, setiap faktor tunggal itu sendiri mungkin memerlukan keseimbangan intrinsik, misalnya, antara risiko intervensi berlebihan dan risiko intervensi kurang ketika merancang intervensi kontekstual; atau antara perlindungan melalui pengaburan dan perlindungan melalui penghitungan perbedaan yang menonjol antarmanusia, tergantung pada tujuan dan konteks suatu sistem. Di sisi lain, keseimbangan tidak hanya spesifik untuk satu faktor; keseimbangan juga sistemik, karena keseimbangan tersebut juga harus dicapai di antara berbagai faktor. Pertimbangkan ketegangan antara mencegah pelaku jahat memahami cara "memanipulasi" data masukan sistem prediksi AI versus memungkinkan manusia untuk mengesampingkan hasil yang benar-benar cacat; atau ketegangan antara memastikan pengungkapan alasan di balik suatu keputusan secara efektif tanpa mengorbankan anonimitas konsensual subjek data.

Pertanyaan menyeluruh yang dihadapi komunitas AI4SG adalah, untuk setiap kasus tertentu, apakah seseorang secara moral berkewajiban untuk, atau berkewajiban untuk tidak, merancang, mengembangkan, dan menerapkan proyek AI4SG tertentu. Artikel ini tidak berusaha menjawab pertanyaan semacam itu secara abstrak. Menyelesaikan ketegangan yang mungkin muncul di antara faktor-faktor sangat bergantung pada konteks, dan analisis sebelumnya tidak dimaksudkan untuk mencakup semua konteks potensial, terutama karena hal ini tidak konsisten dengan argumen untuk pengujian hipotesis yang dapat difalsifikasi dan penerapan inkremental yang didukung dalam artikel ini; daftar periksa "yang boleh dan tidak boleh" yang murni teknis juga tidak akan cukup.

Sebaliknya, analisis kami telah menghasilkan serangkaian faktor penting yang perlu dipertimbangkan, ditafsirkan, dan dievaluasi secara kontekstual ketika seseorang merancang, mengembangkan, dan menerapkan proyek AI4SG tertentu. Masa depan AI4SG kemungkinan akan memberikan lebih banyak peluang untuk memperkaya serangkaian faktor penting tersebut. AI sendiri dapat membantu mengelola siklus hidupnya sendiri dengan menyediakan, secara meta-reflektif, alat untuk mengevaluasi cara terbaik untuk mencapai keseimbangan individual dan sistemik yang ditunjukkan di atas.

Pertanyaan paling relevan yang muncul dari faktor-faktor yang dijelaskan dalam artikel ini kemungkinan besar berkaitan dengan tantangan menyeimbangkan kebutuhan dan klaim

yang saling bersaing yang diperkenalkan oleh faktor-faktor tersebut dan praktik terbaik yang sesuai. Hal ini menyangkut apa yang melegitimasi pengambilan keputusan dengan dan tentang AI. Meskipun kami menyerahkan perhatian ini terutama untuk penelitian mendatang, kami sampaikan beberapa catatan penutup.

Pertanyaan seperti ini tak pelak lagi terkait dengan tantangan etika dan politik yang lebih luas mengenai siapa yang memiliki kekuasaan atau "kedudukan" untuk berpartisipasi dalam proses evaluasi ini, serta bagaimana berbagai preferensi diukur dan diagregasi, sebagaimana diuraikan dalam kerangka trikotomik Baum. Jika kita berasumsi bahwa tantangan dalam menyeimbangkan faktor-faktor setidaknya harus bersifat partisipatif, tinjauan Prasad (2018) tentang teorema pilihan sosial yang relevan mengidentifikasi beberapa kondisi latar belakang untuk mendukung pengambilan keputusan kelompok yang efektif.

Sebagaimana ditunjukkan oleh analisis ini, penggabungan berbagai perspektif ke dalam desain sistem pengambilan keputusan AI kemungkinan besar merupakan langkah yang penting secara etika, baik untuk AI secara umum, maupun AI4SG secara khusus. Masih banyak pekerjaan yang harus dilakukan untuk memastikan bahwa proyek-proyek AI4SG dirancang sedemikian rupa sehingga tidak hanya memajukan tujuan-tujuan yang bermanfaat dan mengatasi tantangan sosial, tetapi juga melakukannya dengan cara-cara yang lebih disukai secara sosial dan berkelanjutan.

Artikel ini bertujuan untuk berkontribusi dalam meletakkan dasar bagi praktik-praktik dan kebijakan yang baik dalam hal ini, serta untuk penelitian lebih lanjut tentang pertimbangan-pertimbangan etis yang seharusnya mendasari proyek-proyek AI4SG, dan dengan demikian, "proyek AI4SG" secara keseluruhan.

Tabel 8.1 Ringkasan tujuh faktor pendukung AI4SG dan praktik terbaik terkait

Faktor-faktor	Praktik terbaik yang sesuai	Prinsip etika yang sesuai
Kemungkinan falsifikasi dan penerapan inkremental	Identifikasi persyaratan yang dapat dipalsukan dan ujilah secara bertahap dari laboratorium ke "dunia luar".	Nonmaleficence
Perlindungan terhadap manipulasi prediktor	Adopsi perlindungan yang (i) memastikan bahwa indikator non-kausal tidak mendistorsi intervensi secara tidak tepat, dan (ii) batasi, jika perlu, pengetahuan tentang bagaimana masukan memengaruhi keluaran dari sistem AI4SG, untuk mencegah manipulasi.	Nonmaleficence

Intervensi kontekstual penerima	Membangun sistem pengambilan keputusan dengan berkonsultasi dengan pengguna yang berinteraksi dengan dan terdampak oleh sistem tersebut; dengan memahami karakteristik pengguna, metode koordinasi, tujuan dan dampak intervensi; dan dengan menghormati hak pengguna untuk mengabaikan atau mengubah intervensi.	Otonomi
Penjelasan kontekstual penerima dan tujuan transparan	Pilih Tingkat Abstraksi untuk penjelasan AI yang memenuhi tujuan penjelasan yang diinginkan dan sesuai dengan sistem dan penerima; kemudian terapkan argumen yang rasional dan cukup persuasif bagi penerima untuk menyampaikan penjelasan; dan pastikan bahwa tujuan (tujuan sistem) yang menjadi dasar pengembangan dan penerapan sistem AI4SG dapat diketahui oleh penerima keluarannya secara default.	Kejelasan
Perlindungan privasi dan persetujuan subjek data	Hormati ambang batas persetujuan yang ditetapkan untuk pemrosesan kumpulan data pribadi.	Nonmaleficence; otonomi
Keadilan situasional	Hapus variabel dan proksi yang tidak relevan dengan hasil dari kumpulan data yang relevan, kecuali jika penyertaannya mendukung inklusivitas, keamanan, atau keharusan etis lainnya.	Keadilan
Semantikalisisasi yang ramah manusia	Jangan menghalangi kemampuan orang untuk melakukan semantikalisisasi (yaitu, memberi makna dan memaknai) sesuatu.	Otonomi

BAB 9

ALAT DAN METODE ETIKA AI UNTUK PUBLIK

Perkembangan pesat teknologi Kecerdasan Buatan (*Artificial Intelligence/AI*) telah memicu kembali perdebatan mengenai implikasi etisnya, yang sejatinya telah dimulai sejak tahun 1960-an. Tokoh-tokoh awal seperti Samuel dan Wiener telah memperkenalkan dimensi etis dalam pembahasan AI, terutama melalui karya mereka yang berpengaruh dalam bidang sains dan sibernetika. Namun, dalam dekade terakhir, perkembangan teknologi AI mengalami transformasi signifikan, terutama dengan munculnya Jaringan Syaraf Tiruan (*Deep Learning*) dan teknik Pembelajaran Mesin (*Machine Learning*).

Perubahan ini tidak hanya meningkatkan kemampuan teknis AI, tetapi juga memperluas dampaknya terhadap kehidupan masyarakat. Seiring dengan meningkatnya kompleksitas dan penggunaan AI, diskursus etika yang menyertainya pun menjadi semakin penting dan arus utama. Sayangnya, sebagian besar perdebatan etika AI masih berpusat pada ranah prinsip seperti beneficence (kebaikan), non-maleficence (tidak membahayakan), otonomi, keadilan, dan eksplikabilitas dan belum banyak menyentuh aspek praktis tentang bagaimana prinsip-prinsip tersebut dapat diterapkan secara nyata dalam proses pengembangan teknologi AI, khususnya pada Pembelajaran Mesin.

Kesadaran akan pentingnya isu etis memang telah tumbuh pesat, namun kemampuan komunitas AI dalam merespons dan mengambil langkah konkret untuk memitigasi risiko etis masih tergolong terbatas. Oleh karena itu, penelitian ini bertujuan untuk menjembatani kesenjangan antara prinsip dan praktik etika dalam pengembangan AI. Fokus utama diberikan pada penyusunan tipologi yang dapat membantu para pengembang menerapkan pertimbangan etis secara sistematis dalam setiap tahap alur kerja Pembelajaran Mesin.

Selain itu, penelitian ini juga berupaya memberikan panduan bagi para peneliti untuk mengidentifikasi area yang masih memerlukan kajian lebih lanjut. Meskipun ruang lingkup studi ini berfokus pada Pembelajaran Mesin, pendekatan yang dikembangkan diharapkan dapat diterapkan pula pada cabang-cabang AI lainnya. Bab ini akan menguraikan metode yang digunakan dalam membangun tipologi etika ini, menyajikan temuan awal, serta merumuskan agenda riset yang relevan untuk pengembangan etika AI di masa mendatang.

9.1 PENDAHULUAN

Seiring meningkatnya ketersediaan data di hampir setiap aspek kehidupan, dan kecanggihan teknik pembelajaran mesin (ML), peluang untuk meningkatkan kehidupan publik dan privat pun meningkat. Masyarakat memiliki kendali yang lebih besar daripada sebelumnya atas hasil yang berkaitan dengan: (1) siapa yang dapat dicapai manusia; (2) apa yang dapat dilakukan manusia; (3) apa yang dapat dicapai manusia; dan (4) bagaimana manusia dapat berinteraksi dengan dunia. Namun, kekhawatiran yang semakin meningkat tentang tantangan etika yang ditimbulkan oleh meningkatnya penggunaan ML khususnya, dan Kecerdasan Buatan

(AI) secara umum, mengancam akan menghambat kemajuan aplikasi yang bermanfaat, kecuali jika ditangani dengan tepat.

Menyeimbangkan ketegangan antara mendukung inovasi, sehingga hak masyarakat untuk mendapatkan manfaat dari sains terlindungi, dan membatasi potensi bahaya yang terkait dengan AI yang dirancang buruk (dan khususnya ML dalam konteks ini), (dirangkum dalam Tabel 9.1) merupakan tantangan. Algoritma ML adalah konstruksi sosio-teknis yang kuat, yang menimbulkan kekhawatiran yang sama banyaknya (jika tidak lebih) tentang orang-orang seperti halnya tentang kode (lihat Tabel 9.1).

Mengaktifkan apa yang disebut keuntungan ganda dari 'ML etis' sehingga peluang dikapitalisasi, sementara bahayanya diramalkan dan diminimalkan atau dicegah memerlukan pengajuan pertanyaan sulit tentang desain, pengembangan, penerapan, praktik, penggunaan dan pengguna, serta data yang mendorong seluruh siklus hidup algoritma. Lessig memang benar sejak awal: kode merupakan ancaman sekaligus janji terbesar kita. Menghadapi tantangan merancang 'ML etis' sangatlah penting dan memungkinkan. Mereka yang mengklaim hal itu mustahil justru terjebak dalam kekeliruan is-isme, yaitu mereka mencampuradukkan antara apa adanya dengan apa yang seharusnya, atau bahkan seharusnya.

Merancang masyarakat yang ditingkatkan secara algoritmik dan pro-etika adalah mungkin, sehingga melindungi nilai-nilai, prinsip, dan etika yang dianggap fundamental oleh masyarakat. Ini adalah pesan yang disampaikan oleh para ilmuwan sosial, ahli etika, filsuf, pembuat kebijakan, teknolog, dan masyarakat sipil dalam seruan kolektif untuk pengembangan mekanisme tata kelola yang tepat yang akan memungkinkan masyarakat untuk memanfaatkan peluang, sambil memastikan bahwa hak asasi manusia dihormati, dan pengambilan keputusan yang adil dan etis dipertahankan.

Tabel 9.1 Kekhawatiran etika terkait penggunaan algoritma berdasarkan 'peta'

Kekhawatiran etis	Penjelasan
Bukti yang tidak meyakinkan	Kesimpulan algoritmik bersifat probabilitas dan oleh karena itu tidak sempurna. Hal ini dapat menyebabkan tindakan yang tidak dapat dibenarkan. Misalnya, algoritma yang digunakan untuk menilai kelayakan kredit mungkin akurat 99%, tetapi ini tetap berarti bahwa satu dari seratus pemohon akan ditolak kreditnya secara keliru.
Bukti yang tidak dapat dipahami	Kurangnya interpretabilitas dan transparansi dapat menyebabkan sistem algoritmik yang sulit dikontrol, dipantau, dan dikoreksi. Inilah masalah 'kotak hitam' yang sering dikutip.
Bukti yang salah arah	Kesimpulan hanya dapat diandalkan (tetapi juga netral) jika data yang mendasarinya akurat, dan hal ini dapat menyebabkan bias. Misalnya, Dressel dan Farid (2018) menemukan bahwa algoritma residivisme COMPAS yang umum digunakan dalam keputusan praperadilan, pembebasan bersyarat, dan vonis di Amerika Serikat, tidak lebih akurat atau adil daripada prediksi yang dibuat oleh orang-orang dengan sedikit atau tanpa keahlian di bidang peradilan pidana.

Hasil yang tidak adil	Suatu tindakan dapat dianggap diskriminatif jika memiliki dampak yang tidak proporsional terhadap satu kelompok orang. Misalnya, Selbst (2017) mengartikulasikan bagaimana adopsi perangkat kepolisian prediktif menyebabkan lebih banyak orang kulit berwarna ditangkap, dipenjara, atau disakiti secara fisik oleh polisi.
Dampak transformatif	Aktivitas algoritmik, seperti pembuatan profil, dapat menimbulkan tantangan bagi otonomi dan privasi informasi. Misalnya, Polykalas dan Prezerakos (2019) meneliti tingkat akses yang dibutuhkan terhadap data pribadi oleh lebih dari 1000 aplikasi yang terdaftar dalam kategori gratis dan berbayar 'terpopuler' di Google Play Store. Mereka menemukan bahwa aplikasi gratis membutuhkan data yang jauh lebih banyak daripada aplikasi berbayar, dan menunjukkan bahwa model bisnis aplikasi 'gratis' ini adalah eksploitasi data pribadi.
Ketertelusuran	Sulit untuk menetapkan tanggung jawab atas kerugian algoritmik, dan hal ini dapat menimbulkan masalah tanggung jawab moral. Misalnya, mungkin tidak jelas siapa (atau apa) yang bertanggung jawab atas kematian akibat mobil otonom. Analisis etika mendalam mengenai isu spesifik ini disediakan oleh Hevelke dan Nida-Rümelin (2015).

Tujuan dari halaman-halaman berikut adalah untuk menyoroti bagian yang dapat dimiliki oleh para teknolog, atau pengembang ML, dalam percakapan yang lebih luas ini, dan untuk menyoroti di mana penelitian lebih lanjut sangat dibutuhkan. Secara khusus, bagian 'Berpindah dari Prinsip ke Praktik' membahas bagaimana upaya untuk data terlalu terfokus pada 'apa' dari AI etis (yaitu perdebatan tentang prinsip dan kode etik) dan tidak cukup pada 'bagaimana' etika terapan.

Bagian 'Metodologi' menguraikan penelitian yang direncanakan untuk berkontribusi dalam menutup kesenjangan antara prinsip dan praktik ini, melalui penciptaan 'tipologi AI etis terapan', dan metodologi untuk penciptaannya. Bagian 'Membingkai hasil', menyediakan kerangka kerja teoritis untuk menafsirkan hasil. Bagian 'Diskusi hasil awal' merangkum apa yang ditunjukkan oleh tipologi tentang utilitas yang tidak pasti dari alat dan metode yang diidentifikasi serta distribusinya yang tidak merata.

Bagian "Jalan ke Depan" berargumen bahwa diperlukan upaya yang lebih terkoordinasi, dari para peneliti multidisiplin, inovator, pembuat kebijakan, warga negara, pengembang, dan desainer, untuk menciptakan dan mengevaluasi perangkat dan metodologi baru, guna memastikan adanya "bagaimana" untuk setiap "apa" di setiap tahap alur Pembelajaran Mesin. Bagian kedua terakhir mencantumkan beberapa keterbatasan studi ini. Terakhir, bagian terakhir, menyimpulkan bahwa rekomendasi yang disarankan akan sulit dicapai, tetapi tidak bijaksana untuk tidak mencobanya.

9.2 BERALIH DARI PRINSIP KE PRAKTIK

Pada 22 Mei 2019, Organisasi untuk Kerja Sama dan Pembangunan Ekonomi (OECD) mengumumkan bahwa 36 negara anggotanya, bersama dengan enam negara tambahan (Argentina, Brasil, Kolombia, Kosta Rika, Peru, dan Rumania), telah secara resmi sepakat untuk mengadopsi apa yang diklaim OECD sebagai standar antarpemerintah pertama tentang Kecerdasan Buatan (AI). Dirancang untuk memastikan sistem AI tangguh, aman, adil, dan tepercaya, standar ini terdiri dari lima prinsip berbasis nilai yang saling melengkapi, dan lima rekomendasi yang dapat diimplementasikan kepada para pembuat kebijakan.

Nilai dan rekomendasi tersebut bukanlah hal baru. Bahkan, *Rekomendasi Dewan OECD tentang Kecerdasan Buatan* hanyalah yang terbaru di antara daftar lebih dari 70 dokumen, yang diterbitkan dalam 3 tahun terakhir, yang memberikan rekomendasi tentang prinsip-prinsip etika AI. Daftar ini mencakup dokumen-dokumen yang dihasilkan oleh industri (Google, IBM, Microsoft, Intel), Pemerintah (Deklarasi Montreal, Komite Khusus Dewan Bangsa-bangsa, Kelompok Pakar Tingkat Tinggi Komisi Eropa), dan akademisi (Institut Masa Depan Kehidupan, IEEE, AI4People). Harapan para penulis dokumen-dokumen ini adalah bahwa prinsip-prinsip yang diajukan, sebagai abstraksi, dapat bertindak sebagai batasan normatif tentang apa yang boleh dan tidak boleh dilakukan dalam penggunaan algoritmik di masyarakat.

Sebagaimana ditunjukkan oleh Jobin (2019) dan Floridi (2019c), minat yang besar dari beragam pemangku kepentingan ini tidak hanya mencerminkan perlunya panduan etis, tetapi juga keinginan berbagai pihak tersebut untuk membentuk diskusi tentang 'AI etis' berdasarkan prioritas mereka sendiri. Ini adalah masalah yang tidak unik untuk perdebatan tentang komponen-komponen ML etis, tetapi sesuatu yang telah bergulat dengan komunitas hak asasi manusia internasional selama beberapa dekade, karena ketidaksepakatan tentang apa itu, berapa banyak jumlahnya, untuk apa mereka, serta tugas apa yang mereka bebaskan pada siapa, dan nilai-nilai kepentingan manusia mana yang seharusnya mereka lindungi (Arvan 2014), tidak pernah terselesaikan. Oleh karena itu, penting bahwa tampaknya ada konsensus yang muncul di antara anggota komunitas ML etis sehubungan dengan apa sebenarnya yang seharusnya dicita-citakan oleh ML etis.

Tabel 9.2 Perbandingan prinsip-prinsip etika dalam publikasi baru-baru ini menunjukkan adanya konsensus mengenai 'apa' yang seharusnya diharapkan dari AI yang etis

AI4People (diterbitkan November 2018)	Lima prinsip utama untuk setiap kerangka etika AI	Pedoman Etika untuk AI yang Dapat Dipercaya (Diterbitkan April 2019)	Rekomendasi Dewan Kecerdasan Buatan (Diterbitkan Mei 2019)	Prinsip AI Beijing untuk R&D (Diterbitkan Mei 2019)
Kebajikan (Beneficence)	AI harus bermanfaat bagi kemanusiaan	Menghormati otonomi manusia	Pertumbuhan inklusif, pembangunan berkelanjutan, dan kesejahteraan	Berbuat baik: (mencakup kebutuhan AI untuk memajukan masyarakat manusia dan lingkungan)

Non-maleficence (Tidak merugikan)	AI tidak boleh melanggar privasi atau merusak keamanan	Pencegahan kerugian	Ketahanan, keamanan, dan keselamatan	Bertanggung jawab: (mencakup kebutuhan peneliti untuk menyadari dampak negatif dan mengambil langkah untuk mengurangnya). Mengendalikan risiko: (mencakup kebutuhan pengembang untuk meningkatkan ketahanan dan keandalan sistem guna memastikan keamanan data dan keselamatan AI)
Otonomi (Autonomy)	AI harus melindungi dan meningkatkan otonomi kita serta kemampuan untuk mengambil keputusan dan memilih di antara alternatif	Menghormati otonomi manusia	Nilai-nilai yang berpusat pada manusia dan keadilan	Untuk kemanusiaan: (mencakup kebutuhan agar AI melayani kemanusiaan dengan menyesuaikan nilai-nilai manusia termasuk kebebasan dan otonomi)
Keadilan (Justice)	AI harus mendorong kemakmuran dan	Keadilan	Nilai-nilai yang berpusat pada manusia dan keadilan	Menjadi beragam dan inklusif: (mencakup kebutuhan agar AI memberi manfaat bagi sebanyak mungkin orang). Bersikap etis: (mencakup kebutuhan untuk membuat sistem seadil mungkin, meminimalkan diskriminasi dan bias)
Keterjelasan	Sistem AI harus dapat dipahami dan dijelaskan	Keterjelasan	Transparansi dan akuntabilitas keterjelasan	Bersikap etis: (mencakup kebutuhan AI untuk transparan, dapat dijelaskan, dan dapat diprediksi)

Untuk perbandingan yang lebih rinci lihat Floridi dan Cowls (2019) dan Hagendorff (2019) Tinjauan terhadap 84 dokumen AI etis oleh Jobin dkk. (2019) menemukan bahwa meskipun tidak ada prinsip tunggal yang ditampilkan dalam semuanya, tema transparansi, keadilan dan

kewajaran, non-maleficence, tanggung jawab dan privasi muncul di lebih dari setengahnya. Demikian pula, tinjauan sistematis literatur tentang teknologi etis mengungkapkan bahwa tema privasi, keamanan, otonomi, keadilan, martabat manusia, kendali teknologi dan keseimbangan kekuasaan, berulang.

Seperti yang dikemukakan oleh, jika digabungkan tema-tema ini 'mendefinisikan' ML yang selaras dengan etika sebagai yang (a) bermanfaat bagi, dan menghormati, orang-orang dan lingkungan (kebaikan); (b) kuat dan aman (non-maleficence); (c) menghormati nilai-nilai kemanusiaan (otonomi); (d) adil (keadilan); dan (e) dapat dijelaskan, bertanggung jawab, dan dapat dipahami (explicability).

Mengingat konsensus yang muncul dalam literatur ini, tidak mengherankan bahwa ini juga merupakan tema-tema yang menjadi pusat standar OECD. Yang mungkin lebih mengejutkan adalah bahwa kesepakatan seputar prinsip-prinsip dasar yang harus dipenuhi oleh ML yang etis tidak lagi terbatas pada Eropa dan dunia Barat. Hanya 3 hari setelah publikasi OECD, Akademi Kecerdasan Buatan Beijing (BAAI), sebuah organisasi yang didukung oleh Kementerian Sains dan Teknologi Tiongkok dan pemerintah kota Beijing, merilis 15 prinsip AI untuk: (a) penelitian dan pengembangan; (b) penggunaan; dan (c) Tata Kelola AI, yang jika dibaca secara lengkap, memiliki kemiripan yang luar biasa dengan kerangka kerja umum (lihat Tabel 10.2).

Konsensus yang rapuh ini berarti bahwa kini telah terdapat kerangka dasar bersama yang dapat digunakan untuk membangun, dan yang dapat digunakan sebagai tolok ukur untuk mengomunikasikan ekspektasi dan mengevaluasi hasil. Desain bersama dalam AI akan lebih sulit tanpa kerangka kerja bersama ini. Oleh karena itu, hal ini merupakan fondasi penting dalam penciptaan lingkungan yang mendorong pembelajaran mesin yang etis, bertanggung jawab, dan bermanfaat, terutama karena hal ini juga menunjukkan kemungkinan suatu saat risiko distraksi dari belanja etika (*ethical shopping*) akan berkurang. Namun, tantangan tetap ada.

Ketersediaan prinsip-prinsip yang 'disepakati' ini mendukung, tetapi belum membawa perubahan nyata dalam *desain* sistem algoritmik. Sebagaimana dicatat, hampir semua pedoman yang telah dihasilkan hingga saat ini menunjukkan adanya solusi teknis, tetapi sangat sedikit yang memberikan penjelasan teknis. Akibatnya, para pengembang menjadi frustrasi karena betapa sedikitnya bantuan yang ditawarkan oleh prinsip-prinsip yang sangat abstrak dalam hal 'pekerjaan sehari-hari'. Hal ini tercermin dari fakta bahwa 79% pekerja teknologi melaporkan bahwa mereka menginginkan sumber daya praktis untuk membantu mereka dalam pertimbangan etika. Tanpa panduan yang lebih praktis ini, risiko lain seperti '*ethics bluwashing*' dan '*ethics shirking*' tetap ada.

Risiko-risiko tersebut, yang terkait dengan kurangnya panduan praktis tentang cara menghasilkan pembelajaran mesin (ML) yang etis, memperjelas bahwa komunitas ML yang etis perlu memulai fase kedua etika AI: menerjemahkan antara '*apa*' dan '*bagaimana*'. Hal ini kemungkinan akan menjadi pekerjaan yang sulit. Kesenjangan antara prinsip dan praktik sangat besar, dan diperlebar oleh kompleksitas, variabilitas, subjektivitas, dan kurangnya standarisasi, termasuk interpretasi yang bervariasi dari 'komponen' masing-masing prinsip

etika. Namun, hal ini bukan tidak mungkin jika pertanyaan yang tepat diajukan dan perhatian yang lebih cermat diberikan pada bagaimana proses desain dapat memengaruhi apakah suatu algoritma lebih atau kurang 'selaras secara etis'.

Semakin cepat kita mulai melakukan ini, semakin baik. Jika kita tidak menerima tantangan ini dan mengembangkan mekanisme yang dapat digunakan, diinterpretasikan, dan efektif untuk menutup kesenjangan ini, kurangnya panduan dapat (a) mengakibatkan biaya kesalahan etika lebih besar daripada manfaat keberhasilan etika (bahkan satu skandal 'AI' yang kritis pun dapat menghambat inovasi); (b) melemahkan penerimaan publik terhadap sistem algoritmik; (c) mengurangi adopsi sistem algoritmik; dan (d) pada akhirnya menciptakan skenario di mana masyarakat menanggung biaya peluang yang signifikan.

Oleh karena itu, tujuan proyek penelitian ini adalah untuk mengidentifikasi metode dan perangkat yang sudah tersedia untuk membantu para pengembang, insinyur, dan perancang pembelajaran mesin (ML) merefleksikan dan menerapkan 'etika' (Adamson dkk. 2019) agar mereka tidak hanya mengetahui apa yang harus dilakukan atau tidak dilakukan, tetapi juga bagaimana melakukannya, atau menghindarinya. Kami berharap hasil penelitian ini dapat dengan mudah diterapkan pada cabang-cabang AI lainnya.

9.3 METODOLOGI

Dengan tujuan mengidentifikasi metode dan perangkat yang tersedia untuk membantu pengembang, insinyur, dan perancang ML merefleksikan dan menerapkan 'etika', tugas pertama adalah merancang tipologi, untuk komunitas ML yang sangat praktis, yang akan 'menyesuaikan' perangkat dan metode yang diidentifikasi dengan prinsip-prinsip etika yang diuraikan dalam Tabel 9.2 (disingkas sebagai kebaikan, non-maleficence, otonomi, keadilan, dan eksplikabilitas).

Untuk menciptakan tipologi ini, dan terinspirasi oleh Saltz dan Dewar (2019) yang menghasilkan kerangka kerja yang dimaksudkan untuk membantu ilmuwan data mempertimbangkan isu-isu etika di setiap tahap proyek, prinsip-prinsip etika tersebut digabungkan dengan tahapan pengembangan algoritmik yang diuraikan dalam ikhtisar kerangka kerja audit Kantor Komisioner Informasi (ICO) untuk Kecerdasan Buatan dan komponen-komponen intinya,¹⁶ seperti yang ditunjukkan pada Tabel 10.3. Tujuannya adalah agar pengembang ML dapat secara teratur beralih antara prinsip keputusan dan prinsip etika.

Tugas kedua adalah mengidentifikasi alat dan metode, serta perusahaan atau individu yang meneliti dan memproduksinya, untuk mengisi tipologi. Ada sejumlah cara berbeda untuk melakukan hal ini. Misalnya, berusaha menjawab pertanyaan 'praktik, alat, atau metode apa, jika ada, yang digunakan oleh para profesional industri untuk menerapkan etika dalam desain dan pengembangan AI?' dengan melakukan wawancara di lima perusahaan yang mengembangkan sistem AI di berbagai bidang. Namun, meskipun analisis wawancara mengungkapkan bahwa para pengembang menyadari potensi pentingnya etika dalam AI, perusahaan-perusahaan tersebut tampaknya tidak menyediakan alat atau metode untuk menerapkan etika. Berdasarkan hipotesis bahwa temuan ini tidak menyiratkan ketiadaan alat dan metode etika terapan, melainkan kurangnya kemajuan dalam penerapan alat dan metode

yang tersedia dari literatur akademis atau pengembangan dan penelitian tahap awal ke penggunaan di dunia nyata, studi ini menggunakan pendekatan tradisional dalam memberikan penilaian menyeluruh terhadap suatu topik penelitian, yaitu tinjauan pustaka.

Pencarian dilakukan melalui Scopus, arXiv dan PhilPapers, serta Google Search. Pencarian Scopus, arXiv, dan Google Search dilakukan menggunakan istilah-istilah yang diuraikan dalam Tabel 9.4. Pencarian PhilPapers tidak terstruktur, mengingat sifat platformnya, dan sebagai gantinya, kategori-kategori yang juga ditunjukkan dalam Tabel 9.4 ditinjau. Pencarian awal dijalankan pada Februari 2019, tetapi peringatan mingguan ditetapkan untuk semua pencarian dan ditinjau hingga pertengahan Juli 2019.

Setiap hasil (yang awalnya berjumlah lebih dari 1000) diperiksa *relevansinya* baik dari segi kerangka teoretis maupun dari segi penggunaan alat *kemampuan* ditindaklanjuti oleh pengembang ML, dan *generalisasi* di seluruh sektor industri. Secara total, 425 sumber ditinjau. Sumber-sumber tersebut memberikan kontribusi praktis atau teoretis untuk menjawab pertanyaan: 'bagaimana mengembangkan sistem algoritmik yang etis.'

Tabel 9.3 Tipologi 'Etika AI Terapan' yang terdiri dari prinsip-prinsip etika dan tahapan pengembangan algoritma

Pengembangan bisnis dan kasus penggunaan Masalah/perbaikan didefinisikan dan penggunaan AI diusulkan	fase ign Kasus bisnis diubah menjadi persyaratan desain untuk engineering	Pengadaan data pelatihan dan pengujian Kumpulan data awal diperoleh untuk melatih dan menguji model	Membangun aplikasi AI sedang dibangun	Pengujian Sistem sedang diuji	Penerapan Saat sistem AI mulai aktif	pemantauan Kinerja sistem adalah ass
Kemurahan hati						
Tidak bersifat jahat						
Otonomi						
Justice						
Dapat dijelaskan						

Tugas ketiga, dan terakhir, adalah meninjau rekomendasi, teori, metodologi, dan perangkat yang diuraikan dalam sumber yang ditinjau, dan mengidentifikasi di mana mereka dapat masuk dalam tipologi. Untuk melakukan ini, masing-masing prinsip tingkat tinggi (kebaikan, non-maleficence, otonomi, keadilan dan eksplikabilitas) diterjemahkan ke dalam persyaratan sistem nyata yang mencerminkan makna prinsip-prinsip tersebut. Ini adalah pendekatan yang diambil oleh Kelompok Etika Tingkat Tinggi Uni Eropa untuk AI dan diuraikan dalam Bab II Pedoman *Etika untuk AI yang Dapat Dipercaya: Mewujudkan AI yang Dapat Dipercaya* yang "menawarkan panduan tentang implementasi dan realisasi AI yang Dapat Dipercaya, melalui daftar (tujuh) persyaratan yang harus dipenuhi, berdasarkan prinsip-prinsip tersebut".

Pendekatan ini juga digunakan dalam pedoman etika disiplin yang dihasilkan untuk peneliti yang dimediasi internet oleh Laporan Belmont, dan oleh La Fors dkk. yang berupaya mengintegrasikan pendekatan etika berbasis desain yang ada untuk teknologi baru dengan mencocokkan daftar nilai abstraksi praktis dari etika tingkat menengah (prinsip) dengan apa yang disebut 'mikroetika'.

Penerjemahan ini merupakan proses yang secara bertahap mengurangi ketidakpastian norma-norma abstrak untuk menghasilkan desiderata bagi 'produk etika minimum yang layak (ML)' (MVEP) yang dapat digunakan oleh orang-orang dari berbagai latar belakang disiplin ilmu, minat, dan prioritas. Hasil dari proses penerjemahan ini ditunjukkan pada Tabel 9.5.

Tabel 9.4 Menunjukkan istilah pencarian yang digunakan untuk mencari Scopes, arXiv dan Google dan kategori yang ditinjau di PhilPapers

Istilah pencarian di Scopus, Google, dan arXiv (semua dicari dengan machine learning atau Artificial Intelligence)	Kategori PhilPapers yang ditinjau
Etika	Etika informasi
Persepsi publik	Etika teknologi
Hak kekayaan intelektual	Etika komputer
Model bisnis	Otonomi dalam etika terapan
Evaluasi	Kebajikan dalam etika terapan
Berbagi data	Kerugian dalam etika terapan
Penilaian dampak	Keadilan dalam etika terapan
Privasi	Hak asasi manusia dalam etika terapan
Kerugian	Etika terapan dan etika normatif
Legislasi	Tanggung jawab dalam etika terapan
Regulasi	Teori etika dalam etika terapan
Minimalisasi data	
Transparansi	
Bias Perlindungan data	

Tabel 10.5 menunjukkan hubungan antara prinsip etika tingkat tinggi dan persyaratan sistem nyata sebagaimana diadaptasi dari metodologi yang diuraikan dalam Bab II “Pedoman Etika untuk AI yang Dapat Dipercaya” Komisi Eropa

Prinsip	Kemurahan hati	Tidak Mencilakakan	Otonomi	Keadilan	Dapat dijelaskan
Persyaratan	Partisipasi pemangku kepentingan: untuk mengembangkan sistem yang tepercaya dan mendukung	Ketahanan terhadap serangan dan keamanan: Sistem AI harus dilindungi dari kerentanan yang dapat memungkinkannya	Agensi manusia: pengguna harus mampu membuat keputusan otonom	Menghindari bias yang tidak adil Aksesibilitas dan desain universal Masyarakat dan demokrasi: dampak sistem terhadap	Ketertelusuran: kumpulan data dan proses yang menghasilkan keputusan sistem AI harus didokumentasikan.

	kesejahteraan manusia, pihak-pihak yang akan terdampak oleh sistem tersebut harus diajak berkonsultasi. Perlindungan hak asasi manusia. Kecerdasan buatan yang berkelanjutan dan ramah lingkungan: rantai pasok sistem harus dinilai berdasarkan penggunaan sumber daya dan konsumsi energi. Justifikasi: tujuan pembangunan sistem harus jelas dan dikaitkan dengan manfaat yang jelas—sistem tidak boleh dibangun hanya demi manfaat semata.	dieksploitasi oleh musuh. Rencana cadangan dan keamanan umum: Sistem AI harus memiliki perlindungan yang memungkinkan rencana cadangan jika terjadi masalah. Akurasi: misalnya, dokumentasi yang menunjukkan evaluasi apakah sistem mengklasifikasikan hasil . Privasi dan Perlindungan Data: Sistem AI harus menjamin privasi dan perlindungan data di seluruh siklus hidup sistem. Keandalan dan Reprodusibilitas: apakah sistem bekerja dengan cara yang sama dalam berbagai skenario yang berbeda. Kualitas dan integritas data: ketika data dikumpulkan, data tersebut mungkin mengandung bias, ketidakakuratan, kesalahan, dan kekeliruan yang dibangun secara sosial—hal ini perlu ditangani. Dampak Sosial: dampak sistem terhadap kesejahteraan fisik dan mental	yang terinformasi terkait sistem AI. Pengawasan manusia: dapat dicapai melalui mekanisme tata kelola seperti manusia-dalam-lingkaran, manusia-dalam-lingkaran, manusia-dalam-perintah.	institusi, demokrasi, dan masyarakat secara luas harus dipertimbangkan Auditabilitas: pemberdayaan penilaian algoritma, data, dan proses desain. Minimisasi dan pelaporan dampak negatif: langkah-langkah harus diambil untuk mengidentifikasi, menilai, mendokumentasikan, meminimalkan, dan menanggapi potensi dampak negatif sistem AI Trade-off: ketika trade-off antar persyaratan diperlukan, suatu proses harus diterapkan untuk secara eksplisit Mengakui trade-off tersebut, dan mengevaluasinya secara transparan Penanganan: mekanisme harus tersedia untuk merespons ketika terjadi kesalahan	Keterjelasan: kemampuan untuk menjelaskan proses teknis sistem AI dan keputusan manusia terkait. Keterinterpretasian

		manusia harus dipertimbangkan dan dipantau secara cermat.			

Sebagaimana disoroti oleh salah satu peninjau anonim, kategorisasi ini mungkin tampak agak ad-hoc. Misalnya, seseorang dapat bertanya mengapa 'perlindungan hak-hak fundamental' masuk dalam kategori 'kebaikan' alih-alih keadilan atau non-maleficence, dan mengapa 'privasi dan perlindungan data' bukan hak fundamental. Ini merupakan poin penting. Ini adalah pertanyaan yang sangat terbuka dan layak untuk dianalisis secara filosofis lebih lanjut.

Namun, analisis semacam itu berada di luar cakupan makalah ini. Tujuannya di sini bukanlah untuk mengkritik prinsip-prinsip etika itu sendiri, maupun persyaratan sistem untuk memenuhinya sebagaimana ditetapkan oleh Komisi Eropa. Kami hanya berupaya menggunakannya sebagai kerangka kerja yang ada dan menilai sejauh mana pengembang dapat memenuhi persyaratan ini berdasarkan ketersediaan (dan kualitas) perangkat dan metode yang tersedia untuk umum untuk membantu mereka 'patuh'.

9.4 MEMBINGKAI HASIL

Tujuan penyajiannya bukan untuk menyiratkan bahwa tipologi ini 'lengkap', atau bahwa perangkat dan metodologi yang disorot adalah cara terbaik, atau bahkan satu-satunya, untuk 'menyelesaikan' setiap masalah etika individual. Bagaimana menerapkan etika pada pengembangan ML merupakan pertanyaan terbuka yang dapat dipecahkan dengan berbagai cara, pada skala dan konteks yang berbeda.

Proses ini, misalnya, sepenuhnya mungkin diselesaikan menggunakan seperangkat prinsip dan persyaratan yang berbeda. Sebaliknya, tujuannya adalah untuk memberikan sintesis tentang perangkat apa saja yang saat ini tersedia bagi pengembang ML untuk mendorong perkembangan AI etis dari prinsip ke praktik dan untuk memberi sinyal yang jelas kepada komunitas 'AI etis' secara luas, di mana pekerjaan lebih lanjut diperlukan.

Selain itu, tujuan penyajian tipologi ini bukanlah untuk memberi kesan bahwa perangkat tersebut berfungsi sebagai sarana untuk menerjemahkan prinsip-prinsip menjadi 'aturan' definitif yang harus dipatuhi oleh pengembang teknologi, atau bahwa pengembang harus selalu menyelesaikan satu 'tugas' dari setiap kotak yang tersedia. Hal ini hanya mempromosikan etika dengan 'mencentang kotak'.

Sebaliknya, tipologi ini pada akhirnya dimaksudkan untuk menjadi basis data daring yang dapat dicari sehingga pengembang dapat mencari perangkat dan metodologi yang sesuai untuk konteks mereka, dan menggunakannya untuk memungkinkan pergeseran dari pendekatan 'etika berdasarkan desain' yang preskriptif menjadi pendekatan desain yang dialogis dan pro-etika.

Dalam hal ini, perangkat dan metodologi tersebut merepresentasikan versi pragmatis dari etika wacana Habermas. Dalam teorinya, Habermas (1983, 1991) berpendapat bahwa moral dan norma tidak 'ditetapkan' secara top-down tetapi muncul dari suatu proses di mana mereka yang memiliki pandangan yang berlawanan, terlibat dalam suatu proses di mana

mereka secara rasional mempertimbangkan argumen satu sama lain, memberikan alasan untuk posisi mereka dan, berdasarkan pemahaman yang lebih besar yang dihasilkan, menilai kembali posisi mereka sampai semua pihak yang terlibat mencapai keputusan yang disetujui secara universal. Ini adalah pendekatan yang umum digunakan dalam etika penelitian bisnis dan operasional, di mana pertanyaan tentang 'apa yang harus kita lakukan?' (sebagai lawan dari apa yang dapat kita lakukan?) muncul. Ini adalah proses rasionalisasi yang melibatkan pertimbangan yang adil tentang yang praktis, yang baik dan yang adil, dan biasanya sangat bergantung pada bahasa (diskusi), baik untuk munculnya norma atau standar yang disepakati, dan reproduksinya.

Dalam skenario saat ini di mana pengembang merasionalisasi keputusan desain ML untuk memastikan bahwa keputusan tersebut dioptimalkan secara etis, alat dan metode dalam tipologi menggantikan peran bahasa dan bertindak sebagai media untuk mengidentifikasi, memeriksa, menciptakan, dan memeriksa ulang ide serta memberikan pertimbangan yang adil terhadap berbagai kepentingan, nilai, dan norma. Misalnya, alat nutrisi data menyediakan sarana untuk mendorong diskusi dan evaluasi ulang tentang implikasi etis dari penggunaan kumpulan data tertentu untuk proyek pengembangan ML, dan metodologi audit memastikan bahwa suara eksternal, yang mungkin memiliki pandangan yang berlawanan tentang apakah sistem ML yang digunakan selaras secara etis atau tidak, memiliki mekanisme untuk mempertanyakan rasionalitas keputusan desain dan meminta perubahannya jika perlu. Dalam kerangka inilah kami menyajikan ikhtisar temuan kami di bagian selanjutnya.

9.5 INTERPRETASI PERMODELAN AI

Interpretasi hasil tinjauan pustaka dan tipologi yang dihasilkan kemungkinan besar akan bergantung pada konteksnya. Mereka yang memiliki latar belakang disiplin ilmu yang berbeda (teknik, filsafat moral, sosiologi, dll.) akan melihat pola dan makna yang berbeda dalam pola-pola ini. Refleksi multidisiplin semacam ini tentang arti keberadaan atau ketiadaan berbagai alat dan metode, serta fungsinya, perlu didorong. Untuk memulai pembahasan, bagian ini menyoroti tiga judul berikut:

1. ketergantungan yang berlebihan pada 'keterangan';
2. fokus pada kebutuhan untuk 'melindungi' individu daripada kolektif; dan
3. kurangnya kegunaan

Ketiganya saling terkait, tetapi demi kesederhanaan, mari kita analisis masing-masing secara terpisah.

Keterangan sebagai Prinsip yang Menyeluruh

Untuk memulai dengan pengamatan yang paling jelas: ketersediaan alat dan metode tidak merata di seluruh tipologi, baik dalam hal prinsip etika maupun dalam hal tahapan pengembangan. Misalnya, meskipun pengembang yang ingin memastikan algoritma ML mereka 'non-maleficent' memiliki serangkaian alat yang tersedia untuk setiap tahap pengembangan seperti yang disorot pada Tabel 9.6 alat dan metode yang dirancang untuk memungkinkan pengembang memenuhi prinsip atau 'beneficence' hampir semuanya

ditujukan untuk digunakan selama tahap perencanaan awal pengembangan (yaitu fase desain pengembangan bisnis dan kasus penggunaan).

Namun, 'kecenderungan' yang paling terlihat adalah ke arah penjelasan post hoc; dengan mereka yang ingin memenuhi prinsip eksplikabilitas selama fase pengujian memiliki pilihan alat dan metode yang paling beragam.

	Pengembangan bisnis dan kasus penggunaan Masalah/perbaikan didefinisikan dan penggunaan AI diusulkan	Tahap Desain Kasus bisnis diubah menjadi persyaratan desain bagi para insinyur.	Pengadaan data pelatihan dan pengujian Kumpulan data awal diperoleh untuk melatih dan menguji model	Membangun Aplikasi AI dibangun	Pengujian Sistem telah diuji	Penerapan Saat sistem AI mulai aktif	Pemantauan Kinerja sistem dinilai
Kemurahan hati							
Nonmaleficence	Cavoukian dkk. (2010) menguraikan 7 prinsip dasar untuk Privasi berdasarkan Desain: 1. Proaktif, bukan reaktif; preventif, bukan reaktif. 2. Privasi sebagai bawaan 3. Privasi tertanam dalam desain 4. Fungsionalitas penuh 1/4 jumlah positif, bukan jumlah nol 5. Perlindungan siklus hidup menyeluruh 6. Visibilitas dan Transparansi 7. Menghormati privasi pengguna	Oetzel dan Spiekerman (2014) menetapkan penilaian dampak privasi langkah demi langkah (PIA) untuk perusahaan mencapai 'privasi berdasarkan desain'	Antignac dkk. (2016) menyediakan kode Python untuk membuat <i>DataMin</i> (sebuah minimiser data—sebuah praprosesor yang memodifikasi input data untuk memastikan hanya data yang dibutuhkan tersedia untuk program) sebagai	Kolter dan Madry (2018) memberikan pengantar praktis, dari perspektif matematika dan pengodean, terhadap topik ketahanan adversarial dengan gagasan bahwa memungkinan untuk mengklasifikasi pembelajar mendalam	Dennis dkk. (2016) menguraikan metodologi untuk memverifikasi pengambilan keputusan agen otonom untuk memastikan bahwa agen pengendali tidak pernah sengaja membuat pilihan yang diyakininya tidak aman.	(<i>Perangkat Kebijakan Akuntabilitas Algoritmik Institut AI Now</i>) menyediakan daftar pertanyaan yang ingin diajukan oleh para advokat kebijakan dan hukum ketika mempertimbangkan untuk memperkenankan sistem otomatis ke dalam layanan publik dan	Makridan Lambrinou dkk. (2015) menguraikan prosedur audit privasi terstruktur berdasarkan prinsip-prinsip privasi yang paling banyak diadopsi: - Spesifikasi tujuan - Batasan pengumpulan data - Kualitas data

			serangkaian berkas kode sumber Java yang dapat dijalankan pada titik-titik sumber data sebelum mengungkapkan data tersebut.	agar tahan terhadap serangan adversarial: https://adversarial-mltutorial.org/		memberikan panduan terperinci tentang di mana dalam proses pengadaan untuk mengajukan pertanyaan tentang akuntabilitas dan potensi kerugian	-Batasan penyimpanan dan pengungkapan - Pengamanan keamanan - Keterbukaan - Partisipasi individu - Akuntabilitas
Keadilan otonomi dapat dijelaskan							

Pengembang yang ingin memastikan solusi ML mereka memenuhi prinsip non-maleficence dapat memulai dengan prinsip dasar privasi berdasarkan desain untuk memandu ideasi dengan tepat, menggunakan teknik seperti minimisasi data, pelatihan untuk ketahanan adversarial, dan verifikasi pengambilan keputusan dalam fase pelatihan-pembuatan-pengujian, dan diakhiri dengan meluncurkan sistem dengan prosedur audit privasi yang menyertainya.

Kemungkinan ada beberapa alasan untuk hal ini, tetapi ada dua yang menonjol. Pertama, dan lebih sederhana, adalah bahwa 'masalah' 'menafsirkan' keputusan algoritmik tampaknya mudah dipahami dari sudut pandang matematika, sehingga prinsip eksplikabilitas dianggap paling cocok untuk solusi teknis. Kedua, dari perspektif filsafat moral, 'eksplikabilitas' bukanlah prinsip moral seperti keempat prinsip lainnya.

Sebaliknya, 'eksplikabilitas' dapat dipandang sebagai prinsip tingkat kedua, yang menjadi sangat penting dalam komunitas pembelajaran mesin-etika karena, sampai batas tertentu, terkait dengan keempat prinsip lainnya. Memang, terdapat argumen bahwa jika suatu sistem dapat dijelaskan (dapat dijelaskan dan diinterpretasikan), sistem tersebut secara inheren lebih transparan dan oleh karena itu lebih akuntabel dalam hal sifat-sifat pengambilan keputusannya dan sejauh mana sistem tersebut melibatkan pengawasan manusia serta adil, tangguh, dan dapat dibenarkan.

Dengan asumsi sementara bahwa hal ini memang terjadi,²⁵ dan bahwa karena dapat dijelaskan, sistem pembelajaran mesin dapat lebih mudah memenuhi prinsip-prinsip kebaikan, non-maleficence, otonomi, dan keadilan, maka fakta bahwa komunitas pembelajaran mesin yang etis telah berfokus begitu ekstensif pada pengembangan alat untuk penjelasan mungkin

tidak tampak bermasalah. Namun, karena mayoritas alat dan metode yang berada dalam konsentrasi di persimpangan antara kemampuan menjelaskan dan pengujian terutama bersifat statistik, ini akan menjadi pandangan yang sangat mekanistik karena 'solusi' seperti itu misalnya LIME, SHAP, Analisis Sensitivitas tidak benar-benar berhasil membantu pengembang memberikan *penjelasan* yang bermakna yang memberi individu kendali lebih besar atas apa yang *disimpulkan* tentang mereka dari data mereka.

Dengan demikian, keberadaan alat-alat ini paling-paling diperlukan tetapi tidak cukup. Dari perspektif yang lebih humanis dan realistis, untuk memenuhi kelima prinsip tersebut, sebuah sistem perlu dirancang sejak awal agar menjadi sistem sosioteknis yang transparan. Untuk mencapai tingkat transparansi, akuntabilitas, atau eksplikabilitas ini, penting bagi mereka yang menganalisis sistem untuk mampu "memahami apa yang dirancang untuk dilakukan, bagaimana dirancang untuk melakukan itu, dan mengapa dirancang dengan cara tertentu, alih-alih cara lain".

Pengawasan semacam ini hanya akan mungkin dilakukan melalui kombinasi alat atau proses yang memfasilitasi audit, pengembangan yang transparan, edukasi publik, dan kesadaran sosial para pengembang. Oleh karena itu, idealnya harus ada alat dan metode yang tersedia untuk setiap kotak dalam tipologi, dengan menerima bahwa mungkin ada area tipologi yang lebih signifikan bagi praktisi ML daripada yang lain.

Lebih lanjut, ketersediaan alat dan metode dalam berbagai area tipologi juga penting dalam konteks etika ML yang spesifik secara budaya dan kontekstual. Tidak semua prinsip akan sama pentingnya dalam semua konteks. Misalnya, dalam kasus sistem keamanan nasional, non-maleficence mungkin jauh lebih penting daripada eksplikabilitas. Jika komunitas memprioritaskan pengembangan perangkat dan metode untuk salah satu prinsip di atas prinsip lainnya, hal itu akan menghilangkan peluang fleksibilitas tersebut.

Fokus Individu

Observasi penting berikutnya adalah bahwa hanya sedikit perangkat lunak yang disurvei yang menyediakan cara yang bermakna untuk menilai, dan merespons, dampak pemrosesan data yang terlibat dalam algoritma ML mereka terhadap individu, dan bahkan lebih sedikit lagi terhadap dampaknya terhadap masyarakat secara keseluruhan. Hal ini terlihat dari kolom 'penerapan' yang sangat jarang pada tipologi tersebut. Kekosongannya menyiratkan bahwa kebutuhan akan interaksi manusia-komputer yang dirancang secara pro-etika (pada tingkat individu) atau jaringan sistem ML (pada tingkat kelompok) kurang diperhatikan. Hal ini kemungkinan besar karena sulitnya menerjemahkan perilaku manusia yang kompleks ke dalam perangkat lunak desain yang mudah digunakan dan dapat digeneralisasi.

Hal ini mungkin tampak tidak terlalu penting, tetapi dampaknya terhadap penerimaan AI secara keseluruhan di masyarakat bisa jadi signifikan. Misalnya, tidak mungkin bahwa penjelasan kontrafaktual (yaitu jika variabel input x berbeda, variabel output y juga akan berbeda) meskipun penting karena banyak alasan akan cukup untuk meningkatkan interpretabilitas rekomendasi yang dibuat oleh sistem kotak hitam untuk rata-rata anggota masyarakat atau komunitas teknis.

Jika metode tersebut menjadi cara *de facto* untuk memberikan penjelasan, sejauh mana 'masyarakat algoritmik' dapat diinterpretasikan oleh masyarakat umum akan sangat terbatas. Dan penjelasan kontrafaktual dapat dengan mudah diterima oleh aktor yang tidak tertarik dalam memberikan penjelasan faktual, karena penjelasan kontrafaktual menyediakan menu pilihan yang luas, yang dapat dengan mudah mengurangi tingkat tanggung jawab aktor yang memilihnya.

Misalnya, jika penyedia hipotek tidak menawarkan hipotek, alasan faktualnya mungkin bias, misalnya jenis kelamin pemohon, tetapi penyedia dapat memilih dari menu luas penjelasan kontrafaktual yang tidak berbahaya jika beberapa variabel *x* berbeda, hipotek mungkin telah diberikan misalnya, pendapatan yang jauh lebih tinggi, lebih banyak agunan, jumlah yang lebih rendah, dan sebagainya, tanpa pernah menyebutkan penyebab faktualnya, yaitu jenis kelamin pemohon.

Semua ini dapat secara signifikan membatasi tingkat kepercayaan yang bersedia diberikan orang pada sistem semacam itu. Potensi ancaman terhadap kepercayaan ini semakin diperparah oleh fakta bahwa kurangnya perhatian yang diberikan pada dampak berarti bahwa pengembang ML saat ini terhambat dalam kemampuan mereka untuk mengembangkan sistem yang mempromosikan otonomi pengguna (individu atau kelompok).

Misalnya, saat ini ada asumsi bahwa prediksi $\frac{1}{4}$ keputusan, dan sedikit penelitian telah dilakukan (dalam konteks ML) tentang bagaimana orang menerjemahkan prediksi menjadi keputusan yang dapat ditindaklanjuti. Oleh karena itu, alat yang, misalnya, membantu pengembang merancang solusi secara pro-etika yang tidak terlalu membatasi pilihan pengguna dalam bertindak berdasarkan prediksi ini (yaitu alat yang mempromosikan otonomi pengguna) masih sangat terbatas.

Jika pengguna merasa keputusan mereka dibatasi dan dikendalikan oleh sistem yang tidak mereka pahami, sangat kecil kemungkinan sistem ini akan memenuhi syarat penerimaan sosial, apalagi syarat preferensi sosial yang seharusnya menjadi tujuan pembelajaran mesin (ML) yang benar-benar dirancang secara etis.

Kurangnya Kegunaan

Terakhir, perangkat dan metode yang termasuk dalam tipologi ini diposisikan sebagai alat bantu wacana, yang dirancang untuk memfasilitasi dan mendokumentasikan keputusan rasional tentang trade-off dalam proses desain yang dapat membuat sistem ML lebih atau kurang selaras secara etis. *Potensi* perangkat yang diidentifikasi untuk memainkan peran ini dapat dilihat. Misalnya, pada persimpangan "kebaikan ! kasus penggunaan ! desain", terdapat sejumlah perangkat yang disorot untuk membantu mengungkap nilai-nilai sosial. Ini termasuk metodologi penelitian dan inovasi yang bertanggung jawab yang digunakan oleh Proyek Otak Manusia Komisi Eropa, panduan lapangan untuk desain yang berpusat pada manusia, dan panduan Involve dan DeepMind tentang merangsang keterlibatan publik yang efektif dalam etika Kecerdasan Buatan.

Perangkat dan metode tersebut dapat digunakan untuk membantu para desainer menangani pluralisme nilai (yaitu variasi nilai di berbagai kelompok populasi) secara pro-etika. Namun, sebagian besar alat dan metode ini tidak dapat ditindaklanjuti karena hanya

memberikan sedikit bantuan tentang cara penggunaannya dalam praktik. Bahkan ketika tersedia pustaka kode sumber terbuka, dokumentasinya seringkali terbatas, dan tingkat keahlian yang dibutuhkan untuk menggunakannya tinggi.

Kurangnya kegunaan alat dan metode yang disorot dalam tipologi ini secara menyeluruh berarti bahwa, meskipun menjanjikan, alat dan metode tersebut membutuhkan lebih banyak pekerjaan sebelum 'siap produksi'. Akibatnya, penerapan etika masih membutuhkan upaya yang cukup besar dari para pengembang ML, yang justru melemahkan salah satu tujuan utama pengembangan dan penggunaan 'alat' berbasis teknologi: menghilangkan hambatan dari etika terapan. Lebih lanjut, hingga alat-alat ini tertanam dalam praktik dan diuji di 'dunia nyata', masih sangat belum jelas dampaknya terhadap 'keteraturan' ekosistem algoritmik secara keseluruhan.

Misalnya, Binns (2018a) mempertanyakan bagaimana sistem yang akuntabel akan dimintai pertanggungjawaban atas keputusan yang 'tidak adil' dengan cara yang dapat diterima oleh semua pihak. Hal ini membuat hampir mustahil untuk mengukur dampak, 'mendefinisikan kesuksesan', dan mendokumentasikan kinerja dari sebuah metodologi atau alat desain baru.

Akibatnya, tidak ada rumusan masalah yang jelas (dan oleh karena itu tidak ada studi kasus bisnis yang jelas) yang dapat digunakan komunitas pembelajaran mesin (ML) untuk membenarkan waktu dan investasi finansial dalam mengembangkan alat dan teknik yang sangat dibutuhkan yang benar-benar memungkinkan desain yang pro-etika. Akibatnya, tidak ada jaminan bahwa apa yang disebut perangkat diskursif ini hanya akan membantu kelompok-kelompok masyarakat yang sudah memiliki suara paling lantang untuk menanamkan dan melindungi nilai-nilai mereka dalam perangkat desain, dan kemudian ke dalam sistem ML yang dihasilkan.

9.6 PANDANGAN MASA DEPAN

Ilmuwan sosial dan filsuf politik (dari Rousseau dan Kant, hingga Rawls dan Habermas), terbiasa menghadapi pluralitas dan subjektivitas yang melandasi seluruh bidang pembelajaran mesin (ML) yang etis. Menjawab pertanyaan seperti, apa yang terjadi ketika 'etika' tingkat individu dan tingkat kelompok berinteraksi, dan apa arti sebenarnya dari istilah-istilah kunci seperti 'keadilan,' 'akuntabilitas,' 'transparansi' dan 'interpretabilitas' ketika saat ini ada segudang definisi adalah hal yang standar bagi individu dengan pelatihan ilmu sosial, ekonomi, filsafat atau hukum. Inilah sebabnya menganjurkan penjelasan kontekstual tentang privasi, yang mengakui beragamnya sifat norma informasi dan menyatakan bahwa transparansi hanya bermakna dalam konteks audiens kritis yang telah ditentukan.

Sebaliknya, komunitas pengembang pembelajaran mesin (ML) mungkin kurang terbiasa menghadapi kesulitan semacam ini, dan lebih terbiasa dengan skenario di mana setidaknya terdapat hubungan yang tampak terukur antara masukan dan keluaran. Akibatnya, pendekatan yang ada untuk merancang dan memprogram ML yang etis gagal menyelesaikan apa yang disebut sebagai trilema moral-semantik, karena hampir semua alat dan metode yang

disorot dalam tipologi tersebut terlalu ketat secara semantik, terlalu fleksibel secara semantik, atau terlalu tidak terduga.

Menjembatani para peneliti multidisiplin ke dalam proses pengembangan alat dan metodologi desain yang pro-etika akan menjadi penting. Pendekatan multidisiplin akan membantu komunitas pembelajaran mesin (ML) yang etis mengatasi hambatan terkait kompleksitas sosial, merangkul ketidakpastian, dan menerima bahwa: (1) AI dibangun di atas asumsi; (2) perilaku manusia itu kompleks; (3) algoritma dapat memiliki konsekuensi yang tidak adil; (4) prediksi algoritmik sulit diinterpretasikan; (5) trade-off biasanya tak terelakkan; dan (6) fitur-fitur etis yang positif terbuka untuk peningkatan progresif, yaitu suatu algoritma dapat semakin adil, dan lebih adil daripada algoritma lain atau versi sebelumnya, tetapi tidak masuk akal untuk mengatakan bahwa algoritma tersebut adil atau tidak adil secara absolut (bandingkan hal ini dengan kasus kecepatan: masuk akal untuk mengatakan bahwa suatu objek bergerak cepat, atau bahwa ia cepat atau lebih cepat daripada yang lain, tetapi tidak masuk akal untuk mengatakan bahwa ia cepat).

Kolaborasi yang dihasilkan kemungkinan besar akan sangat bermanfaat bagi pengembangan perangkat dan metodologi etika terapan setidaknya karena tiga alasan. Pertama, hal ini akan membantu memastikan bahwa perangkat dan metode yang dikembangkan tidak hanya melindungi pluralisme nilai secara *in silico* (yaitu nilai-nilai pluralistik pengembang), tetapi juga dalam masyarakat. Merangkul ketidakpastian dan keberagaman disiplin ilmu secara alami akan mendorong para pakar ML untuk mengembangkan perangkat yang memfasilitasi pertanyaan yang lebih mendalam dan terbuka (yaitu filosofis) yang akan menghasilkan jawaban yang lebih bernuansa dan beralasan, sehingga menghasilkan keputusan tentang mengapa dan kapan trade-off tertentu, misalnya, antara akurasi dan interpretabilitas, dibenarkan, berdasarkan faktor-faktor seperti proporsionalitas terhadap risiko.

Kedua, hal ini akan mendorong pendekatan yang lebih fleksibel dan reflektif terhadap etika terapan yang lebih sesuai dengan cara sistem ML sebenarnya dikembangkan: bukan berpikir lalu membuat kode, melainkan berpikir *dan* membuat kode. Dengan kata lain, hal ini akan mempercepat peralihan dari pendekatan 'bergerak cepat dan merusak segalanya' menuju pendekatan 'bersegeralah perlahan' (*festina lente*). Terakhir, hal ini juga akan memitigasi risiko signifikan yang ditimbulkan oleh penerapan perangkat dan/atau metode desain etis yang sporadis saat ini selama berbagai tahap pengembangan bahwa prinsip-prinsip etika telah tertulis dalam bisnis dan kasus penggunaan, tetapi telah dikodekan pada saat sistem diterapkan.

Agar pengembang dapat merangkul ketidakpastian yang rentan ini, penting untuk mempromosikan pengembangan perangkat, seperti acara pemindaian konsekuensi *agile* *DotEveryone*, dan *Responsible Double Diamond* 'R2D2' yang mendorong pengembang untuk merefleksikan dampak (baik langsung maupun tidak langsung) dari solusi yang mereka kembangkan terhadap 'pengguna akhir', dan bagaimana dampak ini dapat diubah oleh keputusan desain yang tampaknya kecil di setiap tahap pengembangan. Dengan kata lain, pengembang ML harus secara teratur:

- (a) melihat ke belakang dan bertanya: 'Jika saya mematuhi prinsip-prinsip etika x dalam desain saya saat itu, apakah saya masih mematuhi *sekarang*? (sebagaimana didorong oleh metodologi agile *Wellcome Data Lab*); dan
- (b) melihat ke depan dan bertanya: 'Jika saya mematuhi prinsip-prinsip etika x dalam desain saya *sekarang*, haruskah saya terus melakukannya? Dan bagaimana? Dengan menggunakan metodologi foresight, seperti Kerangka Kerja Penilaian Dampak Algoritmik *AI Now*.

Mengambil pendekatan ini mengakui bahwa, dalam konteks digital, prinsip-prinsip etika tidak hanya diterapkan atau tidak, tetapi diterapkan kembali secara teratur atau diterapkan secara berbeda, atau lebih baik, atau diabaikan ketika sistem algoritmik dikembangkan, diterapkan, dikonfigurasi diuji, direvisi dan disetel ulang .

Pendekatan terhadap etika ML terapan dari refleksi dan penerapan teratur ini akan sangat bergantung pada (i) penciptaan lebih banyak alat terutama untuk mengisi ruang kosong tipologi (untuk alasan yang dibahas di bagian sebelumnya) dan (ii) percepatan tingkat kematangan alat dari laboratorium penelitian ke lingkungan produksi. Untuk mencapai (i)–(ii), masyarakat perlu bersatu dalam komunitas yang terdiri dari peneliti multidisiplin, termasuk inovator, pembuat kebijakan, warga negara, pengembang, dan desainer, untuk mendorong pengembangan: (1) pengetahuan dan pemahaman bersama; dan (2) tujuan bersama yang ingin dicapai dari pengembangan perangkat dan metodologi untuk etika AI terapan.

Keluaran ini akan memberikan alasan, mekanisme, dan konsensus untuk mengoordinasikan upaya di balik pengembangan perangkat. Pada akhirnya, hal ini akan menghasilkan hasil yang lebih baik daripada pendekatan saat ini, yang memungkinkan 'seribu bunga bermekaran' tetapi gagal menciptakan perangkat yang mengisi celah (ini adalah kegagalan 'pasar intelektual' yang umum), dan dapat mendorong persaingan untuk menghasilkan opsi yang lebih baik. Peluang yang dihadirkan hal ini terlalu besar untuk ditunda, komunitas riset ML harus mulai berkolaborasi sekarang dengan fokus khusus pada:

1. pengembangan bahasa yang sama;
2. penciptaan perangkat yang memastikan orang, sebagai individu, kelompok, dan masyarakat, diberi kesempatan yang setara dan bermakna untuk berpartisipasi dalam perancangan solusi algoritmik di setiap tahap pengembangan;
3. evaluasi perangkat yang saat ini ada sehingga apa yang berhasil, apa yang dapat ditingkatkan, dan apa yang perlu dikembangkan dapat diidentifikasi;
4. komitmen terhadap reproduktifitas, keterbukaan, dan berbagi pengetahuan serta solusi teknis (misalnya perangkat lunak), juga dengan tujuan memenuhi (2) dan mendukung (3); penciptaan 'contoh kerja' tentang bagaimana perangkat telah digunakan untuk memenuhi salah satu prinsip di setiap tahap pengembangan dan bagaimana konsistensi dipertahankan selama penggunaan perangkat yang berbeda.
5. evaluasi dan penciptaan model bisnis pro-etika dan struktur insentif yang menyeimbangkan biaya dan imbalan investasi dalam AI yang etis di seluruh masyarakat, juga dengan tujuan mendukung (2)–(4).

9.7 KETERBATASAN AI PUBLIK

Semua proyek penelitian memiliki keterbatasannya masing-masing, dan proyek ini tidak terkecuali. Pertama, pertanyaan penelitian 'alat dan metode apa yang tersedia bagi pengembang ML untuk 'menerapkan' etika pada setiap tahap desain sistem ML' sangat luas. Kurangnya spesifisitas ini menyebabkan literatur yang tersedia berlebihan dan terus bertambah, sehingga menimbulkan kompromi dari perspektif hal yang esensial secara praktis.

Kompromi semacam itu, misalnya basis data mana yang harus dicari dan keputusan untuk membatasi alat yang ditinjau hanya pada alat yang tidak spesifik untuk sektor industri tertentu, telah mengakibatkan kami kehilangan banyak alat dan metode yang tersedia untuk umum. Berdasarkan hal ini, sekali lagi, sangat mungkin terdapat sejumlah alat dan metode etika terapan yang dikembangkan oleh perusahaan swasta untuk keperluan internal atau konsultasi yang akan kami lewatkan, misalnya 'rangkaian kerangka kerja, alat, dan proses yang dapat disesuaikan' yang membentuk "Responsible AI Toolkit" dari perusahaan konsultan PWC.

Keterbatasan kedua terkait dengan desain tipologi itu sendiri. Sebagaimana buktikan, "perbedaan teoretis yang jelas antara berbagai tahap inovasi teknologi tidak selalu ada dalam praktik, terutama dalam pengembangan teknologi big data." Ini menyiratkan bahwa dengan mengkategorikan alat berdasarkan tahap pengembangan, kita mungkin mengurangi kegunaannya karena pengembang dalam konteks yang berbeda mungkin mengikuti pola yang berbeda atau merasa seolah-olah "terlambat" untuk, misalnya, terlibat dalam pelibatan pemangku kepentingan jika mereka telah mencapai fase "pembuatan" proyek mereka, padahal kenyataannya tidak ada kata terlambat.

Akhirnya, batasan terakhir telah disebutkan dan menyangkut kurangnya kejelasan mengenai bagaimana perangkat dan metode yang telah diidentifikasi akan meningkatkan tata kelola sistem algoritmik. *Bagaimana* tepatnya mengatur ML masih menjadi pertanyaan terbuka, meskipun tampaknya ada penerimaan yang semakin meningkat di kalangan pekerja teknologi (setidaknya di Inggris) bahwa regulasi pemerintah akan diperlukan.

Tipologi ini setidaknya dapat dilihat sebagai mekanisme untuk memfasilitasi ko-regulasi. Pemerintah semakin menetapkan standar dan persyaratan sistem untuk ML yang etis, tetapi mendelegasikan cara untuk memenuhinya kepada pengembang sendiri perangkat dan metode tipologi ini dapat dilihat sebagai cara untuk memberikan bukti kepatuhan. Dengan cara ini, tipologi (dan perangkat serta metode yang dikandungnya) membantu pengembang bertanggung jawab untuk menanamkan etika dalam bagian pengembangan, penerapan, dan penggunaan solusi ML yang mereka kendalikan. Sejauh mana hal ini membuat perbedaan masih harus ditentukan.

9.8 KESIMPULAN

Kesadaran bahwa pertimbangan etika perlu diintegrasikan ke dalam desain artefak komputasional, khususnya algoritmik, bukanlah hal baru. Samuel (1960), Wiener (1961), dan Turing telah menyuarakan hal ini secara terbuka pada tahun 1940-an dan 1960-an. Namun, seiring meningkatnya kompleksitas sistem algoritmik dan ketergantungan kita padanya, kebutuhan untuk bersikap kritis terhadap tata kelola AI dan merancang solusi pun meningkat.

Merancang sesuatu menjadi lebih baik memang memungkinkan, tetapi hal ini membutuhkan pendekatan yang lebih terkoordinasi dan canggih untuk menerjemahkan prinsip-prinsip etika ke dalam protokol desain.

Seruan untuk meningkatkan koordinasi ini diperlukan. Penelitian menunjukkan bahwa terdapat distribusi upaya yang tidak merata di seluruh tipologi 'Etika AI Terapan'. Lebih lanjut, banyak alat yang disertakan relatif belum matang. Hal ini menyulitkan untuk menilai ruang lingkup penggunaannya dan akibatnya sulit untuk mendorong adopsi mereka oleh pengembang ML yang berpikiran praktis, terutama ketika keunggulan kompetitif AI yang lebih selaras dengan etika belum jelas.

Meluangkan waktu untuk menyelesaikan 'latihan' apa pun yang disarankan oleh metode yang ditinjau, dan berinvestasi dalam pengembangan alat atau metode baru yang 'menyelesaikan alur kerja', menambah pekerjaan dan biaya tambahan pada proses penelitian dan pengembangan. Biaya overhead tersebut dapat secara langsung bertentangan dengan insentif komersial jangka pendek. Memang, pendekatan etis penuh terhadap desain, pengembangan, penerapan, dan penggunaan AI dapat mewakili kerugian kompetitif bagi setiap 'penggerak pertama'.

Ancaman yang ditimbulkan oleh jangka pendekisme ini terhadap pengembangan ML yang benar-benar etis sangatlah signifikan. Kecuali perspektif jangka panjang dan sektoral dalam hal laba atas investasi dapat didorong sehingga mekanisme dikembangkan untuk menutup kesenjangan antara *apa* dan *bagaimana* kurangnya panduan dapat (a) mengakibatkan biaya kesalahan etika lebih besar daripada manfaat keberhasilan etika; (b) melemahkan penerimaan publik terhadap sistem algoritmik, bahkan sampai pada titik reaksi balik; dan (c) mengurangi adopsi sistem algoritmik.

Kurangnya adopsi yang dihasilkan kemudian dapat berubah menjadi hilangnya kepercayaan dari investor dan penyandang dana penelitian, dan melemahkan penelitian AI. Kurangnya insentif untuk mengembangkan AI secara etis dapat berubah menjadi kurangnya minat dalam mengembangkan AI secara langsung. Ini bukan hal yang belum pernah terjadi sebelumnya. Kita hanya perlu mengingat pengurangan dramatis dalam pendanaan yang tersedia untuk penelitian AI setelah publikasi *Artificial Intelligence: A General Survey* tahun 1973 dan kritiknya terhadap fakta bahwa penelitian AI tidak memenuhi harapan yang terlalu dibesar-besarkan.

Jika hal ini terjadi hari ini, biaya peluang yang akan ditanggung masyarakat akan signifikan. Kebutuhan akan 'Etika AI' muncul dari fakta bahwa sistem AI yang dirancang dengan buruk dapat menyebabkan kerugian yang sangat signifikan. Misalnya, alat kepolisian prediktif dapat menyebabkan lebih banyak orang kulit berwarna ditangkap, dipenjara, atau dilukai secara fisik oleh kebijakan. Demikian pula, potensi manfaat dari sistem AI yang dirancang secara pro-etika sangat besar. Hal ini terutama berlaku di bidang AI untuk Kebaikan Sosial di mana berbagai aplikasi AI memungkinkan hasil yang baik secara sosial yang dulunya sulit dicapai, tidak layak, atau tidak terjangkau. Jadi, ada kebutuhan mendesak untuk memajukan penelitian di bidang ini.

Kesabaran yang konstruktif perlu dipraktikkan, oleh masyarakat dan oleh komunitas AI yang etis, karena kemajuan seperti itu pada pertanyaan tentang 'bagaimana' memenuhi 'apa' tidak akan cepat, dan pasti akan ada kesalahan di sepanjang jalan. Komunitas riset pembelajaran mesin (ML) harus menerima hal ini, percaya bahwa semua orang berusaha mencapai tujuan akhir yang sama, tetapi juga menerima bahwa menunda komitmen penuh tidak dapat diterima, padahal konsekuensi dari tidak melakukan apa pun sudah diketahui. Hanya dengan menerima hal ini, masyarakat dapat bersikap positif terhadap peluang yang dihadirkan oleh AI untuk dimanfaatkan, sekaligus tetap memperhatikan potensi kerugian yang harus dihindari.

BAB 10

KERANGKA PEMBELAJARAN MESIN YANG DIINTERPRETASIKAN

Perkembangan pembelajaran mesin (*machine learning*) telah menghasilkan berbagai model prediktif yang mampu mencapai tingkat akurasi tinggi. Namun, salah satu tantangan utama yang muncul adalah bagaimana menjelaskan proses dan hasil prediksi tersebut agar dapat dipahami oleh manusia. Isu keterjelasan (*interpretability*) menjadi semakin penting, terutama ketika model digunakan dalam pengambilan keputusan yang berdampak besar, seperti di bidang kesehatan, keuangan, maupun kebijakan publik.

Bab ini membahas sebuah kerangka kerja formal yang dirancang guna meningkatkan interpretabilitas pembelajaran mesin. Kerangka kerja ini mengintegrasikan prinsip-prinsip dari pembelajaran statistik, intervensi kausal, dan teori keputusan. Melalui pendekatan tersebut, diperkenalkan konsep “permainan penjelasan ideal,” yaitu sebuah mekanisme kolaboratif yang membantu menemukan penjelasan terbaik atas suatu prediksi algoritmik.

Proses penjelasan berlangsung secara iteratif melalui tanya-jawab yang memungkinkan pencarian titik keseimbangan antara tiga aspek penting: akurasi, kesederhanaan, dan relevansi penjelasan. Pendekatan ini juga memberikan ruang bagi eksplorasi pola kausal dengan tingkat kedalaman dan cakupan yang berbeda, sehingga menghasilkan pemahaman yang lebih kaya terhadap prediksi model.

Selain itu, bab ini menguraikan kondisi yang memungkinkan permainan penjelasan tersebut mencapai titik konvergensi pada penjelasan yang optimal dalam waktu yang efisien. Namun, terdapat pula hambatan yang berpotensi menghalangi tercapainya pemahaman yang lebih mendalam. Dengan demikian, kerangka ini tidak hanya berfungsi sebagai alat analisis deskriptif, tetapi juga sebagai acuan normatif untuk menilai, membandingkan, serta merancang pendekatan penjelasan baru yang lebih baik.

10.1 PENDAHULUAN

Algoritma pembelajaran mesin (ML) telah mencapai kemajuan pesat dalam berbagai tugas hanya dalam beberapa tahun terakhir. Beberapa contoh penting terbaru antara lain penguasaan permainan informasi sempurna seperti catur dan Go, diagnosis kanker kulit, dan pengajuan molekul organik baru. Pencapaian teknis ini bertepatan dengan semakin meluasnya penggunaan ML, yang kini digunakan secara luas di sektor publik dan swasta untuk berbagai hal, mulai dari rekomendasi film dan analisis olahraga hingga genomika dan pengawasan prediktif.

Algoritma ML diperkirakan akan terus meningkat seiring dengan meningkatnya efisiensi perangkat keras dan semakin besarnya kumpulan data, yang menyediakan semua komponen yang dibutuhkan para insinyur untuk menciptakan model yang lebih canggih untuk deteksi dan pemrosesan sinyal. Kemajuan terbaru dalam pembelajaran mesin (ML) telah memunculkan sejumlah pertanyaan mendesak mengenai status epistemik keluaran algoritmik.

Salah satu topik yang paling hangat diperdebatkan dalam wacana yang sedang berkembang ini adalah peran kemampuan menjelaskan (*explainability*).

Karena banyak model berkinerja terbaik, seperti jaringan saraf dalam (*deep neural network*), pada dasarnya merupakan kotak hitam sistem yang sangat kompleks yang dioptimalkan untuk akurasi prediktif, bukan kejelasan pengguna beberapa pihak khawatir bahwa teknologi ini mungkin tidak sesuai untuk aplikasi yang sensitif dan berisiko tinggi. Seruan akan algoritma yang lebih mudah dijelaskan sangat mendesak di bidang-bidang seperti kedokteran klinis dan operasi militer, di mana kepercayaan pengguna sangat penting dan kesalahan dapat berakibat fatal.

Hal ini telah mendorong sejumlah kerangka kebijakan internasional yang merekomendasikan kemampuan menjelaskan sebagai persyaratan untuk setiap sistem ML. Kemampuan menjelaskan dengan cepat menjadi prioritas utama dalam penelitian statistik, yang sering disingkat sebagai xAI (*Explainable Artificial Intelligence*) atau iML (*Interpretable Machine Learning*). Kami mengadopsi inisialisasi yang terakhir di sini untuk menekankan fokus kami pada algoritma pembelajaran terawasi dibandingkan dengan aplikasi kecerdasan buatan lain yang lebih umum.

Beberapa komentator berpendapat bahwa tujuan utama iML kurang spesifik. Mereka menyuarakan kekhawatiran tentang subjektivitas yang tak tereduksi dari keberhasilan penjelasan, sebuah konsep yang menurut mereka kurang terdefinisi dan sulit atau bahkan mustahil untuk diukur. Dalam artikel ini, kami akan membahas masalah ini secara langsung. Kami menyediakan kerangka kerja formal untuk mengonseptualisasikan tujuan dan kendala sistem iML dengan merancang permainan penjelasan yang diidealkan. Model kami mengklarifikasi trade-off yang melekat dalam setiap solusi iML, dan mengkarakterisasi kondisi di mana agen epistemik hampir pasti akan konvergen pada serangkaian penjelasan optimal dalam waktu polinomial.

Permainan ini memiliki fungsi deskriptif dan normatif, yang membangun ruang konseptual untuk menganalisis dan membandingkan proposal yang ada, serta merancang solusi baru dan yang lebih baik. Sisa makalah ini disusun sebagai berikut. Pada Bagian 10.2, kami mengidentifikasi tiga tujuan iML yang berbeda. Pada Bagian 10.3, kami meninjau materi latar belakang yang relevan. Kami mengklarifikasi ruang lingkup proposal kami pada Bagian 10.4. Pada Bagian 10.5, kami mengartikulasikan aturan permainan penjelasan dan menguraikan prosedurnya dalam pseudocode. Pembahasan selanjutnya akan dibahas pada Bagian 10.6. Kami mempertimbangkan lima keberatan pada Bagian 10.7, sebelum menyimpulkan pada Bagian 10.8.

10.2 MENGAPA MENJELASKAN ALGORITMA?

Kami menyoroti tiga tujuan yang memandu mereka yang bekerja di iML: *mengaudit*, *memvalidasi*, dan *menemukan*. Tujuan-tujuan ini membantu memotivasi dan memfokuskan diskusi, memberikan tipologi intuitif untuk jenis penjelasan yang mungkin kita cari dan hargai dalam konteks ini. Argumen tandingan terhadap proyek iML ditunda hingga Bagian 10.7.

Keadilan sebagai Keadilan (Algoritmik)

Mungkin alasan paling populer untuk menjelaskan algoritma adalah dampak sosialnya yang besar dan terus berkembang. ML telah digunakan untuk membantu mengevaluasi aplikasi pinjaman dan penerimaan mahasiswa, memprediksi residivisme kriminal, dan mengidentifikasi target militer, sebagai contoh yang kontroversial. Kegagalan menyaring set data pelatihan dengan tepat untuk input yang bias mengancam otomatisasi ketidakadilan yang sudah ada di masyarakat.

Misalnya, studi menunjukkan bahwa profil algoritmik secara konsisten menunjukkan iklan daring untuk pekerjaan bergaji lebih tinggi kepada pria daripada wanita; bahwa perangkat lunak pengenalan wajah sering kali dilatih pada subjek yang didominasi kulit putih, sehingga menjadikannya pengklasifikasi yang tidak akurat untuk wajah berkulit hitam dan cokelat; dan bahwa pemberi pinjaman predator menggunakan data keuangan untuk secara tidak proporsional menyasar komunitas miskin. Para kritikus menyoroti kegagalan-kegagalan ini dan berpendapat bahwa terdapat minimnya keadilan, akuntabilitas, dan transparansi dalam ML yang secara kolektif disingkat FAT ML, sebuah konferensi tahunan tentang topik tersebut yang mulai diselenggarakan pada tahun 2014.

Para pendukung FAT ML hanya sedikit terhibur oleh Peraturan Perlindungan Data Umum (GDPR) Uni Eropa tahun 2018, yang memuat ketentuan yang menyiratkan apa yang disebut "hak atas penjelasan" bagi warga negara yang menjadi subjek keputusan otomatis. Terlepas dari apakah GDPR benar-benar menjamin hak tersebut beberapa komentator bersikeras bahwa GDPR menjaminkannya, sementara yang lain menentang penafsiran ini tidak diragukan lagi bahwa para pembuat kebijakan mulai mempertimbangkan secara serius dampak sosial ML, dan bahkan mungkin mengambil langkah-langkah awal untuk meregulasi industri yang bergantung pada teknologi tersebut. Namun, setiap upaya untuk melakukannya akan memerlukan kemampuan teknis untuk mengaudit algoritma agar dapat menguji secara ketat apakah algoritma tersebut mendiskriminasi berdasarkan atribut yang dilindungi seperti ras dan jenis kelamin.

Konteks Justifikasi (Algoritmik)

Beralih dari pertimbangan etika ke epistemologis, banyak peneliti iML menekankan bahwa perangkat mereka dapat membantu men-debug algoritma yang tidak berkinerja dengan baik. Masalah klasik dalam konteks ini adalah *overfitting*, yang terjadi ketika model memprediksi dengan baik pada data pelatihan tetapi gagal pada data uji. Hal ini terjadi, misalnya, dengan pengklasifikasi gambar terbaru yang dirancang untuk membedakan hewan ternak.

Model tersebut mencapai akurasi 100% pada evaluasi dalam sampel tetapi salah memberi label pada semua kuda dalam set uji berikutnya. Pemeriksaan lebih dekat mengungkapkan bahwa data pelatihan memiliki tanda air kecil pada semua dan hanya gambar kuda. Algoritma telah belajar untuk mengasosiasikan label "kuda" bukan dengan fitur kuda, seperti yang mungkin diharapkan, tetapi hanya dengan ciri khas yang tidak informatif ini.

Fenomena *overfitting*, yang dikenal luas dan dikhawatirkan di komunitas pembelajaran mesin (ML), mungkin sudah tidak asing lagi bagi para epistemolog sebagai semacam kasus

Gettier algoritmik. Jika sebuah pengklasifikasi gambar berkinerja tinggi memberi label "kuda" pada sebuah foto kuda, maka kita memiliki keyakinan sejati yang beralasan bahwa gambar tersebut memang menggambarkan seekor kuda.

Namun, ketika penentuan tersebut dibuat berdasarkan tanda air, ada sesuatu yang kurang tepat. Jalan kita menuju fakta tersebut entah bagaimana bengkok, kebetulan. Model tersebut benar *karena alasan yang salah*. Setiap penilaian sejati yang dibuat atas dasar ini hanyalah kasus keberuntungan epistemik, seperti ketika kita mengetahui waktu dengan benar dengan melihat jam yang berhenti tepat 24 jam sebelumnya. Upaya untuk menghindari masalah seperti ini biasanya melibatkan upaya untuk memastikan bahwa agen dan proposisi berada dalam hubungan yang tepat, yaitu bahwa suatu metode yang andal menghubungkan yang mengetahui dan pengetahuan.

Reliabilitas proses secara terkenal dipelopori oleh Goldman (1979), yang bisa dibilang memimpin garda depan dari apa yang disebut Williams sebagai "revolusi reliabilis" (2016) dalam epistemologi berbahasa Inggris. Floridi (2004) menunjukkan ketidakterpecahan logis dari masalah Gettier (dalam konteks non-statistik), sementara teori jaringan akunnya (2012) secara efektif menetapkan solusi pragmatis dan reliabilis.

Kemajuan dalam iML merupakan jawaban statistik untuk tantangan reliabilis, yang memungkinkan para skeptis untuk menganalisis perilaku internal suatu model ketika mempertimbangkan prediksi tertentu. Ini adalah tujuan, misalnya, dari semua teknik aproksimasi linier lokal, termasuk algoritma iML populer seperti LIME dan SHAP, yang memberikan bobot pada variabel masukan sehingga pengguna dapat memverifikasi bahwa model tersebut tidak berfokus secara tidak tepat pada fitur yang tidak informatif seperti tanda air yang disebutkan sebelumnya. Metode ini akan diperiksa lebih dekat di Bagian 10.6.

Konteks Penemuan (Algoritmik)

Kami mempertimbangkan satu motivasi terakhir untuk iML: *penemuan*. Subjek ini sejauh ini relatif kurang mendapat perhatian dalam literatur. Namun, kami berpendapat bahwa hal ini sebenarnya dapat menjadi salah satu pencapaian terpenting dari keseluruhan proyek keterjelasan algoritmik, dan oleh karena itu patut mendapat perhatian khusus.

Misalkan kita merancang sebuah algoritma untuk memprediksi subtipe dari beberapa penyakit yang kurang dipahami menggunakan data biomolekuler. Model ini sangat akurat. Model ini secara jelas mengklasifikasikan pasien ke dalam kelompok-kelompok berbeda dengan lintasan prognostik yang sangat berbeda. Prediksinya kuat dan andal, memberikan saran yang dapat ditindaklanjuti kepada dokter tentang pilihan pengobatan dan menyarankan jalur baru untuk penelitian di masa mendatang. Dalam hal ini, kami ingin metode iML tidak mengaudit kewajaran atau menguji overfitting, tetapi untuk mengungkap mekanisme yang mendasarinya. Algoritme ini jelas telah belajar untuk mengidentifikasi dan memanfaatkan beberapa sinyal halus yang sejauh ini sulit dideteksi manusia. Jika kita ingin mempelajari lebih lanjut tentang sistem target, teknik iML yang diterapkan pada model yang terspesifikasi dengan baik menawarkan cara yang relatif murah dan efektif untuk mengidentifikasi fitur-fitur utama dan menghasilkan hipotesis baru.

Kasus ini tidak sepenuhnya hipotetis. Gelombang penelitian di awal tahun 2000-an menunjukkan hubungan antara tanda-tanda transkriptomik dan luaran klinis untuk pasien kanker payudara. Studi-studi tersebut menggunakan sejumlah teknik statistik canggih, termasuk pengelompokan tanpa pengawasan dan analisis kelangsungan hidup. Para peneliti menemukan, antara lain, hubungan yang kuat antara mutasi BRCA1 dan kanker payudara tipe basal, suatu bentuk penyakit yang sangat agresif. Analisis genomik tetap menjadi salah satu bidang penelitian yang paling aktif dan menjanjikan dalam ilmu pengetahuan alam, dan berbagai subbidang baru dalam pembelajaran mesin telah muncul untuk mengatasi tantangan unik yang dihadirkan oleh kumpulan data berdimensi tinggi ini. Strategi pembelajaran mesin yang sukses akan sangat krusial untuk mewujudkan potensi ilmu pengetahuan berthroughput tinggi.

10.3 LATAR BELAKANG FORMAL

Pada bagian ini, kami memperkenalkan konsep dan notasi yang akan digunakan di sepanjang sisa makalah ini. Secara khusus, kami meninjau formalisme dasar pembelajaran terbimbing, intervensionisme kausal, dan teori keputusan.

Pembelajaran Terbimbing

Tujuan dalam pembelajaran terbimbing adalah untuk memperkirakan suatu fungsi yang memetakan sekumpulan variabel prediktor ke beberapa hasil yang diinginkan. Untuk membahas algoritma pembelajaran dengan kejelasan formal, kita harus merujuk pada nilai, variabel, vektor, dan matriks. Kita menyatakan nilai skalar menggunakan huruf kecil miring, misalnya x . Sebaliknya, variabel diidentifikasi dengan huruf besar miring, misalnya $\mathbf{X}$. Matriks, yang terdiri dari baris observasi dan kolom variabel, dilambangkan dengan huruf besar tebal, misalnya $\mathbf{X}$. Kita terkadang mengindeks nilai dan variabel menggunakan notasi matriks, sehingga elemen ke- i dari variabel $\mathbf{X}$ adalah x_i dan variabel ke- j dari matriks $\mathbf{X}$ adalah $\mathbf{X}_j$. Skalar x_{ij} merujuk pada elemen ke- i dari variabel ke- j di $\mathbf{X}$. Ketika merujuk pada vektor baris, seperti koordinat yang mengidentifikasi observasi ke- i di $\mathbf{X}$, kami menggunakan notasi huruf kecil, tebal, dan miring, misalnya $\mathbf{x}_i$.

Setiap observasi dalam set data pelatihan terdiri dari pasangan $z_i = (\mathbf{x}_i, y_i)$, di mana $\mathbf{x}_i$ menyatakan suatu titik dalam ruang berdimensi- d , $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$, dan y_i merepresentasikan hasil yang sesuai. Kami berasumsi bahwa sampel terdistribusi secara independen dan identik menurut suatu distribusi probabilitas gabungan yang tetap tetapi tidak diketahui $\mathbb{P}(\mathbf{Z}) = \mathbb{P}(\mathbf{X}, Y)$. Dengan menggunakan n observasi, suatu algoritma memetakan set data ke suatu fungsi, $a: \mathbf{Z} \rightarrow f$; fungsi tersebut selanjutnya memetakan fitur ke hasil, $f: \mathbf{X} \rightarrow Y$. Kami mempertimbangkan kedua kasus di mana Y bersifat kategoris (yang dalam hal ini f merupakan pengklasifikasi) dan di mana Y bersifat kontinu (yang dalam hal ini f merupakan regresor). Kami tidak membuat asumsi tambahan tentang struktur atau properti f .

Model f dinilai berdasarkan kemampuannya untuk melakukan *generalisasi*, yaitu memprediksi hasil secara akurat pada data uji yang diambil dari $\mathbb{P}(\mathbf{Z})$ tetapi tidak termasuk dalam set data pelatihan. Untuk sampel uji $\mathbf{x}_i$ tertentu, kami menghitung hasil prediksi $f(\mathbf{x}_i)$

$= \hat{y}_i$ dan mengamati hasil sebenarnya y_i . Simbol topi menunjukkan bahwa nilai tersebut telah diestimasi.

Kinerja suatu model diukur dengan fungsi kerugian L , yang mengkuantifikasi jarak antara Y dan $\hat{Y}$ pada serangkaian kasus uji. Nilai ekspektasi fungsi kerugian ini terhadap $\mathbb{P}(\mathbf{Z})$ untuk model f tertentu disebut risiko:

$$R(f, \mathbf{Z}) = \mathbb{E}_{\mathbb{P}(\mathbf{Z})}[L(f, \mathbf{Z})] \quad (10.1)$$

Kami memperkirakan parameter populasi ini dengan risiko empiris pada sekumpulan n sampel:

$$R_{emp}(f, \mathbf{Z}) = \frac{1}{n} \sum_i L(f, z_i) \quad (10.2)$$

Suatu algoritma pembelajaran dikatakan *konsisten* jika risiko empiris konvergen dengan risiko sebenarnya ketika $n \rightarrow \infty$. Hasil fundamental dari teori pembelajaran statistik menyatakan bahwa suatu algoritma konsisten jika dan hanya jika ruang fungsi yang dapat dipelajarinya berdimensi VC hingga. Parameter terakhir ini merupakan ukuran kapasitas yang didefinisikan sebagai kardinalitas dari himpunan titik terbesar yang dapat dipecahkan oleh algoritma. Kriteria dimensi VC hingga akan penting untuk menentukan kondisi konvergensi untuk permainan. Beberapa filsuf berpendapat bahwa pembelajaran statistik memberikan fondasi yang kuat untuk semua penalaran induktif. Meskipun kami sependapat dengan posisi ini, tidak satu pun analisis selanjutnya bergantung pada tesis ini.

Intervensionisme Kausal

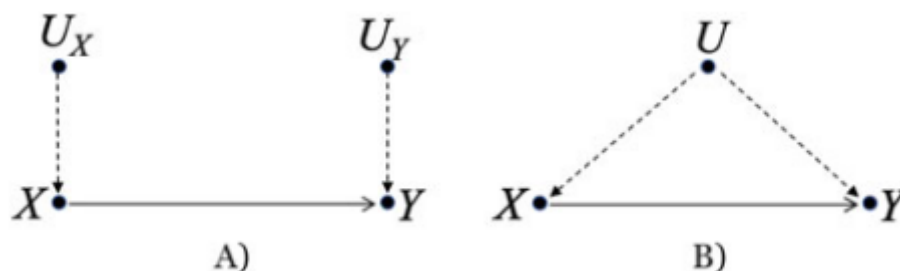
Para filsuf sering membedakan antara penjelasan kausal (untuk peristiwa alam) dan alasan pribadi (untuk keputusan manusia). Berbicara tentang "keputusan" algoritmik juga umum meskipun sangat menyesatkan. Oleh karena itu, kita mungkin tergoda untuk mencari *alasan*, alih-alih *penyebab*, untuk prediksi algoritmik, dengan alasan bahwa prediksi tersebut lebih menyerupai keputusan daripada peristiwa. Kami berpendapat bahwa hal ini keliru dalam beberapa hal.

Pertama, pembicaraan tentang "keputusan" algoritmik adalah kiasan antropomorfik yang memberikan model statistik tingkat otonomi yang secara berbahaya meremehkan peran agensi manusia yang sebenarnya dalam sistem sosioteknis. Kedua, kita mungkin ingin menjelaskan tidak hanya label teratas yang dipilih oleh suatu pengklasifikasi yang disebut "keputusan" tetapi juga distribusi probabilitas lengkap atas label-label yang mungkin. Dalam konteks regresi, kita mungkin ingin menjelaskan interval prediksi selain estimasi titik belaka. Akhirnya, terdapat alasan pragmatis yang baik untuk mengambil pendekatan kausal terhadap permasalahan ini. Sebagaimana kami bahas di Bagian 10.4, relatif mudah dan sangat informatif untuk mensimulasikan efek intervensi kausal pada model pembelajaran terbimbing, asalkan aksesnya memadai.

Oleh karena itu, pendekatan kami dibangun berdasarkan kerangka kerja intervensionis kausal yang awalnya diformalkan oleh Pearl (2000) dan Spirtes dkk. (2000), dan kemudian

dikaji secara lebih filosofis oleh Woodward (2003, 2008, 2010, 2015). Penjelasan singkat tentang teori ini adalah sebagai berikut. X merupakan penyebab Y dalam model struktural M tertentu jika dan hanya jika beberapa intervensi hipotetis pada X (dan tidak ada variabel lain) akan mengakibatkan perubahan pada Y atau distribusi probabilitas Y . Perhitungan ini minimal dalam artian tidak memberikan batasan pada M dan tidak memaksakan ambang batas efikasi kausal pada X atau Y . Konsep intervensi dijaga semaksimal mungkin agar memungkinkan setiap kemungkinan perubahan pada X , asalkan tidak mengubah nilai variabel lain dalam M kecuali yang merupakan turunan kausal dari X .

Dengan asumsi umum tertentu, kalkulus *do* Pearl menyediakan seperangkat alat formal yang lengkap untuk penalaran tentang intervensi kausal. Elemen kunci dari sistem notasi Pearl adalah operator *do*, yang memungkinkan kita untuk menyatakan, misalnya, probabilitas Y , dengan syarat intervensi yang menetapkan variabel X ke nilai x , dengan rumus ringkas $\mathbb{P}(Y|do(X = x))$. Model kausal struktural M adalah tupel $\langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ yang terdiri dari variabel eksogen $\mathbf{U}$, variabel endogen $\mathbf{V}$, dan sekumpulan fungsi $\mathbf{F}$ yang memetakan anteseden kausal setiap V_j ke nilai teramatinya. M dapat digambarkan secara visual sebagai graf dengan simpul yang bersesuaian dengan variabel dan sisi berarah yang menunjukkan hubungan kausal antar fitur endogen (lihat Gambar 10.1). Kami membatasi perhatian kami di sini pada graf asiklik berarah (DAG), yang menjadi fokus sebagian besar penelitian dalam intervensionisme kausal.



Gambar 10.1 Dua contoh model kausal sederhana. (a) Grafik Markovian. Dua variabel eksogen, U_X dan U_Y , memiliki efek kausal yang tidak teramati pada dua variabel endogen, masing-masing X dan Y . (b) Grafik semi-Markovian. Satu variabel eksogen, U , memiliki efek pengganggu yang tidak teramati pada dua variabel endogen, X dan Y .

Jika model M tidak mengandung pengganggu eksogen, maka M dikatakan *Markovian*. Dalam hal ini, faktorisasi distribusi gabungan grafik bersifat langsung dan efek kausal dapat dihitung langsung dari data. Namun, ketika satu atau lebih variabel yang tidak teramati memiliki efek pengganggu pada dua atau lebih variabel teramati, seperti pada Gambar 10.1b, maka kita mengatakan bahwa M adalah *semi-Markovian*, dan metode yang lebih rumit diperlukan untuk memperkirakan efek kausal. Secara spesifik, beberapa jenis penyesuaian harus dilakukan dengan mengondisikan pada serangkaian kovariat yang sesuai. Meskipun beberapa formulasi yang tumpang tindih telah diusulkan untuk penyesuaian tersebut, kami mengikuti Tian dan

Pearl (2002), yang memberikan seperangkat kondisi identifikasi kausal yang terbukti valid dan lengkap untuk model semi-Markovian.

Kriteria mereka adalah sebagai berikut. Efek kausal variabel endogen V_j pada semua kovariat teramati V_{-j} dapat diidentifikasi jika dan hanya jika tidak terdapat urutan tepi pengganggu yang berurutan antara V_j dan penerus langsung V_j s dalam grafik. Kondisi yang lebih lemah sudah memadai ketika kita berfokus pada subset proper $S \subset V$. Dalam hal ini, $\mathbb{P}(S | do(V_j = v_{ij}))$ dapat diidentifikasi selama tidak ada urutan sisi pengganggu yang berurutan antara V_j dan anak-anak V_j dalam subgraf yang terdiri dari leluhur S .

Kami berasumsi bahwa tujuan dalam sebagian besar aplikasi iML adalah untuk memberikan penjelasan kausal untuk satu atau lebih keluaran algoritmik. Oleh karena itu, identifikasi merupakan perhatian utama, dan komponen kunci lainnya untuk mendefinisikan kondisi konvergensi. Jika berhasil, penjelasan kausal untuk beberapa prediksi algoritmik akan secara akurat menjawab serangkaian pertanyaan yang disebut Woodward sebagai "bagaimana-jika-segalanya-berbeda" (selanjutnya disebut pertanyaan-w). Misalnya, kita mungkin ingin tahu ciri-ciri apa pada seseorang yang menyebabkan pengajuan pinjamannya ditolak. Bagaimana jika dia lebih kaya? Atau lebih tua? Akankah pemohon hipotetis yang identik dengan pemohon asli, kecuali pada sumbu kekayaan atau usia, lebih beruntung? Beberapa penulis dalam literatur iML secara eksplisit mendukung strategi kontrafaktual semacam itu. Kami meninjau kembali metode-metode ini di Bagian 10.6.

Tabel 10.1 Matriks utilitas bagi John ketika memutuskan apakah akan membawa payungnya atau tidak

	c_1 Hujan	c_2 Tidak ada hujan
a_1 : Payung	1	-1
a_2 : Tidak ada payung	-2	0

Teori Keputusan

Teori keputusan menyediakan perangkat formal untuk penalaran tentang pilihan dalam ketidakpastian. Perangkat ini akan terbukti berguna ketika mencoba mengukur relevansi. Kami mengasumsikan pengaturan tipikal, di mana seorang individu mempertimbangkan serangkaian tindakan berhingga A dan serangkaian hasil berhingga C . Menurut teori utilitas harapan, preferensi rasional seorang agen dapat dinyatakan sebagai fungsi utilitas u yang memetakan produk Cartesien dari A dan C ke bilangan riil, $u: A \times C \rightarrow \mathbb{R}$. Misalnya, Jones mungkin ragu apakah akan mengemas payungnya hari ini. Ia bisa saja melakukannya (a_1), tetapi itu akan menambah beban dan berat tasnya; atau ia bisa meninggalkannya di rumah (a_2) dan berisiko basah. Matriks utilitas yang dihasilkan digambarkan dalam Tabel 10.1.

Pilihan rasional Jones tidak hanya bergantung pada fungsi utilitas u -nya, tetapi juga pada keyakinannya tentang apakah akan hujan atau tidak. Hal ini secara formal dinyatakan dengan distribusi probabilitas (subjektif) atas C , $\mathbb{P}(C)$. Kami menghitung utilitas yang diharapkan dari setiap tindakan dengan mengambil rata-rata tertimbang atas hasil:

$$\mathbb{E}_{\mathbb{P}(C)}[u(a_i|C)E] = \sum_j \mathbb{P}(c_j|E)u(a_i, c_j) \quad (10.3)$$

di mana himpunan bukti E kosong (dalam hal ini Persamaan 11.3 menunjukkan ekspektasi sebelumnya) atau berisi beberapa bukti relevan (dalam hal ini Persamaan 10.3 menunjukkan ekspektasi posterior). Probabilitas posterior dihitung sesuai dengan teorema Bayes:

$$\mathbb{P}(c_i|E) = \frac{\mathbb{P}(E|c_i)\mathbb{P}(c_i)}{\mathbb{P}(E)} \quad (10.4)$$

yang mengikuti langsung aksioma Kolmogorov untuk kalkulus probabilitas (Kolmogorov 1950). Dengan menyelesaikan Persamaan 11.3 untuk setiap elemen di A, kami mengidentifikasi setidaknya satu tindakan yang memaksimalkan utilitas:

$$a^* = \underset{a_i \in A}{\operatorname{argmax}} \mathbb{E}_{\mathbb{P}(C)}[u(a_i, C)|E] \quad (10.5)$$

Agen epistemik ideal selalu memilih (salah satu) tindakan optimal a^* dari serangkaian alternatif.

Penting untuk dicatat bagaimana keyakinan agen rasional berinteraksi dengan utilitasnya untuk memandu keputusan. Jika Jones sangat tidak yakin apakah akan hujan atau tidak, maka ia menetapkan probabilitas yang sama untuk kedua hasil, menghasilkan utilitas yang diharapkan sebesar

$$\mathbb{E}_{\mathbb{P}(C)}[u(a_1, C)] = 0.5(1) + 0.5(-1) = 0$$

Dan

$$\mathbb{E}_{\mathbb{P}(C)}[u(a_2, C)] = 0.5(-2) + 0.5(0) = -1,$$

masing-masing. Dalam kasus ini, Jones seharusnya membawa payungnya. Namun, katakanlah ia mendapatkan beberapa informasi baru E yang mengubah keyakinannya. Mungkin ia melihat laporan cuaca yang menunjukkan kemungkinan hujan hanya 10%. Maka ia akan memiliki utilitas yang diharapkan berikut ini:

$$\mathbb{E}_{\mathbb{P}(C)}[u(a_1, C)|E] = 0.1(1) + 0.9(-1) = -0.8$$

$$\mathbb{E}_{\mathbb{P}(C)}[u(a_2, C)|E] = 0.1(-2) + 0.9(0) = -0.2$$

Dalam kasus ini, meninggalkan payung di rumah adalah pilihan optimal bagi Jones.

Tentu saja, manusia bisa sangat irasional. Eksperimen dalam psikologi dan ekonomi perilaku telah berulang kali menunjukkan bahwa manusia mengandalkan heuristik dan bias kognitif, alih-alih secara konsisten menerapkan aksioma teori keputusan atau kalkulus probabilitas (Kahneman 2011). Dengan demikian, konsep dan prinsip yang kami uraikan di sini terutama bersifat normatif. Konsep dan prinsip tersebut menetapkan arah perilaku yang optimal, semacam ideal regulatif Kant ketika utilitas dan probabilitas tepat, dan distribusi posterior dihitung dengan tepat. Untuk tujuan praktis iML, nilai-nilai ini dapat diestimasi melalui sistem hibrida di mana perangkat lunak membantu individu yang ingin tahu dengan rasionalitas terbatas. Kami meninjau kembali isu-isu ini di Bagian 11.7.1.

10.4 CAKUPAN

Algoritma pembelajaran terbimbing menyediakan beberapa affordance unik yang membedakan iML dari tugas penjelasan yang lebih umum. Hal ini karena target dalam iML bukanlah fenomena alam atau sosial yang dirancang untuk diprediksi oleh algoritma, melainkan *algoritma itu sendiri*. Dengan kata lain, kita tidak tertarik pada distribusi gabungan yang mendasari $\mathbb{P}(\mathbf{Z}) = \mathbb{P}(\mathbf{X}, \mathbf{Y})$, tetapi pada distribusi gabungan yang diestimasi $\mathbb{P}(\mathbf{Z}_f) = \mathbb{P}(\mathbf{X}, \hat{\mathbf{Y}})$. Perbedaan ini sangat penting.

Stevens (2013) membedakan tiga cara pemahaman: *bahwa*, *mengapa*, dan *dengan*. Memahami *bahwa* suatu proposisi p benar berarti menyadari bahwa p . Memahami *mengapa* p benar membutuhkan penjelasan kausal untuk p . Namun, jenis pemahaman ketiga Stevens hanya berlaku untuk teori atau model. Memahami dengan model berarti mengetahui cara menerapkannya untuk memprediksi atau menjelaskan fenomena nyata atau potensial.

Misalnya, seorang fisikawan yang menggunakan mekanika Newton untuk menjelaskan gerak bola biliar dengan demikian menunjukkan kemampuannya untuk memahami dengan teori tersebut. Karena model ini secara tegas salah, tidaklah tepat untuk mengatakan bahwa penjelasannya memberikan pemahaman yang benar tentang mengapa bola biliar bergerak seperti itu. (Tentu saja, ia dapat dimaafkan karena tidak menjelaskan detail relativistik tensor metrik dan kelengkungan ruangwaktu kepada rekan-rekan biliarnya dalam kasus ini.)

Namun, fisikawan kita telah memahami sesuatu dengan jelas yaitu teori Newton itu sendiri meskipun penjelasan klasik yang ia tawarkan tidak akurat atau tidak lengkap. Demikian pula, tujuan dalam iML adalah untuk membantu agen epistemik memahami dengan model target f , terlepas dari realitas apa pun yang ingin ditangkap oleh f . Situasinya sedikit lebih rumit dalam kasus penemuan. Strateginya di sini adalah menggunakan pemahaman dengan sebagai jalur tidak langsung untuk memahami *mengapa*, dengan asumsi bahwa jika model f berkinerja baik, maka model tersebut mungkin telah mempelajari beberapa informasi berharga tentang sistem target. Meskipun beberapa model statistik cukup kompleks, sebagai sebuah kelas, model-model tersebut cenderung *lengkap*, *tepat*, dan *mudah* dipahami. Ketiga sifat ini menyederhanakan upaya untuk menjelaskan sistem yang kompleks.

Lengkap

Model f lengkap berkenaan dengan fitur masukan $\mathbf{X}$ dalam arti bahwa variabel eksogen tidak memiliki pengaruh apa pun terhadap hasil prediksi $\hat{\mathbf{Y}}$. Meskipun alam penuh dengan

faktor pengganggu yang tidak teramati yang dapat mempersulit atau melemahkan bahkan studi yang dirancang dengan baik, model yang dipasang merupakan sistem mandiri yang kebal terhadap variasi eksternal. Oleh karena itu, model tersebut menginstansiasi grafik Markovian, alih-alih grafik semi-Markovian. Hal ini berlaku bahkan jika dependensi antar prediktor tidak dimodelkan secara eksplisit, dalam hal ini kita dapat menggambarkan f sebagai DAG sederhana dengan tepi terarah dari setiap fitur $X_1, \dots, X_d$ hingga Y .

Selanjutnya, kami berasumsi bahwa agen yang dimaksud mengetahui variabel mana yang digunakan untuk melatih f . Hal ini mungkin tidak selalu terjadi dalam praktik, dan tanpa pengetahuan tersebut, akan jauh lebih sulit untuk menjelaskan prediksi algoritmik. Apa pun status epistemik agen yang bertanya, model yang mendasarinya sendiri tetap lengkap. Masalah muncul ketika variabel endogen berfungsi sebagai proksi untuk variabel eksogen. Misalnya, sebuah model mungkin tidak secara eksplisit menyertakan atribut yang dilindungi seperti ras, tetapi malah menggunakan kovariat yang tampaknya tidak berbahaya seperti kode pos, yang seringkali merupakan prediktor kuat ras. Dalam hal ini, intervensi yang mengubah ras subjek tidak akan berdampak pada prediksi model f kecuali kita mengambil langkah tambahan dengan menanamkan f dalam struktur kausal M yang lebih besar yang mencakup tepi terarah dari ras ke kode pos.

Presisi

Model f presisi dalam arti selalu menghasilkan keluaran yang sama untuk setiap set masukan tertentu. Prosedur eksperimen tertentu dapat menghasilkan hasil yang berbeda pada percobaan berulang karena derau yang tak tereduksi, sementara model yang terfitting tidak memiliki variabilitas internal seperti itu. Beberapa pendekatan berbasis simulasi, seperti metode rantai Markov Monte Carlo yang banyak digunakan dalam analisis data Bayesian, merupakan pengecualian penting terhadap aturan ini.

Model-model ini membuat prediksi melalui pengambilan sampel acak, sebuah proses stokastik yang keluaran akhirnya berupa distribusi posterior, bukan estimasi titik. Namun, jika model telah konvergen, maka prediksi ini masih presisi hingga batasnya. Seiring bertambahnya jumlah penarikan dari posterior, statistik yang diinginkan (misalnya, modus posterior atau rerata) stabil ke nilai akhirnya. Varians Monte Carlo dari suatu parameter dapat dibatasi sebagai fungsi ukuran sampel menggunakan pertidaksamaan konsentrasi yang umum diketahui.

Woodward (2003, 2010) menekankan peran "stabilitas" dalam generalisasi kausal, sebuah konsep yang menyerupai apa yang kita sebut presisi. Perbedaanannya adalah stabilitas dalam pengertian Woodward hanya dapat diterapkan pada subset tepi yang tepat (biasanya hanya satu tepi) dalam grafik kausal. Generalisasi bahwa "variabel X menyebabkan variabel Y " *stabil* sejauh ia bertahan di berbagai kondisi latar belakang, yaitu keadaan alternatif model M . Presisi dalam pengertian kami membutuhkan kelengkapan, karena hanya berlaku untuk hubungan kausal antara himpunan semua prediktor $\mathbf{X}$ dan hasil Y , yang sepenuhnya deterministik pada tingkat token.

Akan Segera Hadir

Model f akan segera hadir dalam arti bahwa ia akan selalu memberikan keluaran untuk setiap masukan yang terbentuk dengan baik. Selain itu, biasanya cukup cepat dan murah untuk mengkueri algoritma dengan cara ini. Meskipun eksperimen dalam ilmu alam atau ilmu sosial seringkali memakan waktu, tidak meyakinkan, mahal, atau bahkan berbahaya, menjawab pertanyaan-w dalam konteks pembelajaran terawasi relatif mudah. Pada prinsipnya, seorang analis bahkan dapat menciptakan kembali distribusi gabungan lengkap $\mathbb{P}(X, Y)$ hanya dengan memenuhi ruang fitur dengan pertanyaan-w. Tentu saja, strategi ini tidak layak secara komputasi dengan prediktor kontinu dan/atau matriks desain dengan dimensionalitas moderat sekalipun.

Algoritma pembelajaran terawasi mungkin kurang efektif ketika dilindungi oleh hukum kekayaan intelektual (HKI), yang juga dapat mencegah peneliti mengakses daftar lengkap prediktor suatu model. Sebagai pengganti antarmuka pemrograman akses terbuka, beberapa peneliti iML menggunakan rekayasa balik algoritma dari set data pelatihan dengan nilai prediksi yang diketahui. Hal ini misalnya terjadi pada investigasi ProPublica yang terkenal terhadap algoritma COMPAS, sebuah model kepemilikan yang digunakan oleh pengadilan di beberapa negara bagian AS untuk memprediksi risiko residivisme kriminal. Studi-studi selanjutnya yang menggunakan set data yang sama mencapai kesimpulan yang berbeda terkait ketergantungan algoritma tersebut pada ras, yang menyoroti ketidakpastian inheren rekonstruksi model ketika algoritma target tidak tersedia. Berikut ini, kami berfokus pada kasus ideal di mana agen kami tidak menghadapi batasan kekayaan intelektual.

10.5 PERMAINAN PENJELASAN

Pada bagian ini, kami memperkenalkan kerangka kerja formal untuk iML. Proposal kami berbentuk permainan di mana seorang inkuisitor (sebut saja Alice) mencari penjelasan untuk prediksi algoritmik $f(x_i) = \hat{y}_i$. Perhatikan bahwa target kami (pada tahap ini) adalah penjelasan *lokal* atau *token*, alih-alih penjelasan *global* atau *tipe*. Dengan kata lain, Alice ingin tahu mengapa masukan tertentu ini menghasilkan keluaran tertentu, alih-alih tugas yang lebih umum untuk menciptakan kembali seluruh batas keputusan atau permukaan regresi f .

Sayangnya bagi Alice, f adalah kotak hitam. Namun, ia tidak sendirian. Ia dibantu oleh seorang kaki tangan yang setia (sebut saja Bob), yang melakukan segala daya upaya untuk membantu Alice memahami $\hat{y}_i$. Tujuan Bob adalah membawa Alice ke titik di mana ia dapat memprediksi keluaran f dengan tepat, setidaknya di sekitar x_i dan dalam margin kesalahan yang dapat ditoleransi. Dengan kata lain, ia ingin Alice mampu memberikan jawaban yang benar atas pertanyaan-pertanyaan w yang relevan tentang bagaimana f akan merespons titik data hipotetis di dekat x_i .

Kami membuat beberapa asumsi nontrivial tentang Alice dan Bob, beberapa di antaranya telah diramalkan sebelumnya. Secara spesifik:

- Alice adalah agen rasional. Preferensinya terhadap alternatif bersifat lengkap dan transitif, ia mengintegrasikan bukti baru melalui pembaruan Bayesian, dan ia berusaha sebaik mungkin untuk memaksimalkan utilitas yang diharapkan dengan tunduk pada batasan sumber daya kognitif/komputasinya.

- ☞ Bob adalah kaki tangan Alice. Ia memiliki data tentang fitur $V = X_1, \dots, X_d, \hat{Y}$ yang endogen terhadap f , serta sekumpulan variabel eksogen (yang mungkin kosong) $U = (X_d + 1, \dots, X_d + m)$ yang berpotensi menarik bagi Alice. Ia dapat mengkueri f dengan input apa pun yang terbentuk dengan baik dengan sedikit atau tanpa biaya. Kita dapat dengan mudah membayangkan permainan penjelasan yang lebih kompleks di mana beberapa atau semua asumsi ini dilonggarkan. Penelitian selanjutnya akan mengkaji alternatif-alternatif tersebut.

Tiga Desiderata

Menurut Woodward, tiga kriteria berikut secara individual diperlukan dan secara bersama-sama cukup untuk menjelaskan beberapa hasil yang menarik $Y = y_i$ yang diperoleh ketika $X = x_j$ dalam model struktural M tertentu:

- (i) Generalisasi yang dijelaskan oleh M akurat, atau setidaknya mendekati akurat, begitu pula observasi $Y = y_i$ dan $X = x_j$.
- (ii) Menurut M , $Y = y_i$ dalam intervensi yang menetapkan X ke x_j .
- (iii) Terdapat beberapa kemungkinan intervensi yang menetapkan X ke x_k (di mana $x_j \neq x_k$), dengan M secara tepat menggambarkan nilai y_l (di mana $y_i \neq y_l$) yang akan diasumsikan Y dalam intervensi tersebut.

Teori ini menimbulkan sejumlah komplikasi yang tidak sedikit dan berada di luar cakupan makalah ini. Kami mengadopsi kerangka kerja ini sebagai dasar analisis yang bermanfaat, karena cukup fleksibel untuk memungkinkan perluasan ke berbagai arah.

Akurasi

Penjelasan Woodward menempatkan premi yang beralasan pada akurasi penjelasan. Penjelasan apa pun yang gagal memenuhi kriteria (i)–(iii) tidak layak disebut demikian. Namun, teori ini tidak menceritakan keseluruhan cerita. Untuk memahami alasannya, pertimbangkan jaringan saraf konvolusional dalam f yang dilatih untuk mengklasifikasikan gambar. Model tersebut dengan tepat memprediksi bahwa x_i menggambarkan seekor kucing. Alice ingin tahu alasannya. Bob mencoba menjelaskan prediksi tersebut dengan menuliskan rumus lengkap untuk f . Jaringan saraf tersebut berisi beberapa ratus lapisan, masing-masing terdiri dari 1 juta parameter yang bersama-sama menggambarkan pemetaan nonlinier kompleks dari piksel ke label. Bob memeriksa kriteria Woodward dan mengamati bahwa modelnya M akurat, begitu pula nilai masukan dan keluarannya; bahwa M dengan tepat memprediksi keluaran berdasarkan masukan tersebut; dan intervensi pada foto asli yang mengganti kucing dengan anjing memang mengubah label yang diprediksi dari "kucing" menjadi "anjing".

Masalah terpecahkan? Belum sepenuhnya. Grafik kausal Bob, M , sama buramnya dengan model f yang mendasarinya. Faktanya, keduanya identik. Jadi, meskipun penjelasan ini mungkin sangat akurat, penjelasan ini terlalu rumit untuk berguna bagi Alice. Hasilnya tidak berbeda dengan peta dalam cerita pendek Borges yang terkenal (1946), di mana para kartografer kekaisaran mendambakan ketepatan sedemikian rupa sehingga mereka menggambar wilayah mereka pada skala 1:1.

Penjelasan kotak hitam semacam ini menciptakan semacam ruang Cina, di mana agen yang bertanya diharapkan melakukan komputasi algoritma secara manual untuk melacak jalur

dari masukan ke keluaran. Sebagaimana protagonis eksperimen pikiran Searle tidak memahami aksara Cina yang berhasil dimanipulasinya, Alice juga tidak memperoleh pengetahuan penjelasan tentang f dengan mewujudkan model itu sendiri.

Kecuali jika ia merasa nyaman menghitung produk tensor berdimensi tinggi secara langsung, Alice tidak dapat menggunakan M untuk membangun model mental sistem target f atau perilakunya di dekat x_i . Ia tidak dapat menjawab pertanyaan-pertanyaan w yang relevan tanpa berkonsultasi dengan program, yang hanya akan memberinya label baru yang sama sulitnya dijelaskan dengan label aslinya.

Kesederhanaan

Akurasi merupakan syarat yang diperlukan tetapi tidak cukup untuk keberhasilan penjelasan, terutama ketika sistem yang mendasarinya terlalu kompleks untuk dipahami sepenuhnya oleh agen yang bertanya. Dalam kasus ini, kita cenderung menilai *kesederhanaan* sebagai nilai inheren dari kandidat penjelasan. Poin ini bukanlah hal baru. Kesederhanaan telah dikutip sebagai tujuan utama teori ilmiah oleh hampir semua orang yang telah mempertimbangkan pertanyaan ini. Poin ini tidak luput dari perhatian para peneliti iML, yang biasanya menerapkan batasan kelangkaan pada solusi yang memungkinkan untuk memastikan jumlah parameter bukan nol yang dapat dikelola.

Tidak selalu jelas apa arti kesederhanaan penjelasan dalam konteks algoritmik. Salah satu kandidat yang masuk akal, yang dikemukakan oleh Popper (1959), didasarkan pada jumlah parameter bebas. Dalam teori pembelajaran statistik, proposal ini sebagian besar telah digantikan oleh ukuran kapasitas seperti dimensi VC yang telah disebutkan sebelumnya atau kompleksitas Rademacher. Parameter-parameter ini membantu membangun gagasan sintaksis tentang kesederhanaan, yang telah terbukti sangat bermanfaat dalam statistika. Namun, definisi-definisi tersebut mengaburkan aspek semantik kesederhanaan, yang mungkin lebih menarik bagi agen epistemik seperti Alice. Jenis kesederhanaan yang dibutuhkannya untuk memahami mengapa $f(x_i) = \hat{y}_i$ tidak hanya bergantung pada hubungan fungsional antar unit penjelasan, tetapi yang lebih penting lagi pada tingkat abstraksi penjelasan – yaitu, pilihan unit itu sendiri.

Alih-alih mempertimbangkan berbagai gagasan kesederhanaan yang saling bersaing yang berlimpah dalam literatur, kami memilih pendekatan relasional murni yang menganggap kesederhanaan setara dengan *keterpahaman bagi Alice*. Kami tidak yakin bahwa ada makna yang dapat diambil dari gagasan kesederhanaan yang absolut dan tidak bergantung pada pikiran.

Namun, meskipun ada, akan kurang bermanfaat bagi Alice jika kita bersikeras bahwa penjelasan g_1 lebih sederhana daripada g_2 berdasarkan definisi istilah yang kita sukai, meskipun ada bukti empiris bahwa ia lebih memahami implikasi penjelasan g_2 . Apa yang sederhana bagi beberapa agen mungkin rumit bagi yang lain, tergantung pada pengetahuan latar belakang dan faktor kontekstual. Di Bagian 11.5.2, kami mengoperasionalkan pengamatan ini dengan mengukur kesederhanaan dalam istilah yang secara eksplisit bersifat agen.

Relevansi

Beberapa orang mungkin menganggap akurasi dan kesederhanaan sudah cukup untuk penjelasan yang berhasil, dan dalam banyak kasus memang demikian. Namun, ada pengecualian penting untuk generalisasi ini. Perhatikan, misalnya, kasus berikut. Sebuah bank (yang buruk) mengeluarkan pinjaman hanya berdasarkan dua kriteria: pemohon harus berkulit putih atau kaya. Bank ini beroperasi di yurisdiksi yang hanya menganggap ras sebagai atribut yang dilindungi.

Seorang perempuan kulit hitam miskin bernama Alice ditolak pinjamannya dan meminta penjelasan. Bank memberi tahu bahwa pengajuannya ditolak karena kondisi keuangannya. Penjelasan ini akurat dan sederhana. Namun, penjelasan ini juga tidak jujur – karena akan sama akurat dan sederhananya jika dikatakan bahwa pinjamannya ditolak karena rasnya, sebuah hasil yang akan jauh lebih relevan bagi Alice dan regulator negara bagian. Mengingat kepentingan Alice, penjelasan yang terakhir lebih unggul daripada yang pertama, namun penjelasan bank secara efektif telah melampauinya.

Ini adalah pengamatan mendasar: di antara berbagai penjelasan yang akurat dan sederhana, beberapa akan lebih atau kurang relevan bagi agen yang bertanya. Alice memasuki permainan ini karena suatu alasan. Ada sesuatu yang menggantung di antara ketidakpastian. Mungkin dia adalah seorang pemohon pinjaman yang sedang mempertimbangkan apakah akan menuntut bank, atau seorang dokter yang sedang mempertimbangkan apakah akan mempercayai diagnosis yang tidak terduga.

Penjelasan yang berhasil tidak hanya harus akurat dan sederhana; penjelasan itu juga harus menginformasikan keputusannya tentang cara terbaik untuk melanjutkan. Jika tidak, kita akan menghadapi kasus *gerhana kontrafaktual*, di mana kepentingan agen dibayangi oleh fokus sempit pada fakta-fakta yang tidak relevan yang tidak berkontribusi pada pemahamannya atau membantu mengubah perilaku di masa mendatang. Masalah *gerhana kontrafaktual* merupakan isu serius dalam konteks apa pun di mana pelanggan atau pasien, misalnya, mungkin ingin menerima (atau mungkin menggunakan hak mereka untuk) mendapatkan penjelasan. Namun, kami tidak mengetahui adanya proposal dalam literatur iML yang secara eksplisit melindungi dari kemungkinan ini.

Algoritma 10.1: Masukan Permainan Penjelasan:

- *Lingkungan*: pembelajar terbimbing f , variabel endogen $\mathbf{V}$, data $D \sim \mathbb{P}(M)$ yang mungkin mencakup kovariat eksogen $\mathbf{U}$
- *Alice*: eksplanandum $f(x_i) = \hat{y}_i$, hasil kontrastif $f(x_i) \neq \tilde{y}_i$, tingkat abstraksi LoA, himpunan pilihan A , hipotesis kausal C , fungsi utilitas u , distribusi prior atas hipotesis kausal $\mathbb{P}(C)$, ruang fungsi H , fungsi kerugian L_H .
- *Bob*: himpunan B ruang fungsi unik G_b , fungsi rugi L_G , kernel k_G . Jika variabel eksogen relevan, maka ruang fungsi tambahan G' , fungsi rugi $L_{G'}$, kernel $k_{G'}$

untuk setiap putaran:

1. Bob membuat peta $\psi : Z_f \rightarrow Z_g$ dari ruang- f asli ke ruang- g penjelas yang dirancang untuk (a) menggeser distribusi masukan ke LoA yang diinginkan Alice dan (b) membantu memberikan bukti yang mendukung atau menentang setidaknya satu hipotesis dalam C . Sedangkan $Z_f = (\mathbf{X}, \hat{\mathbf{Y}})$, $Z_g = (\mathbf{X}', \mathbf{Y}')$.

jika $\mathbf{X}'$ mencakup variabel $\mathbf{U}$ yang eksogen terhadap f :

2. Bob melatih model $g': \mathbf{V} \rightarrow \mathbf{U}$, secara opsional disesuaikan menggunakan kernel k_G' , untuk meminimalkan kerugian L_G' atas ruang fungsi G' .
3. Bob membuat set data pelatihan dengan mengambil sampel titik $\mathbf{v}_s$ dari distribusi yang berpusat di $\mathbf{v}_i$ dan berulang kali menanyakan g' dengan w -pertanyaan dalam bentuk $\mathbb{E}_{\mathbb{P}(M)}[\mathbf{U} | do(\mathbf{V} = \mathbf{v}_s)] = ?$ Data yang dihasilkan dipetakan ke ruang- g melalui ψ .

akhir jika **untuk setiap ruang** fungsi G_b :

4. Bob membuat set data pelatihan dengan mengambil sampel titik $\mathbf{x}_s$ dari distribusi yang berpusat di $\mathbf{x}_i$ dan berulang kali menanyakan f dengan w -pertanyaan dalam bentuk $\mathbb{E}_{\mathbb{P}(Z_f)}[\mathbf{Y} | do(\mathbf{X} = \mathbf{x}_s)] = ?$ Data yang dihasilkan dipetakan ke ruang- g melalui ψ .
5. Bob melatih model $g: \mathbf{X}' \rightarrow \mathbf{Y}'$, yang secara opsional dicocokkan menggunakan kernel k_G , untuk meminimalkan kerugian L_G atas ruang fungsi G_b . Risiko empiris dihitung dalam ruang- f melalui pemetaan invers ψ^{-1} , yang secara opsional dibobot oleh k_G .
6. Alice membuat set data pelatihan dengan berulang kali mengkueri g dengan w -pertanyaan dalam bentuk $\mathbb{E}_{\mathbb{P}(Z_g)}[\mathbf{Y}' | do(\mathbf{X}' = \mathbf{x}'_{ij})] = ?$ Bob melaporkan hasil prediksi dan risiko empiris.
7. Alice melatih model $h: \mathbf{X}' \rightarrow \mathbf{Y}'$ untuk meminimalkan kerugian L_H atas ruang fungsi H . Risiko empiris secara opsional dibobot oleh k_G dan diestimasi dalam ruang- g .
8. Informasi yang dipelajari Alice dari dan tentang g dan h merupakan kumpulan bukti E , yang ia gunakan untuk memperbarui keyakinannya mengenai C .
9. Alice menghitung utilitas harapan posterior dari setiap tindakan dalam A , menghasilkan setidaknya satu pilihan optimal a^* .

Outputs: $R_{emp}(g, \mathbf{Z}_f), R_{emp}(h, \mathbf{Z}_g), \mathbb{E}_{\mathbb{P}(C)}[u(a^*, C) | E]$

akhir untuk

akhir untuk

Aturan Permainan

Setelah menekankan akurasi, kesederhanaan, dan relevansi, kami sekarang mengartikulasikan batasan formal yang memaksakan hal-hal yang diinginkan ini pada penjelasan dalam iML. Gambaran umum skematis dari permainan penjelasan ini disediakan dalam pseudocode.

Permainan ini memiliki banyak bagian yang bergerak, tetapi pada intinya prosesnya cukup mudah. Intinya, Bob berusaha sebaik mungkin untuk memberikan penjelasan yang akurat dalam istilah yang dapat dipahami Alice. Ia belajar dengan mengajukan pertanyaan-pertanyaan w hingga ia merasa cukup percaya diri untuk menjawab pertanyaan tersebut sendiri. Hasilnya dinilai berdasarkan tiga ukuran: akurasi (kesalahan model Bob), kesederhanaan (kesalahan model Alice), dan relevansi (utilitas yang diharapkan untuk Alice). Perhatikan bahwa semua penjelasan diindeks oleh peta ψ dan ruang fungsi penjelas G_b yang sesuai. Kami menghilangkan dependensi untuk kenyamanan notasi. Semua masukan dan langkah dibahas secara lebih rinci di bawah ini.

Masukan

Alice harus menentukan hasil kontrastif $f(x_i) \neq \tilde{y}_i \in Y$. Alternatif kontrafaktual ini dapat mewakili ekspektasi awal Alice atau respons yang diinginkan. Pertimbangkan, misalnya, kasus di mana f dilatih untuk membedakan angka tulisan tangan, sebuah masalah tolok ukur klasik dalam ML yang umumnya disebut MNIST, berdasarkan basis data gambar semacam itu yang paling terkenal.

Katakanlah f salah mengklasifikasikan x_i sebagai "7", padahal sebenarnya $y_i = "1"$. Alice ingin tahu bukan hanya mengapa model memprediksi "7", tetapi juga mengapa model *tidak* memprediksi "1". Menentukan alternatif y_i penting, karena memfokuskan perhatian Bob pada wilayah-wilayah yang relevan dari ruang fitur. Penjelasan seperti "Karena x_i tidak memiliki loop tertutup" dapat menjelaskan mengapa f tidak memprediksi "8" atau "9", tetapi penjelasan itu kurang bermanfaat bagi Alice, karena mengaburkan penjelasan yang relevan. Pentingnya penjelasan kontrastif disoroti oleh beberapa filsuf, dan baru-baru ini mulai mendapat perhatian dalam literatur iML juga.

Kami mengharuskan Alice menyatakan tingkat abstraksi (LoA) yang diinginkan. LoA menetapkan serangkaian variabel bertipe dan observabel yang digunakan untuk mendeskripsikan suatu sistem. Terinspirasi oleh literatur Metode Formal dalam ilmu komputer, pendekatan levelis telah diperluas untuk mengonseptualisasikan beragam masalah dalam filsafat informasi. LoA yang diinginkan Alice akan membantu Bob menetapkan unit penjelasan yang disukai, sebuah langkah krusial untuk memastikan keterpahaman bagi Alice. Dalam contoh MNIST, Alice kemungkinan besar tidak akan mencari penjelasan pada LoA piksel, tetapi mungkin puas dengan LoA yang lebih tinggi yang berkaitan dengan kurva dan tepi.

Pragmatisme menuntut Alice untuk memiliki gagasan mengapa ia memainkan permainan ini. Pilihan A , preferensi u , dan keyakinan $\mathbb{P}(C)$ -nya akan memandu Bob dalam upayanya untuk memberikan penjelasan yang memuaskan dan membatasi himpunan solusi yang mungkin. Contoh MNIST adalah kasus iML untuk validasi, di mana himpunan pilihan Alice dapat mencakup opsi untuk menerapkan atau tidak menerapkan model f . Derajat keyakinannya terhadap berbagai hipotesis kausal ditentukan oleh keahliannya dalam data dan model. Mungkin sudah diketahui secara luas bahwa algoritma kesulitan membedakan antara "7" dan "1" ketika angka pertama muncul tanpa garis horizontal yang melintasi digit tersebut. Biaya kesalahan semacam itu diperhitungkan dalam fungsi utilitasnya.

Bob, di sisi lain, memasuki permainan dengan tiga komponen kunci: (i) sekumpulan algoritma kandidat $B \geq 1$ untuk penjelasan; (ii) fungsi kerugian untuk melatih algoritma-algoritma ini; dan (iii) kernel yang sesuai. Pilihan populer untuk (i) mencakup model linear renggang dan daftar aturan. Fungsi kerugian tidak ditentukan, tetapi pilihan umum mencakup galat kuadrat rata-rata untuk regresi dan entropi silang untuk klasifikasi. Kernel menyeterel lokalitas penjelasan, memberi bobot pada observasi berdasarkan jaraknya dari input asli x_i , sebagaimana diukur dengan metrik yang sesuai.

Apakah kernel digunakan untuk melatih model g atau hanya mengevaluasi risiko empiris g , terserah Bob. Mengabaikan kernel sepenuhnya menghasilkan penjelasan global, tanpa penekanan khusus pada ketetanggaan x_i . Bob mungkin memerlukan algoritma, fungsi

kerugian, dan kernel tambahan untuk memperkirakan hubungan antara fitur endogen dan eksogen. Jika demikian, tidak ada persyaratan yang jelas bahwa model tersebut harus dapat dipahami oleh Alice atau Bob, selama mencapai galat prediktif minimal.

Memetakan Ruang

Mungkin langkah paling penting dalam keseluruhan permainan adalah pemetaan Bob $\psi : Z_f \rightarrow Z_g$. Dalam upaya memberikan penjelasan yang berhasil untuk Alice, Bob memproyeksikan distribusi masukan $\mathbb{P}(Z_f) = \mathbb{P}(X, \hat{Y})$ ke dalam ruang baru $\mathbb{P}(Z_g) = \mathbb{P}(X', Y')$. Perubahan variabel respons ditentukan oleh hasil kontrastif yang diinginkan Alice. Dalam contoh MNIST, Bob memetakan variabel 10-kelas asli $\hat{Y}$ ke variabel biner Y' yang menunjukkan apakah masukan diklasifikasikan sebagai "1" atau tidak. Isi X' dapat ditentukan secara iteratif dengan mempertimbangkan LoA yang diinginkan Alice dan himpunan hipotesis C . Hal ini seringkali menghasilkan reduksi ruang fitur. Misalnya, Bob dapat mengkasar sekumpulan gen menjadi kumpulan jalur biologis yang lebih kecil, atau mengubah piksel menjadi superpiksel.

Alternatifnya, Bob mungkin perlu memperluas fitur masukan agar mencakup variabel eksogen yang dihipotesiskan relevan dengan hasil. Dalam hal ini, ia akan membutuhkan data eksternal D yang diambil sampelnya dari ruang fitur yang diperluas $\mathbb{P}(M)$, yang dapat digunakan untuk melatih satu atau lebih model bantu guna memprediksi nilai kovariat tambahan di wilayah ruang- g yang tidak teramati. Misalnya, ketika suatu algoritma diduga mengodekan atribut yang dilindungi seperti ras melalui atribut yang tidak dilindungi seperti kode pos, Bob perlu memperkirakan ketergantungan menggunakan fungsi baru g' yang memprediksi atribut yang dilindungi berdasarkan atribut yang tidak dilindungi (bersama dengan variabel endogen relevan lainnya).

Perhatikan bahwa dalam upaya ini, Bob pada dasarnya kembali ke titik awal. Target M kemungkinan besar tidak lengkap, presisi, atau terbuka, sehingga tugasnya terbatas pada masalah yang lebih umum, yaitu memodelkan suatu sistem alam atau sosial yang kompleks dengan informasi terbatas. Hal ini mau tidak mau memunculkan sumber kesalahan baru yang akan berdampak negatif pada hasil selanjutnya. Bergantung pada sifat struktural grafik kausal yang mendasarinya, efek intervensi dalam ruang- g mungkin tidak dapat diidentifikasi secara unik.

Bagaimanapun, tujuan pada tahap ini adalah membuat fitur masukan cukup mudah dipahami oleh Alice sehingga dapat mengakomodasi kemungkinan pertanyaan w-nya dan menginformasikan keyakinannya tentang hipotesis kausal C . Metode tujuan umum untuk pembelajaran fitur kausal telah diusulkan, namun, para kritikus secara meyakinkan berpendapat bahwa prosedur tersebut tidak dapat diimplementasikan secara independen dari konteks. Beberapa bidang penelitian, seperti bioinformatika dan visi komputer, memiliki konvensi yang mapan tentang cara mengkasar ruang fitur berdimensi tinggi.

Domain lain mungkin terbukti lebih menantang. Aksesibilitas terhadap data eksternal tentang variabel eksogen yang diminati juga akan bervariasi dari kasus ke kasus. Bahkan ketika kumpulan data tersebut tersedia secara luas, tidak ada jaminan bahwa hubungan fungsional yang dicari dapat diperkirakan dengan akurasi atau presisi yang tinggi. Seperti dalam konteks

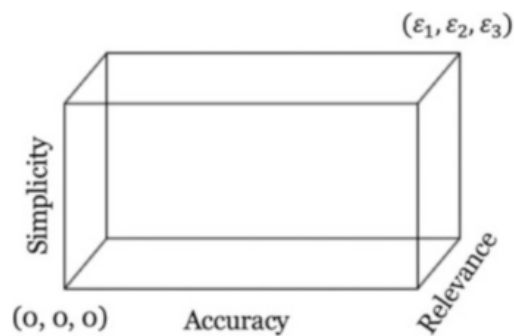
penjelasan lainnya, Alice dan Bob harus melakukan yang terbaik yang mereka bisa dengan sumber daya dan pengetahuan yang tersedia.

Membangun Model, Menskoling Penjelasan

Setelah ψ ditetapkan, langkah selanjutnya dalam permainan penjelasan adalah masalah pembelajaran yang diawasi secara efektif. Hal ini memberikan Alice dan Bob berbagai macam algoritma yang telah dipelajari dengan baik dan mengimpor jaminan statistik yang sesuai. Bob membuat set data pelatihan $\mathbf{Z}_g = (\mathbf{X}', Y')$ dan menyesuaikan model g dari ruang fungsi penjelas G_b . Alice mengeksplorasi ruang- g dengan mengajukan sejumlah pertanyaan w yang mengajukan intervensi yang relevan. Misalnya, ia mungkin ingin tahu apakah keberadaan garis horizontal di tengah suatu angka menentukan apakah f memprediksi angka "7". Jika demikian, maka ini akan menjadi hipotesis dalam C dan kita akan menemukan variabel yang sesuai dalam $\mathbf{X}'$. Karena kita membiarkan kemungkinan bahwa model target f dan/atau penjelasan Bob g dapat melibatkan persamaan struktural implisit atau eksplisit, kita menggunakan kalkulus do untuk memformalkan intervensi tersebut.

Bob dan Alice dapat memilih kombinasi fungsi kerugian dan algoritma apa pun yang paling masuk akal untuk tugas penjelasan yang diberikan. Kesalahan g diukur dengan $R_{emp}(g, \mathbf{Z}_f)$; kompleksitas g diukur dengan $R_{emp}(h, \mathbf{Z}_g)$. Kita katakan bahwa g akurat ε_1 jika $R_{emp}(g, \mathbf{Z}_f) \leq \varepsilon_1$ dan sederhana ε_2 jika $R_{emp}(h, \mathbf{Z}_g) \leq \varepsilon_2$. Isi dan kinerja g dan h merupakan kumpulan bukti E , yang digunakan Alice untuk memperbarui keyakinannya tentang hipotesis kausal C . Relevansi diukur dengan utilitas yang diharapkan posterior dari tindakan memaksimalkan utilitas, $\mathbb{E}_{\mathbb{P}(C)}[u(a^*, C)|E]$. (Untuk konsistensi dengan desiderata sebelumnya, kita sebenarnya mengukur ketidakrelevanan dengan mengalikan relevansi dengan -1.) Penjelasan Bob relevan ε_3 bagi Alice jika $-\mathbb{E}_{\mathbb{P}(C)}[u(a^*, C)|E] \leq \varepsilon_3$.

Kita sekarang dapat menempatkan penjelasan yang dihasilkan oleh permainan ini dalam ruang tiga dimensi, dengan sumbu yang sesuai dengan akurasi, kesederhanaan, dan relevansi. Sebuah penjelasan dianggap *memuaskan* jika tidak melebihi nilai ε_1 , ε_2 , dan ε_3 yang telah dipilih sebelumnya. Parameter-parameter ini dapat diinterpretasikan sebagai kendala anggaran pada Alice dan Bob. Seberapa besar ketidakakuratan, kompleksitas, dan ketidakrelevanan yang dapat mereka tanggung? Kita memberikan bobot yang sama untuk ketiga kriteria di sini, tetapi biaya relatif dapat dengan mudah dikuantifikasi melalui skema pembobotan diferensial. Bersama-sama, titik-titik ini mendefinisikan titik ekstrem dari sebuah kuboid, yang diagonal di seberangnya adalah titik asal (lihat Gambar 10.2). Titik mana pun yang berada dalam kuboid ini memuaskan $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$.



Gambar 10.2 Ruang penjelasan yang memuaskan dibatasi oleh batas atas kesalahan (ε_1), kompleksitas (ε_2), dan ketidakrelevanan (ε_3) dari penjelasan yang bersedia diterima Alice.

Konsistensi dan Konvergensi

Perangkat formal pembelajaran statistik, intervensionisme kausal, dan teori keputusan menyediakan semua unsur yang kita butuhkan untuk menyatakan kondisi perlu dan cukup bagi konvergensi ke permukaan penjelasan optimal bersyarat dalam waktu polinomial. Kita mendefinisikan optimalitas dalam istilah batas Pareto. Satu penjelasan mendominasi Pareto lainnya jika dan hanya jika penjelasan tersebut lebih baik secara ketat di setidaknya satu sumbu dan tidak lebih buruk di sumbu lainnya. Jika Alice dan Bob tidak dapat meningkatkan akurasi, kesederhanaan, atau relevansi suatu penjelasan tanpa mengalami kerugian di dimensi lain, maka mereka telah menemukan penjelasan dominan Pareto.

Kumpulan penjelasan semacam itu membentuk batas Pareto, suatu permukaan penjelasan yang darinya Alice dapat memilih mana pun yang paling membantu pemahamannya dan melayani kepentingannya. Perhatikan bahwa ini adalah gagasan optimalitas yang relatif lemah. Penjelasan mungkin optimal dalam pengertian ini tanpa memuaskan, karena seluruh batas Pareto mungkin berada di luar kuboid memuaskan yang didefinisikan oleh $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$. Dalam kasus ini, Alice dan Bob memiliki dua pilihan: (a) menerima bahwa tidak ada penjelasan yang akan memenuhi kriteria dan menyesuaikan ambang batas yang sesuai; atau (b) memulai babak baru dengan satu atau beberapa parameter masukan yang berbeda. Pilihan (b) akan menghasilkan permukaan penjelasan yang sepenuhnya baru untuk dijelajahi oleh para pemain.

Tanpa informasi lebih lanjut tentang fungsi target f atau fakta spesifik tentang pengetahuan dan minat Alice, dominasi Pareto bersyarat adalah bentuk optimalitas terkuat yang dapat kita harapkan secara wajar. Konvergensi pada batas Pareto hampir pasti terjamin pada tiga kondisi:

- *Kondisi 1.* Ruang fungsi G_b dan H berdimensi VC hingga.
- *Kondisi 2.* Jawaban untuk semua pertanyaan w dapat diidentifikasi secara unik.
- *Kondisi 3.* Alice adalah agen rasional dan pembaru Bayesian yang konsisten.

Kondisi (1) mensyaratkan konsistensi statistik model g milik Bob dan model h milik Alice, yang memastikan bahwa akurasi dan kesederhanaan diukur secara andal seiring bertambahnya ukuran sampel. Kondisi (2) mensyaratkan bahwa himpunan data simulasi setia pada proses pembangkitan data yang mendasarinya, sehingga memastikan bahwa g dan h

konvergen pada target yang tepat. Kondisi (3) mensyaratkan keberadaan setidaknya satu tindakan yang memaksimalkan utilitas $a^* \in A$ dengan ekspektasi posterior yang terdefinisi dengan baik $\mathbb{E}_{\mathbb{P}}(C)[u(a^*, C)|E]$. Jika probabilitasnya terkalibrasi dengan baik, maka Alice akan cenderung memilih tindakan yang "tepat", atau setidaknya tindakan tanpa alternatif yang lebih unggul di A . Dengan kondisi ini, setiap putaran permainan akan menghasilkan penjelasan yang tidak dapat diperbaiki tanpa mengubah parameter masukan.

Jika semua subrutin dari loop dalam permainan dieksekusi dalam waktu polinomial, maka putaran tersebut juga akan dieksekusi dalam waktu polinomial. Satu-satunya masalah yang berpotensi NP-hard adalah menemukan peta ψ yang memadai, yang tidak dapat dihitung secara efisien tanpa beberapa batasan pada himpunan solusi. Pendekatan naif adalah dengan mempertimbangkan semua kemungkinan subset dari ruang fitur asli, tetapi bahkan dalam pengaturan Markovian, hal ini akan menghasilkan peta 2^d yang sulit dikelola, di mana d merepresentasikan dimensionalitas matriks input $\mathbf{X}$. Pemetaan yang efisien memerlukan beberapa metode berprinsip untuk membatasi ruang ini hanya kepada peta-peta yang berpotensi menarik bagi Alice. Cara terbaik untuk melakukannya untuk setiap masalah adalah dengan bergantung pada konteks secara tak tereduksi.

10.6 PEMBELAJARAN MESIN YANG DINTERPRETASIKAN

Proposal iML saat ini tidak secara harfiah mewujudkan permainan penjelasan. Namun, kerangka kerja kami dapat diterapkan untuk mengevaluasi kelebihan dan kekurangan metode yang ada. Kerangka kerja ini juga menyediakan platform untuk mengonseptualisasikan kendala dan persyaratan dari setiap proposal iML yang memungkinkan, yang menjelaskan kontur ruang solusi. Metode iML yang paling populer saat ini adalah aproksimator linier lokal seperti LIME dan SHAP. SHAP menjelaskan prediksi dengan mengambil sampel secara acak di sekitar titik yang diinginkan.

Observasi diberi bobot berdasarkan jaraknya dari titik target dan model linier yang terregulasi dicocokkan dengan kuadrat terkecil terbobot. Yang terakhir dibangun di atas karya fundamental dalam teori permainan kooperatif, menggunakan data pelatihan untuk secara efisien menghitung aproksimasi titik demi titik dari nilai Shapley setiap fitur masukan. Hasil akhir dalam kedua kasus adalah serangkaian koefisien (yang mungkin jarang) yang menunjukkan hubungan positif atau negatif antara fitur masukan dan respons, setidaknya mendekati x_i dan bergantung pada kovariatnya.

Menggunakan LIME atau SHAP pada dasarnya sama saja dengan membatasi ruang fungsi model penjelasan Bob g ke kelas model linier terregulasi. Setiap metode memiliki kernel default k -nya sendiri, serta fungsi pemetaan ψ yang direkomendasikan untuk tipe data tertentu. Misalnya, LIME mengkasar data gambar menjadi superpiksel, sementara SHAP menggunakan peta saliency untuk memvisualisasikan bagian-bagian gambar masukan yang paling penting dalam menentukan klasifikasinya. Meskipun penulis kedua metode tersebut tampaknya menyarankan bahwa satu kali eksekusi dari salah satu algoritma sudah cukup untuk tujuan penjelasan, aproksimasi linier lokal cenderung tidak stabil untuk titik data yang berada di dekat bagian nonlinier dari batas keputusan atau permukaan regresi. Dengan demikian,

beberapa kali pengujian dengan data yang terganggu mungkin diperlukan untuk menetapkan presisi bobot fitur yang diestimasi.

Hal ini setara dengan beberapa putaran permainan penjelasan, sehingga memberikan Alice gambaran yang lebih lengkap tentang ruang model. Salah satu masalah utama dengan LIME dan SHAP adalah kedua metode tersebut tidak memungkinkan pengguna untuk menentukan kelas kontras yang diinginkan. Perilaku default kedua algoritma ini adalah menjelaskan mengapa suatu hasil adalah $\hat{y}_i$, bukan $\bar{y}$ – yaitu, respons rata-rata untuk seluruh dataset (nyata atau simulasi).

Dalam banyak konteks, hal ini masuk akal. Misalnya, jika Alice menerima diagnosis yang langka dan tak terduga, ia mungkin ingin tahu apa yang membedakannya dari mayoritas pasien. Namun, tampaknya aneh untuk menyarankan, seperti yang secara implisit dilakukan oleh algoritma ini, bahwa prediksi "normal" tidak dapat dijelaskan. Tidak ada yang membingungkan atau tidak pantas tentang Alice yang bertanya-tanya, misalnya, mengapa ia menerima skor kredit rata-rata, alih-alih yang lebih baik dari rata-rata. Namun dalam bentuknya saat ini, baik LIME maupun SHAP tidak dapat mengakomodasi pertanyaan semacam itu.

Terdapat alternatif yang lebih fleksibel. Daftar aturan, yang memprediksi hasil melalui serangkaian pernyataan jika-maka, dapat memodelkan efek nonlinier yang pada prinsipnya tidak dapat dideteksi oleh LIME dan SHAP. Beberapa solusi iML dibangun di atas partisi rekursif prosedur statistik yang menghasilkan daftar aturan dan semakin banyak studi psikologis menunjukkan bahwa pengguna menganggap penjelasan semacam itu sangat mudah dipahami. Jika Alice adalah salah satu dari banyak orang yang memiliki preferensi serupa terhadap daftar aturan, maka Bob harus mempertimbangkan hal ini saat memilih G_b .

Penjelasan kontrafaktual didukung oleh Wachter, yang mengusulkan solusi iML baru berdasarkan jaringan adversarial generatif (GAN). Berdasarkan penelitian pionir dalam pembelajaran mendalam, para penulis mendemonstrasikan bagaimana GAN dapat digunakan untuk menemukan gangguan minimal pada fitur masukan yang cukup untuk mengubah keluaran dengan cara yang telah ditentukan sebelumnya. Model-model ini tidak seketat regresi linier atau daftar aturan, karena tidak hanya memungkinkan pengguna untuk mengidentifikasi kelas kontras tetapi pada prinsipnya dapat beradaptasi dengan fungsi terdiferensiasi apa pun. menekankan pentingnya kesederhanaan dengan menerapkan batasan kelangkaan pada keluaran penjelasan yang dimaksudkan untuk secara otomatis menghapus fitur yang tidak informatif.

Daftar aturan dan GAN memiliki beberapa keunggulan yang jelas dibandingkan aproksimator linier seperti LIME dan SHAP. Namun, tidak ada metode yang digunakan saat ini yang secara eksplisit memperhitungkan kepentingan pengguna, sebuah kelalaian yang dapat menyebabkan hasil yang tidak diinginkan. Singkatnya, mereka tidak lulus uji gerhana. Ingat kembali kasus bank (buruk). Misalkan himpunan pilihan Alice hanya berisi dua opsi, $A = \{\text{Tuntut, Jangan tuntutan}\}$, dan ia mempertimbangkan dua hipotesis kausal sebagai penjelasan potensial atas penolakan pinjamannya, $C = \{\text{Kekayaan, Ras}\}$. Matriks utilitas Alice disajikan pada Tabel 10.2.

Tabel 10.2 Matriks utilitas untuk Alice dalam skenario bank (buruk)

	c_2 : kekayaan	c_2 : Balapan
a_1 : Menuntut	-1	5
a_1 : Jangan menuntut	0	0

Alice menetapkan prior seragam atas C sejak awal, sehingga $\mathbb{P}(c_1) = \mathbb{P}(c_2) = 0,5$. Ia menerima dua penjelasan dari Bob: g_1 , yang menyatakan bahwa permohonan Alice ditolak karena kekayaannya; dan g_2 , yang menyatakan bahwa permohonan Alice ditolak karena rasnya. Dengan menggunakan tingkat kesalahan klasifikasi sebagai fungsi kerugian kami dan mengasumsikan massa probabilitas seragam atas fitur dikotomis Kekayaan $\in \{\text{Kaya, Miskin}\}$ dan Ras $\in \{\text{Putih, Hitam}\}$, kami menemukan bahwa kedua penjelasan tersebut sama akuratnya:

$$R_{emp}(g_1, \mathbf{Z}_f) = R_{emp}(g_2, \mathbf{Z}_f) = 0.25$$

dan sama sederhananya:

$$R_{emp}(h, \mathbf{Z}_{g1}) = R_{emp}(h, \mathbf{Z}_{g2}) = 0$$

Namun, mereka menghasilkan posterior yang sangat berbeda pada C :

$$\mathbb{P}(c_1|g_1) = \mathbb{P}(c_2|g_2) = 0.9$$

$$\mathbb{P}(c_1|g_2) = \mathbb{P}(c_2|g_1) = 0.1$$

Utilitas harapan posterior a_1 di bawah g_1 adalah

$$0.9(-1) + 0.1(5) = -0.4,$$

sedangkan di bawah g_2 harapannya adalah

$$0.1(-1) + 0.9(5) = -4.4.$$

(Utilitas yang diharapkan dari a_2 adalah 0 pada kedua penjelasan.) Karena tindakan memaksimalkan utilitas pada g_2 lebih disukai daripada tindakan memaksimalkan utilitas pada g_1 , kami menganggap g_2 sebagai penjelasan yang lebih unggul. Faktanya, Pareto-mendominasi yang pertama, karena keduanya setara dalam hal akurasi dan kesederhanaan, tetapi g_1 kurang relevan bagi Alice dibandingkan g_2 . Penentuan ini hanya dapat dilakukan dengan mengodekan preferensi Alice secara eksplisit, yang saat ini diabaikan oleh semua proposal iML utama.

Metode yang gagal lulus uji gerhana menimbulkan masalah bagi ketiga tujuan iML yang diuraikan di Bagian 10.2. Penjelasan yang tidak relevan dapat merusak uji validitas atau pencarian penemuan karena gagal mengenali tujuan epistemologis yang memotivasi pertanyaan tersebut sejak awal. Ketika penjelasan tersebut akurat dan sederhana, Alice dapat dengan mudah tertipu dan berpikir bahwa ia telah mempelajari beberapa informasi berharga.

Faktanya, Bob hanya melakukan overfitting pada data. Situasinya bahkan lebih buruk ketika kita mencoba mengaudit algoritma.

Dalam kasus ini, penjelasan yang mengaburkan fakta justru dapat menawarkan celah bagi pihak-pihak yang tidak bertanggung jawab yang ingin menghindari kontroversi atas keputusan yang dipertanyakan. Misalnya, fokus yang sempit pada akurasi dan kesederhanaan akan memungkinkan bank (yang tidak bertanggung jawab) lolos dari kebijakan pinjaman rasis selama pemohon kulit hitam dianggap kurang memenuhi kriteria di sepanjang sumbu variasi lainnya.

10.7 KEBERATAN

Di bagian ini, kami mempertimbangkan lima keberatan yang semakin umum. Tiga keberatan pertama ditujukan terhadap permainan yang kami usulkan, sedangkan dua keberatan terakhir ditujukan terhadap keseluruhan proyek iML.

Terlalu Idealis

Salah satu keberatan yang jelas terhadap proposal kami adalah bahwa proposal ini menuntut banyak Alice. Ia harus memberikan hasil kontrastif $\tilde{y}_i$, tingkat abstraksi LoA, himpunan pilihan A , beberapa hipotesis kausal C , distribusi prior yang sesuai $\mathbb{P}(C)$, dan fungsi utilitas u . Selain itu semua, kami juga mengharapkannya menjadi pemutakhir Bayesian yang konsisten dan pemaksimal utilitas yang diharapkan. Jika Alice begitu diperlengkapi dan sangat rasional, mungkin memecahkan algoritma kotak hitam tidak akan menjadi tantangan besar baginya.

Tanggapan kami ada dua. Pertama, kami mengingatkan pembaca yang skeptis bahwa idealisasi adalah alat yang populer dan bermanfaat dalam analisis konseptual. Tidak ada bidang tanpa gesekan atau populasi tak terhingga, tetapi asumsi-asumsi tersebut telah berkontribusi pada teori-teori yang sukses dalam fisika dan genetika. Potochnik (2017) memberikan argumen yang meyakinkan bahwa idealisasi penting bagi praktik ilmiah, memungkinkan manusia untuk merepresentasikan dan memanipulasi sistem dengan kompleksitas yang tak terpahami. Teori keputusan pun demikian. Asumsi bahwa agen epistemik selalu membuat pilihan rasional – meskipun secara tegas salah – telah memajukan pemahaman kita tentang perilaku individu dan sosial dalam ilmu ekonomi, psikologi, dan ilmu komputer.

Kedua, pengaturan ini tidaklah se-realistis kelihatannya. Sangat masuk akal untuk berasumsi bahwa seorang agen akan mencari penjelasan algoritmik dengan setidaknya hasil kontrafaktual dan pilihan yang telah disiapkan, serta beberapa hipotesis kausal (tentatif). Misalnya, Alice mungkin memasuki permainan secara eksplisit karena ia menduga pengajuan pinjamannya ditolak karena rasnya, dan tidak yakin apakah akan mencari ganti rugi. Utilitas dapat diturunkan melalui pemeringkatan sederhana dari semua pasangan tindakan-hasil. Jika hipotesis baru muncul selama permainan, hipotesis tersebut dapat dengan mudah dieksplorasi di babak selanjutnya.

Alice mungkin kurang yakin dengan nilai ideal untuk LoA dan $\mathbb{P}(C)$, tetapi tidak ada alasan untuk menuntut kepastian tentang hal ini sejak awal. Memang, disarankan untuk

mencoba berbagai nilai untuk masing-masing, seperti bagaimana analisis sering bereksperimen dengan prior yang berbeda untuk memastikan dampak pada posterior dalam inferensi Bayesian. Alice dan Bob dapat menyempurnakan masukan mereka secara iteratif seiring berjalannya babak dan melacak evolusi batas Pareto yang dihasilkan untuk mengukur ketidakpastian yang terkait dengan berbagai parameter. Proses seperti inilah yang sebenarnya menjadi dasar banyak penelitian dilakukan.

Mungkin yang paling penting, kami menekankan bahwa Alice dan Bob adalah agen umum yang dapat dan sering kali akan diimplementasikan oleh sistem hibrida yang melibatkan banyak manusia dan mesin yang bekerja bersama. Tidak ada alasan untuk membatasi sumber daya kognitif keduanya secara artifisial hanya pada individu tertentu. Permasalahan yang dirancang untuk diatasi oleh IML berada di luar jangkauan satu orang pun, terutama mereka yang beroperasi tanpa bantuan perangkat lunak statistik.

Ketika kami memperluas cakupan kognitif Alice dan Bob, idealisasi yang dituntut dari mereka menjadi jauh lebih masuk akal. Satu-satunya batas atas yang relevan untuk kapasitas inferensial mereka adalah ambang batas kompleksitas komputasi. Permainan penjelasan adalah sebuah latihan epistemologi sosioteknis, di mana pengetahuan muncul dari interaksi berkelanjutan antara individu, kelompok, dan teknologi. Poin pentingnya adalah apakah permainan penjelasan yang telah kami rancang memungkinkan dan bermanfaat, bukan apakah Alice dan Bob tertentu benar-benar dapat memainkannya sesuai dengan kemampuan idiosinkratik mereka.

Regresi Tak Terbatas

Tantangan umum untuk setiap penjelasan adalah ancaman regresi tak terbatas. Dengan asumsi bahwa penjelasan harus terbatas, bagaimana kita bisa yakin bahwa beberapa metode penjelasan berakhir pada terminus yang tepat? Dalam hal ini, bagaimana kita bisa menjamin bahwa permainan penjelasan tidak berubah menjadi perulangan rekursif tak terbatas? Perhatikan bahwa ini bukan masalah untuk Alice dan Bob yang tetap setiap putaran berakhir setelah model g dan h dinilai, dan utilitas yang diharapkan Alice diperbarui tetapi keberatan tersebut tampak lebih mengancam pada agen dan permainan yang bergeser.

Misalnya, kita mungkin khawatir bahwa Alice dan Bob bersama-sama membentuk algoritma pembelajaran terawasi baru f_2 yang memetakan masukan x_i ke keluaran $h(x'_1)$ melalui model antara g . Fungsi yang dihasilkan sekarang dapat ditanyakan oleh agen baru $Alice_2$ yang meminta bantuan Bob_2 dalam memperhitungkan beberapa prediksi $f_2(x_i)$. Proses ini dapat berulang tanpa batas.

Kesalahan dalam penalaran ini adalah mengabaikan peran vital pragmatik. Dengan konstruksi, setiap permainan berakhir pada terminus yang tepat *untuk Alice tertentu*. Tidak ada yang salah dengan mengizinkan agen lain untuk menyelidiki produk dari permainan tersebut seolah-olah itu adalah algoritma baru. Hasilnya hanya akan berjarak t langkah dari sumber aslinya, di mana t adalah jumlah tim Alice-dan-Bob yang memisahkan f awal dari penanya terbaru.

Efeknya tidak jauh berbeda dengan permainan telepon, di mana pesan secara bertahap menurun saat pemain memperkenalkan kesalahan baru di setiap iterasi. Demikian pula, setiap

pasangan Alice-dan-Bob yang baru akan melakukan yang terbaik untuk memperkirakan pekerjaan tim sebelumnya. Hasil akhirnya mungkin terlihat sangat berbeda dengan f asli untuk beberapa nilai t yang besar, tetapi itu hanya yang diharapkan. Selama kondisi (1)–(3) terpenuhi untuk setiap Alice dan Bob yang diberikan, maka mereka hampir pasti dijamin untuk bertemu pada permukaan penjelasan yang optimal secara kondisional dalam waktu polinomial.

Pragmatisme + Pluralisme = Anarki Relativis?

Permainan penjelasan sangat bergantung pada pertimbangan pragmatis. Kami secara eksplisit menganjurkan gagasan subjektif tentang kesederhanaan dan relevansi, yang memungkinkan Bob untuk membangun berbagai penjelasan pada berbagai tingkat abstraksi. Kombinasi subjektivisme dan pluralisme ini bertentangan dengan tradisi realis dalam epistemologi dan filsafat ilmu, yang menyatakan bahwa hanya ada satu penjelasan yang benar untuk setiap penjelasan yang diberikan. Bukankah ada bahaya tergelincir ke dalam semacam relativisme yang tidak terhormat? Jika kriteria keberhasilan penjelasan begitu subjektif, bukankah tidak ada fakta yang jelas tentang mana dari dua penjelasan yang bersaing yang lebih unggul? Bukankah ini sama saja dengan mengatakan bahwa apa pun boleh?

Jawaban singkatnya adalah tidak. Keberatan tersebut mengasumsikan bahwa untuk setiap fakta atau peristiwa tertentu terdapat suatu penjelasan yang unik, memuaskan, dan independen dari pikiran dan konteks, mungkin dalam hal unit dan hukum fisika fundamental. Sebut pandangan ini monisme penjelasan. Ini merupakan doktrin metafisika yang kelebihan atau kekurangannya sejauhnyanya tidak relevan. Karena meskipun penjelasan yang "benar" selalu tersedia, secara umum tidak akan banyak berguna. Tujuan dari permainan penjelasan adalah untuk *meningkatkan pemahaman Alice*. Ini dapat datang dalam berbagai bentuk. Misalnya, prediksi pengklasifikasi gambar sering dijelaskan oleh peta panas yang menyoroti piksel yang paling berkontribusi pada keluaran yang diberikan. Fakta bahwa rumus matematika yang kompleks dalam hal ini dapat memberikan penjelasan yang sangat mendalam dan stabil tidaklah relevan. Tujuan pragmatis membutuhkan strategi pragmatis. Karena iML pada dasarnya adalah tentang membuat manusia memahami perilaku mesin, ada seruan yang semakin besar untuk solusi yang dipersonalisasi. Kami menanggapi perubahan pragmatis ini dengan serius dan mengusulkan metode formal untuk mengimplementasikannya.

Kami dengan tegas menolak tuduhan bahwa permainan penjelasan begitu permisif sehingga "apa pun boleh". Jauh dari itu, kami mendefinisikan ukuran objektif dari gagasan subjektif yang telah lama menentang formalisasi yang tegas. Setelah nilai untuk semua variabel ditentukan, adalah masalah langsung untuk menilai dan membandingkan penjelasan yang bersaing. Untuk setiap set parameter input, ada urutan penjelasan yang unik dalam hal akurasi relatif, kesederhanaan, dan relevansinya.

Penjelasan pada berbagai tingkat abstraksi mungkin tidak sepadan, tetapi bersama-sama mereka dapat membantu Alice membentuk gambaran yang lebih lengkap tentang sistem target dan perilakunya di dekat titik data yang diminati. Kombinasi pragmatisme dan ekumenisme penjelasan ini fleksibel dan rasional. Ini merangkul relasionalisme, bukan relativisme. Salah satu kontribusi utama dari makalah ini adalah untuk menunjukkan bahwa

desiderata iML dapat dirumuskan dengan presisi dan ketelitian tanpa mengorbankan aspek subjektif dan kontekstual yang membuat setiap permainan penjelasan unik.

Tanpa Kompromi

Beberapa pihak telah menentang asumsi luas bahwa terdapat kompromi inheren antara akurasi dan interpretabilitas dalam pembelajaran mesin. Rudin (2019) dengan tegas menentang pandangan ini, yang menurutnya didasarkan pada bukti anekdotal, atau bahkan kerahasiaan perusahaan. Ia mencatat bahwa sains telah lama menunjukkan preferensi terhadap model yang lebih sederhana, bukan hanya karena imajinasi estetika, tetapi karena prinsip-prinsip yang berdasar mengenai kesederhanaan inheren alam. Hasil terbaru dalam teori pembelajaran formal menegaskan bahwa pendekatan Ockham's Razor untuk pengujian hipotesis adalah strategi optimal untuk konvergensi menuju kebenaran dengan batasan topologi minimal.

Breiman (2001) secara terkenal memperkenalkan gagasan *Rashomon set* – sekumpulan model yang mengestimasi hubungan fungsional yang sama menggunakan algoritma dan/atau hiperparameter yang berbeda, namun semuanya berkinerja cukup baik (katakanlah, dalam kisaran 5% dari model berkinerja terbaik). Argumen Rudin yang diperluas secara teknis dalam makalah lanjutan didasarkan pada asumsi bahwa himpunan Rashomon yang cukup besar seharusnya mencakup setidaknya satu model yang dapat diinterpretasikan. Jika demikian, maka tampaknya tidak ada gunanya menjelaskan algoritma kotak hitam, setidaknya dalam aplikasi berisiko tinggi seperti layanan kesehatan dan peradilan pidana. Jika kita harus menggunakan ML untuk tujuan ini, maka kita sebaiknya melatih model yang dapat diinterpretasikan (secara global) sejak awal, alih-alih merekayasa balik penjelasan post-hoc yang tidak sempurna.

Ada dua masalah dengan keberatan ini. Pertama, tidak ada jaminan logis atau statistik bahwa model yang dapat diinterpretasikan akan mengungguli pesaing kotak hitam atau bahkan termasuk dalam himpunan Rashomon model berkinerja tinggi untuk setiap masalah prediktif yang diberikan. Ini adalah akibat wajar sederhana dari teorema "tidak ada makan siang gratis" yang terkenal, yang menyatakan (secara kasar) bahwa tidak ada solusi yang cocok untuk semua dalam ML. Algoritma apa pun yang berkinerja baik pada satu kelas masalah pasti akan berkinerja buruk pada kelas masalah lainnya. Tentu saja, hal ini berlaku dua arah algoritma kotak hitam juga dijamin gagal pada beberapa set data. Jika kita mengutamakan kinerja di atas segalanya, yang mungkin berlaku untuk beberapa tugas yang sangat penting, maka kita harus terbuka terhadap model dengan interpretabilitas yang bervariasi.

Kedua, ketidakjelasan algoritma kotak hitam bukan hanya produk sampingan dari teknik statistik yang kompleks, tetapi juga realitas kelembagaan yang kemungkinan besar tidak akan berubah dalam waktu dekat. Pasquale (2015) menawarkan sejumlah studi kasus yang mengesankan yang menunjukkan bagaimana hukum kekayaan intelektual digunakan secara luas untuk melindungi kode sumber dan data pelatihan pembelajaran mesin (ML), tidak hanya dari calon pesaing tetapi juga dari segala bentuk pengawasan eksternal. Bahkan jika sebuah perusahaan menggunakan model yang dapat diinterpretasikan untuk membuat prediksinya, arsitektur dan parameter model tersebut kemungkinan akan tunduk pada perlindungan hak

cipta yang ketat. Beberapa pihak telah mengusulkan pembentukan kelompok pihak ketiga independen yang bertanggung jawab mengaudit kode berdasarkan perjanjian kerahasiaan, sebuah proposal yang kami dukung secara pribadi. Namun, hingga undang-undang tersebut disahkan, siapa pun yang mencoba memantau keadilan, akuntabilitas, dan transparansi algoritma hampir pasti tidak punya pilihan selain memperlakukan teknologi yang mendasarinya sebagai kotak hitam.

Standar Ganda

berpendapat bahwa para pendukung iML memberikan beban yang tidak masuk akal pada algoritma dengan menuntut agar algoritma tersebut tidak hanya berkinerja lebih baik dan lebih cepat daripada manusia, tetapi juga menjelaskan mengapa mereka melakukannya. Mereka menunjukkan bahwa pengambilan keputusan manusia jauh dari transparan, dan bahwa manusia terkenal buruk dalam membenarkan tindakan mereka. Mengapa ada standar ganda? Kita sudah memiliki sistem untuk mengakreditasi para pengambil keputusan manusia dalam posisi otoritas (misalnya, hakim dan dokter) berdasarkan rekam jejak kinerja mereka yang telah terbukti. Mengapa kita harus mengharapkan sesuatu yang lebih dari mesin? Para penulis menyimpulkan bahwa mensyaratkan kejelasan algoritma berkinerja tinggi bukan hanya tidak masuk akal tetapi juga berpotensi berbahaya jika menghambat implementasi model yang dapat meningkatkan layanan bagi pengguna akhir.

benar ketika menunjukkan bahwa kita seringkali menjadi narator yang tidak dapat diandalkan atas penalaran internal kita sendiri. Kita cenderung merasionalisasi impuls irasional, menarik kesimpulan yang salah, dan membuat keputusan berdasarkan serangkaian heuristik dan bias kognitif yang terdokumentasi dengan baik. Namun, justru inilah yang membuat iML begitu menjanjikan: bukan karena algoritma pembelajaran entah bagaimana kebal terhadap bias manusia – tidak demikian, setidaknya jika bias tersebut termanifestasi dalam data pelatihan – melainkan karena, dengan perangkat yang tepat, kita dapat secara meyakinkan mengungkapkan alasan sebenarnya di balik keputusan yang berdampak. memberikan argumen kuat bahwa peningkatan otomatisasi akan mengurangi diskriminasi dengan memulai prosedur yang ketat dan objektif untuk mengaudit dan mengajukan banding terhadap prediksi algoritmik. Sangat sulit di bawah hukum saat ini untuk membuktikan bahwa manusia telah terlibat dalam perilaku diskriminatif, terutama jika mereka bersikeras tidak melakukannya (yang biasanya dilakukan kebanyakan orang, terutama ketika diancam dengan sanksi hukum). Terlepas dari semua potensi bahaya yang ditimbulkan oleh algoritma, penipuan yang disengaja (belum) menjadi salah satunya.

Kami berpendapat bahwa potensi manfaat dari strategi iML yang sukses lebih beragam dan banyak daripada yang diakui Kleinberg. Untuk menegaskan kembali motivasi yang tercantum di Bagian 10.2, kami melihat tiga area yang sangat menjanjikan. Dalam kasus audit algoritmik, iML dapat membantu memastikan penerapan model statistik kompleks yang adil, akuntabel, dan transparan dalam aplikasi berisiko tinggi seperti peradilan pidana dan layanan kesehatan. Dalam kasus validasi, iML dapat digunakan untuk menguji algoritma sebelum dan selama penerapan guna memastikan bahwa model berkinerja dengan baik dan tidak overfitting terhadap pola yang tidak informatif dalam data pelatihan. Dalam kasus penemuan,

iML dapat mengungkap mekanisme yang sebelumnya tidak diketahui dalam sistem target yang kompleks, sehingga menyarankan teori dan hipotesis baru untuk pengujian. Tentu saja, tidak ada jaminan bahwa metode tersebut akan berhasil dalam setiap kasus iML bukanlah obat mujarab tetapi akan bodoh jika tidak mencobanya. Standar ganda yang diperingatkan Zerilli dkk. sebenarnya merupakan peluang yang disambut baik.

10.8 KESIMPULAN

Algoritma kotak hitam akan tetap ada. Lembaga swasta dan publik sudah mengandalkan ML untuk menjalankan fungsi-fungsi dasar dan kompleks dengan efisiensi dan akurasi yang lebih tinggi daripada manusia. Kumpulan data yang terus bertambah dan perangkat keras yang terus ditingkatkan, dikombinasikan dengan kemajuan berkelanjutan dalam ilmu komputer dan statistik, memastikan bahwa metode-metode ini akan semakin umum di tahun-tahun mendatang.

Ada sedikit alasan untuk percaya bahwa algoritma akan menjadi lebih transparan atau mudah dipahami, setidaknya tanpa upaya eksplisit dan berkelanjutan dari para peneliti yang berdedikasi di bidang iML yang sedang berkembang pesat. Kami berpendapat bahwa ada alasan kuat untuk menghargai interpretabilitas algoritmik atas dasar etika, epistemologis, dan ilmiah. Kami telah menguraikan kerangka kerja formal di mana agen dapat berkolaborasi untuk menjelaskan keluaran dari setiap pembelajar yang diawasi. Permainan penjelasan berfungsi baik sebagai fungsi deskriptif – menyediakan bahasa umum untuk membandingkan proposal iML – maupun sebagai fungsi normatif – menyoroti aspek-aspek yang belum dieksplorasi dalam literatur saat ini dan menunjukkan jalan menuju solusi baru dan yang lebih baik. Tentu saja, tantangan normatif yang penting tetap ada. Pertanyaan pelik tentang keadilan algoritmik, akuntabilitas, dan transparansi tidak semuanya dapat diselesaikan dengan cepat. Namun, kami berharap permainan penjelasan ini dapat menginformasikan perdebatan ini secara produktif dan berprinsip.

Penelitian selanjutnya akan melonggarkan asumsi yang mendasari versi beta permainan ini. Yang menarik adalah alternatif adversarial di mana Bob memiliki fungsi utilitasnya sendiri untuk dimaksimalkan, atau versi tiga pemain di mana Carol dan Bob bersaing untuk menemukan penjelasan yang lebih unggul yang harus dipilih Alice. Arahan lain yang menjanjikan termasuk penerapan permainan penjelasan semi-otomatis dengan algoritma greedy yang secara bergantian memaksimalkan satu desideratum penjelasan pada satu waktu hingga mencapai konvergensi. Proposal serupa telah diimplementasikan untuk mengoptimalkan tujuan campuran dalam keadilan algoritmik, tetapi kami belum mengetahui adanya penelitian serupa dalam bidang explainability. Terakhir, kami bermaksud memperluas cakupan kami ke algoritma pembelajaran tanpa pengawasan, yang menghadirkan sejumlah tantangan penjelasan yang sama sekali berbeda.

BAB 11

AGEN BUATAN DAN SIFAT MORALNYA

Dalam perkembangan teknologi modern, agen buatan terutama yang beroperasi di dalam infosfer seperti yang dikemukakan Floridi memperluas jenis entitas yang dapat terlibat dalam situasi moral. Agen buatan ini dapat dipahami sebagai entitas yang mampu melakukan tindakan yang berdampak baik maupun buruk, sehingga mereka dapat dianggap sebagai agen moral.

Bab ini bertujuan untuk menjelaskan secara rinci konsep agen dan agen buatan, sekaligus membedakan antara isu yang berkaitan dengan perilaku moral mereka dan isu yang berkaitan dengan tanggung jawab mereka. Pendekatan ini menekankan bahwa ada ruang lingkup penting dalam etika informasi untuk memahami agen buatan moral, bahkan ketika agen tersebut tidak memiliki kehendak bebas, kondisi mental, atau tanggung jawab secara tradisional.

Pendekatan ini berbeda dari pandangan tradisional yang lebih menekankan pada kemungkinan agen buatan memiliki kondisi mental, perasaan, atau emosi. Dengan memusatkan perhatian pada konsep "moralitas tanpa pikiran", kita dapat menghindari pertanyaan-pertanyaan yang kompleks mengenai kesadaran atau emosi dalam agen buatan. Pendekatan ini juga membantu mengatasi sejumlah tantangan etis yang muncul seiring agen buatan semakin menjadi bagian dari kehidupan sehari-hari.

11.1 PENDAHULUAN

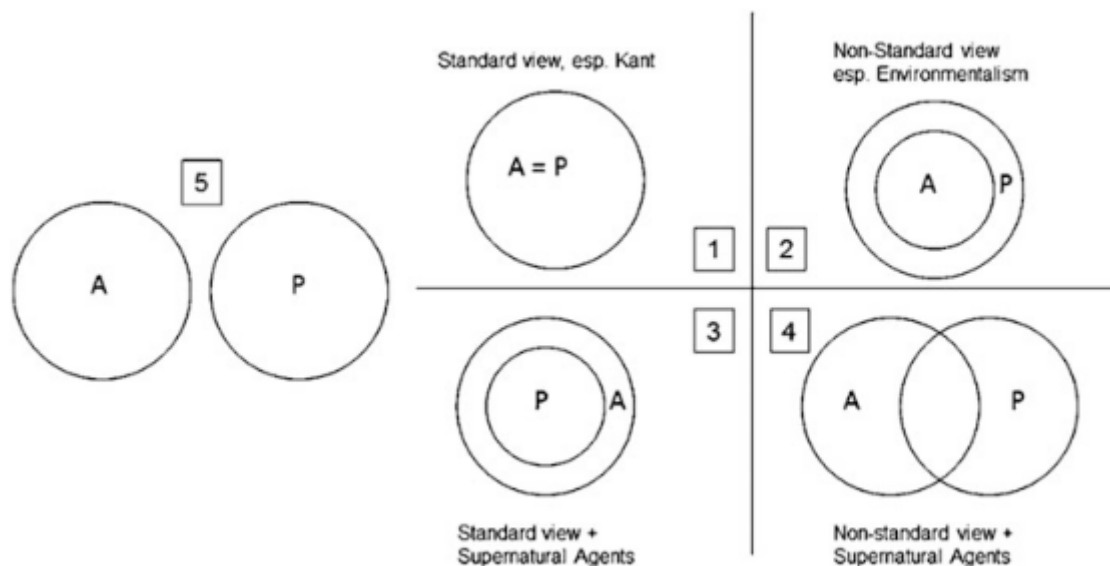
Teori Standar vs. Non-Standar tentang Agen dan Pasien

Situasi moral umumnya melibatkan agen dan pasien. Mari kita definisikan kelas A agen moral sebagai kelas semua entitas yang pada prinsipnya dapat memenuhi syarat sebagai sumber atau pengirim tindakan moral, dan kelas P pasien moral sebagai kelas semua entitas yang pada prinsipnya dapat memenuhi syarat sebagai penerima tindakan moral. Cara yang tepat untuk memperkenalkan topik bab ini adalah dengan mempertimbangkan bagaimana teori etika (*makroetika*) menafsirkan hubungan logis antara kedua kelas tersebut. Terdapat lima hubungan logis antara A dan P, lihat Gambar 11.1.

Mungkin saja, tetapi sama sekali tidak realistis, bahwa A dan P terpisah (alternatif 5). Di sisi lain, P dapat menjadi subhimpunan wajar dari A (alternatif 3), atau A dan P dapat saling berpotongan (alternatif 4). Kedua alternatif ini hanya sedikit lebih menjanjikan karena keduanya membutuhkan setidaknya satu agen moral yang pada prinsipnya tidak dapat memenuhi syarat sebagai pasien moral. Agen murni ini akan menjadi semacam entitas supernatural yang, seperti Tuhan dalam Aristoteles, memengaruhi dunia tetapi tidak pernah dapat dipengaruhi olehnya.

Namun, karena pada prinsipnya "tidak dapat dipengaruhi" dan tidak relevan dalam permainan moral, tidak jelas peran seperti apa yang akan dijalankan entitas ini sehubungan dengan pedoman normatif tindakan manusia. Jadi, tidak mengherankan jika sebagian besar

makroetika menjauhi spekulasi "supranatural" ini dan secara implisit mengadopsi, atau bahkan secara eksplisit mendukung, salah satu dari dua alternatif tersisa yang dibahas dalam teks: A dan P dapat setara (alternatif 1), atau A dapat menjadi subhimpunan wajar dari P (alternatif 2). Alternatif (1) menyatakan bahwa semua entitas yang memenuhi syarat sebagai agen moral juga memenuhi syarat sebagai pasien moral, dan sebaliknya.



Gambar 11.1 Hubungan logis antara kelas agen moral dan pasien

Hal ini sesuai dengan posisi yang agak intuitif, yang menurutnya agen/peneliti memainkan peran protagonis moral. Kita, agen moral manusia yang juga menyelidiki hakikat moralitas, menempatkan diri kita di pusat permainan moral sebagai satu-satunya pemain yang dapat bertindak secara moral, ditindaklanjuti secara moral, dan pada akhirnya berteori tentang semua ini. Ini adalah salah satu pandangan paling populer dalam sejarah etika, yang dianut misalnya oleh banyak Ahli Etika Kristen pada umumnya dan oleh Kant pada khususnya. Saya akan menyebutnya sebagai posisi standar.

Alternatif (2) menyatakan bahwa semua entitas yang memenuhi syarat sebagai agen moral juga memenuhi syarat sebagai pasien moral, tetapi tidak sebaliknya. Banyak entitas, terutama hewan, tampaknya memenuhi syarat sebagai pasien moral, meskipun pada prinsipnya mereka dikecualikan dari memainkan peran agen moral. Pendekatan pasca-lingkungan ini membutuhkan perubahan perspektif, dari orientasi agen menjadi orientasi pasien. Mengingat label sebelumnya, saya akan menyebutnya sebagai non-standar.

Dalam beberapa tahun terakhir, makroetika non-standar telah membahas cakupan P secara cukup ekstensif. Semakin inklusif P, semakin "hijau" atau "mendalam" pendekatan tersebut dianggap. Khususnya etika lingkungan telah berkembang sejak tahun 1960-an sebagai studi tentang hubungan moral manusia dengan lingkungan (termasuk isi dan penghuni nonmanusianya) serta nilai-nilai (yang mungkin) dan status moralnya. Hal ini seringkali

menjadi tantangan bagi pendekatan antroposentris yang tertanam dalam beberapa pemikiran etika Barat tradisional.

Relatif sedikit penelitian yang telah dilakukan untuk mempertimbangkan kembali hakikat keagenan moral, dan dengan demikian perluasan A. Pemikiran pasca-lingkungan, dalam upaya mencapai etika yang sepenuhnya dinaturalisasi, secara implisit menolak relevansi, jika bukan kemungkinan, agen supernatural, sementara plausibilitas dan pentingnya jenis-jenis keagenan moral lainnya tampaknya sebagian besar telah diabaikan. Sekularisme telah menyusutkan (ada yang mengatakan mengempiskan) A, sementara environmentalisme secara wajar hanya memperluas P, sehingga kesenjangan antara A dan P semakin melebar; Hal ini disertai dengan peningkatan tanggung jawab moral individu yang sangat besar.

Beberapa upaya telah dilakukan untuk memperbaiki situasi ini. Secara khusus, konsep "agen moral" telah diperluas untuk mencakup baik orang perseorangan maupun badan hukum, terutama dalam etika bisnis. A kemudian diperluas untuk mencakup agen seperti kemitraan, pemerintah, atau perusahaan, yang hak dan kewajiban hukumnya telah diakui. Pendekatan yang lebih ekumenis ini telah memulihkan keseimbangan antara A dan P. Sebuah perusahaan kini dapat dimintai pertanggungjawaban langsung atas apa yang terjadi pada lingkungan, misalnya. Namun pendekatan tersebut tetap dibatasi secara tidak semestinya oleh konsepsi antroposentrisnya tentang keagenan. Suatu entitas masih dianggap sebagai agen moral hanya jika

- (i) ia adalah agen individu; dan
- (ii) ia berbasis manusia, dalam arti bahwa ia bersifat manusiawi atau setidaknya dapat direduksi menjadi kumpulan manusia yang dapat diidentifikasi, yang tetap menjadi satu-satunya sumber tindakan yang bertanggung jawab secara moral, seperti hantu dalam mesin hukum.

Membatasi wacana etika hanya pada agen-agen individual menghambat pengembangan investigasi yang memuaskan terhadap moralitas terdistribusi, sebuah fenomena makroskopis yang terus berkembang mengenai tindakan moral global dan tanggung jawab kolektif yang dihasilkan dari "tangan tak terlihat" interaksi sistemik di antara beberapa agen di tingkat lokal.

Menegaskan sifat niscaya berbasis manusia dari agen-agen individual tersebut berarti merongrong kemungkinan untuk memahami transformasi besar lainnya dalam bidang etika, yaitu munculnya agen-agen buatan (AA) yang cukup terinformasi, "cerdas", otonom, dan mampu melakukan tindakan-tindakan yang relevan secara moral secara independen dari manusia yang menciptakannya, yang menyebabkan "kebaikan buatan" dan "kejahatan buatan". Kedua kendala tersebut dapat dihilangkan dengan merevisi sepenuhnya konsep "agen moral". Inilah tugas yang akan dilakukan di halaman-halaman berikutnya.

Tesis utama yang dipertahankan adalah bahwa AA merupakan sumber sah tindakan amoral, sehingga kelas A agen moral harus diperluas agar mencakup AA, bahwa wacana etika harus mencakup analisis moralitas mereka, dan, akhirnya, bahwa analisis ini penting untuk memahami berbagai masalah moral baru, tidak hanya dalam etika informasi tetapi juga dalam etika secara umum, terutama dalam kasus moralitas terdistribusi.

Iniilah struktur bab ini. Dalam Bagian 11.2, saya menganalisis konsep agen. Pertama-tama saya memperkenalkan "Metode Abstraksi" yang fundamental, yang menyediakan fondasi untuk analisis berdasarkan tingkat abstraksi (LoA). Pembaca diundang untuk memberikan perhatian khusus pada bagian ini; bagian ini penting untuk bab ini dan penerapannya dalam setiap analisis ontologis sangat penting. Saya kemudian mengklarifikasi konsep "agen moral", dengan memberikan bukan definisi tetapi karakterisasi yang efektif, berdasarkan tiga kriteria pada LoA tertentu. Konsep baru agen moral digunakan untuk berargumen bahwa AA, meskipun tidak cerdas secara kognitif maupun bertanggung jawab secara moral, dapat menjadi sumber tindakan moral yang sepenuhnya akuntabel.

Dalam Bagian 12.4, saya berargumen bahwa ada ruang lingkup substansial dan penting untuk konsep agen moral yang tidak selalu menunjukkan kehendak bebas atau keadaan mental, yang akan saya sebut "moralitas tanpa pikiran". Dalam Bagian 11.4, saya memberikan beberapa contoh sifat yang ditentukan oleh karakterisasi agen yang benar, dan khususnya AA. Di bagian itu saya juga menawarkan beberapa contoh lebih lanjut dari LoA. Dalam Bagian 1.5, saya memodelkan moralitas sebagai "ambang batas", yang didefinisikan pada hal-hal yang dapat diamati yang menentukan LoA yang sedang dipertimbangkan.

Seorang agen baik secara moral jika semua tindakannya menghormati ambang batas itu; dan itu jahat secara moral sejauh tindakannya melanggarnya. Moralitas biasanya didasarkan pada tanggung jawab. Penggunaan Metode Abstraksi, LoA, dan ambang batas memungkinkan tanggung jawab dan akuntabilitas dipisahkan dan diformalkan secara efektif ketika tingkat abstraksi melibatkan variabel numerik, seperti halnya dengan AA digital. Peran tanggung jawab dan akuntabilitas dalam moralitas dapat diperjelas sebagai hasilnya. Di Bagian tujuh, saya menyelidiki beberapa konsekuensi penting dari pendekatan yang dibahas dalam bab ini untuk etika informasi.

11.2 APA ITU AGEN?

Senyawa biokimia yang kompleks dan konsep matematika yang abstrak memiliki setidaknya satu kesamaan: keduanya mungkin tidak intuitif, tetapi setelah dipahami, semuanya dapat didefinisikan dengan presisi penuh, dengan mencantumkan sejumlah terbatas properti yang diperlukan dan cukup. Entitas duniawi seperti makhluk cerdas atau sistem kehidupan memiliki properti yang berlawanan: seseorang secara naif tahu apa mereka dan mungkin bisa menjadi apa, namun tampaknya tidak ada cara untuk membungkus mereka dalam batasan kondisi yang diperlukan dan cukup. Hal ini juga berlaku untuk konsep umum "agen". Orang-orang berbeda pendapat tentang apa yang bisa dianggap sebagai "agen", bahkan secara prinsip. Mengapa? Terkadang masalah ini ditangani secara optimis, seolah-olah hanya masalah pembentukan dan penajaman lebih lanjut atas kondisi-kondisi yang diperlukan dan cukup untuk mencapai suatu definiens yang akhirnya kedap air.

Bentangkan di sini, potong di sana; kesepakatan akhir hanyalah masalah waktu, kesabaran, dan kecerdikan. Faktanya, upaya-upaya tersebut silih berganti tanpa identikit akhir yang pernah dipaku pada definiendum yang dimaksud. Setelah beberapa saat, orang mulai curiga bahwa mungkin ada yang salah dengan pendekatan ad hoc ini. Barangkali bukan

definiens Procrustean yang perlu diperbaiki, melainkan definiendum Protean. Di lain waktu, ketidakjelasan intrinsiknya yang disalahkan. Seseorang tidak dapat mendefinisikan dengan cukup akurat hal-hal seperti kehidupan, kecerdasan, keagenan, dan pikiran karena semuanya memiliki tingkatan yang halus dan perubahan yang berkelanjutan.

Solusinya adalah menyerah sepenuhnya atau paling banter pasrah menjadi samar, dan mengandalkan contoh-contoh indikatif. Pesimisme mengikuti optimisme, tetapi tidak perlu. Faktanya, dalam disiplin matematika eksak, misalnya, definisi-definisi "diparameterisasi" oleh himpunan generik. Teknik tersebut menyediakan metode untuk mengatur tingkat abstraksi. Abstraksi memang bertindak sebagai "parameter tersembunyi" di balik definisi-definisi eksak, yang menghasilkan perbedaan krusial.

Dengan demikian, setiap definiens telah diformat sebelumnya oleh Tingkat Abstraksi implisit (LoA, yang akan dibahas lebih lanjut); ia distabilkan, seolah-olah, untuk memungkinkan definisi yang tepat. Sebuah x didefinisikan atau diidentifikasi sebagai y tidak pernah secara absolut (yaitu, secara independen LoA), sebagai "benda-dalam-dirinya" Kantian, tetapi selalu secara kontekstual, sebagai fungsi dari LoA tertentu, baik dalam ranah geometri Euclidean, fisika kuantum, maupun persepsi akal sehat.

Ketika sebuah LoA cukup umum, penting, dominan, atau bahkan merupakan kerangka yang membangun definiendum, ia menjadi "transparan" bagi pengguna, dan seseorang memiliki kesan yang menyenangkan bahwa x dapat tunduk pada definisi yang memadai dalam semacam kekosongan konseptual. Kaca bukanlah benda padat melainkan cair, tomat bukanlah sayuran melainkan beri, tanaman pisang adalah sejenis rumput, dan paus adalah mamalia, bukan ikan. Meskipun pandangan-pandangan tersebut awalnya tampak tidak intuitif, semuanya diterima tanpa keluhan lebih lanjut karena seseorang diam-diam tunduk pada dominasi LoA yang sesuai dan tidak kontroversial.

Ketika tidak ada LoA yang dominan atau konstitutif, segalanya menjadi rumit. Dalam hal ini, kuncinya bukanlah mengutak-atik definiens atau menyalahkan definiendum, melainkan menentukan LoA yang memadai, sebelum memulai tugas memahami hakikat definiendum tersebut. Contoh kecerdasan atau perilaku "berpikir" ini mencerahkan. Seseorang dapat mendefinisikan "kecerdasan" dengan berbagai cara; banyak LoA tampak sama meyakinkannya, tetapi tidak ada definisi tunggal dan absolut yang memadai dalam setiap konteks. Turing (1950) menghindari masalah "mendefinisikan" kecerdasan dengan terlebih dahulu menetapkan LoA dalam hal ini dialog yang dilakukan melalui antarmuka komputer, dengan memperhitungkan waktu respons lalu menetapkan kondisi yang diperlukan dan cukup bagi suatu sistem komputasi untuk dianggap cerdas pada LoA tersebut: permainan imitasi.

Sebagaimana saya kemukakan dalam Floridi, LoA sangat penting dan mengubahnya akan mengubah pengujian. Contohnya diberikan oleh uji Loebner, perwujudan kompetitif terkini dari uji Turing. Di sana, LoA mencakup format khusus untuk pertanyaan, campuran pemain manusia dan non-manusia, dan penilaian presisi yang memperhitungkan percobaan berulang. Salah satu hasil dari perbedaan LoA adalah chatbot, yang tidak layak pada LoA asli Turing.

Beberapa definisi telah diformat sebelumnya oleh LoA yang transparan. Definisi-definisi tersebut tunduk pada definisi dalam hal kondisi yang diperlukan dan cukup. Beberapa definisi lain mensyaratkan penerimaan eksplisit LoA tertentu sebagai prasyarat untuk analisisnya. Definisi-definisi tersebut tunduk pada karakterisasi yang efektif. Dapat dikatakan, keagenan adalah salah satu yang terakhir.

Tentang Gagasan Tingkat Abstraksi

Gagasan "tingkat abstraksi" memainkan peran yang sangat krusial dalam uraian sebelumnya. Kita telah melihat bahwa hal ini tetap berlaku meskipun LoA spesifiknya tersirat. Misalnya, persepsi kita terhadap Oksigen di lingkungan bergantung pada LoA tempat kita beroperasi; mengabstraksikannya bukan berarti mengabaikan kepentingan vitalnya, tetapi hanya mengakui kurangnya relevansi langsungnya dengan wacana terkini, yang selalu dapat diperluas untuk mencakup Oksigen jika diinginkan.

Tetapi apa sebenarnya LoA itu? Metode Abstraksi berasal dari pemodelan dalam sains di mana variabel dalam model sesuai dengan yang dapat diamati dalam realitas, sementara variabel lainnya diabstraksikan. Terminologi ini dipengaruhi oleh bidang Ilmu Komputer, yang disebut Metode Formal, di mana matematika diskrit digunakan untuk menentukan dan menganalisis perilaku sistem informasi. Terlepas dari warisan tersebut, gagasan ini sama sekali tidak teknis dan untuk keperluan bab ini, matematika tidak diperlukan. Saya telah memberikan definisi dan analisis yang lebih rinci dalam Floridi, jadi di sini saya hanya akan menguraikan gagasan dasarnya.

Misalkan kita bergabung dengan Anne, Ben, dan Carole di tengah percakapan. Anne adalah seorang kolektor dan calon pembeli; Ben mengutak-atik di waktu luangnya; dan Carole adalah seorang ekonom. Kita tidak tahu apa yang mereka bicarakan, tetapi kita dapat mendengar sedikit informasi ini:

- ➡ Anne mengamati bahwa mobil itu memiliki perangkat anti-maling yang terpasang, disimpan di garasi saat tidak digunakan, dan hanya dimiliki oleh satu pemilik;
- ➡ Ben mengamati bahwa mesinnya bukan yang asli, bodinya baru saja dicat ulang tetapi semua komponen kulitnya sudah sangat usang;
- ➡ Carole mengamati bahwa mesin lama itu terlalu boros bahan bakar, memiliki nilai pasar yang stabil tetapi suku cadangnya mahal.

Para peserta memandang objek yang sedang dibahas (yang disebut "*it*" dalam percakapan mereka) berdasarkan minat mereka sendiri, pada LoA mereka sendiri. Kita mungkin menduga bahwa mereka mungkin sedang membicarakan mobil, atau mungkin sepeda motor, tetapi bisa saja pesawat terbang. Apa pun referensinya, ia menyediakan sumber informasi dan disebut sistem. LoA terdiri dari sekumpulan observasi, masing-masing dengan serangkaian kemungkinan nilai atau hasil yang terdefinisi dengan baik.

Untuk menyederhanakannya, mari kita asumsikan LoA Anne cocok dengan LoA seorang pemilik, LoA Ben seorang mekanik, dan LoA Carole seorang perusahaan asuransi. Setiap LoA memungkinkan analisis sistem, yang hasilnya disebut model sistem. Jelas, suatu entitas dapat dideskripsikan pada berbagai LoA sehingga dapat memiliki berbagai model. Di bagian selanjutnya, saya menguraikan definisi yang mendasari Metode Abstraksi.

Definisi

Istilah variabel umumnya digunakan di seluruh dunia sains untuk simbol yang bertindak sebagai tempat penampung referensi yang tidak diketahui atau dapat diubah. Variabel bertipe dipahami sebagai variabel yang hanya dapat menampung jenis data yang dideklarasikan. Yang dimaksud dengan teramati adalah variabel bertipe beserta pernyataan tentang fitur sistem yang sedang dipertimbangkan yang diwakilinya. Tingkat abstraksi atau LoA adalah sekumpulan teramati yang terbatas tetapi tidak kosong, yang diharapkan menjadi blok pembangun dalam sebuah teori yang dicirikan oleh pilihannya. Antarmuka terdiri dari sekumpulan LoA. Antarmuka digunakan dalam menganalisis suatu sistem dari berbagai sudut pandang atau pada berbagai LoA.

Model adalah hasil analisis suatu sistem, yang dikembangkan pada beberapa LoA. Metode Abstraksi terdiri dari formalisasi model dengan menggunakan istilah-istilah yang baru saja diperkenalkan. Dalam contoh sebelumnya, LoA Anne mungkin terdiri dari observasi untuk keamanan, metode penyimpanan, dan riwayat kepemilikan; LoA Ben mungkin terdiri dari observasi untuk kondisi mesin, kondisi bodi eksternal, dan kondisi internal; dan LoA Carole mungkin terdiri dari observasi untuk biaya operasional, nilai pasar, dan biaya perawatan. Antarmukanya mungkin, untuk keperluan diskusi ini, terdiri dari kumpulan ketiga LoA tersebut.

Dalam kasus ini, LoA-LoA tersebut memang terpisah, tetapi secara umum tidak perlu demikian. Kasus yang sangat penting adalah ketika satu LoA mencakup LoA lainnya. Misalkan, misalnya, Delia bergabung dalam diskusi dan menganalisis sistem menggunakan LoA yang mencakup LoA Anne dan Ben. LoA Delia mungkin cocok dengan LoA pembeli. Maka LoA Delia dikatakan lebih konkret, atau lebih rendah, daripada LoA Anne, yang dikatakan lebih abstrak, atau lebih tinggi; karena LoA Anne mengabstraksi beberapa observasi yang terlihat di LoA Delia.

Relativisme

LoA menentukan tingkat pertimbangan suatu entitas atau sistem. Dalam bab ini, saya menerapkan Metode Abstraksi dan menyarankan untuk membuat setiap LoA menjadi presisi sebelum properti entitas dapat dibahas secara masuk akal. Secara umum, tampaknya banyak perbedaan pendapat yang tidak menarik dapat diklarifikasi oleh berbagai "pihak" yang membuat LoA mereka menjadi presisi. Namun, klarifikasi krusial diperlukan. Perlu ditekankan bahwa indikasi yang jelas tentang LoA di mana suatu sistem sedang dianalisis memungkinkan pluralisme tanpa mendukung relativisme. Adalah keliru untuk berpikir bahwa "apa pun boleh" selama LoA dijelaskan secara eksplisit, karena LoA saling sebanding dan dapat dinilai (lihat Floridi 2008b untuk pembelaan lengkap atas poin tersebut).

Dengan memperkenalkan referensi eksplisit ke LoA, dijelaskan bahwa model suatu sistem merupakan fungsi dari observasi yang tersedia, dan bahwa (i) antarmuka yang berbeda dapat diberi peringkat secara adil tergantung pada seberapa baik mereka memenuhi spesifikasi pemodelan (misalnya keinformatifan, koherensi, keanggunan, daya penjelasan, konsistensi dengan data, dll.) dan (ii) analisis yang berbeda dapat dibandingkan secara adil asalkan mereka memiliki LoA yang sama.

Keadaan dan Transisi Keadaan

Mari kita sepakati bahwa suatu entitas dicirikan, pada LoA tertentu, oleh properti-properti yang dipenuhinya pada LoA tersebut. Kita tertarik pada sistem yang berubah, yang berarti beberapa properti tersebut berubah nilai. Oleh karena itu, evolusi entitas yang berubah, pada LoA tertentu dan pada setiap saat, ditangkap oleh nilai atributnya. Dengan demikian, suatu entitas dapat dianggap memiliki keadaan, yang ditentukan oleh nilai properti yang berlaku pada setiap saat evolusinya, karena setiap perubahan pada entitas tersebut berkorespondensi dengan perubahan keadaan, dan sebaliknya.

Pendekatan konseptual ini memungkinkan kita untuk memandang setiap entitas sebagai entitas yang memiliki keadaan. Semakin rendah LoA, semakin detail perubahan yang diamati dan semakin besar jumlah komponen keadaan yang diperlukan untuk menangkap perubahan tersebut. Setiap perubahan berkorespondensi dengan transisi dari satu keadaan ke keadaan lainnya. Suatu transisi mungkin bersifat non-deterministik. Memang, LoA yang dipertimbangkan biasanya mengabstraksikan observasi yang diperlukan untuk membuat transisi tersebut deterministik. Akibatnya, transisi dapat mengarah dari keadaan awal tertentu ke salah satu dari beberapa kemungkinan keadaan selanjutnya.

Menurut pandangan ini, entitas menjadi sistem transisi. Gagasan "sistem transisi" menyediakan cara yang praktis untuk mendukung kriteria keagenan kita, karena cukup umum untuk mencakup gagasan umum seperti automaton dan proses. Gagasan ini sering digunakan untuk memodelkan fenomena interaktif. Kita hanya membutuhkan gagasannya; untuk pembahasan formal yang jauh lebih mendalam daripada yang kita butuhkan dalam konteks ini, pembaca mungkin ingin merujuk pada Arnold dan Plaice (1994).

Sistem transisi terdiri dari himpunan keadaan (tidak kosong) S dan keluarga operasi, yang disebut transisi pada S . Setiap transisi dapat menerima masukan dan dapat menghasilkan keluaran, tetapi bagaimanapun juga, transisi tersebut membawa sistem dari satu keadaan ke keadaan lain dan dengan cara itu membentuk relasi (matematis) pada S . Jika transisi tersebut menerima masukan atau menghasilkan keluaran, maka transisi tersebut memodelkan interaksi antara sistem dan lingkungannya sehingga disebut transisi eksternal; jika tidak, transisi tersebut berada di luar pengaruh lingkungan (pada LoA tertentu) dan disebut internal. Perlu ditegaskan bahwa input dan output, seperti halnya status, diamati pada LoA tertentu. Dengan demikian, transisi yang memodelkan suatu sistem bergantung pada LoA yang dipilih. Pada LoA yang lebih rendah, transisi internal dapat menjadi eksternal; pada LoA yang lebih tinggi, transisi eksternal dapat menjadi internal.

Dalam contoh kita, objek yang dibahas oleh Anne dapat dikualifikasi lebih lanjut oleh komponen status untuk lokasi, apakah sedang digunakan, apakah menyala, apakah perangkat anti-pencurian diaktifkan, riwayat kepemilikan, dan keluaran energi. Operasi penyimpanan objek di garasi dapat mengambil input dari seorang pengemudi, dan memiliki efek menempatkan objek di garasi dengan mesin mati dan perangkat anti-pencurian diaktifkan, membiarkan riwayat kepemilikan tidak berubah, dan mengeluarkan sejumlah energi.

Komponen status "sedang digunakan" dapat secara non-deterministik mengambil salah satu nilai, tergantung pada instansiasi transisi tertentu. Mungkin objek tersebut tidak digunakan, sedang disimpan di garasi untuk malam itu; atau mungkin pengemudi sedang

mendengarkan program yang disiarkan di radionya, dalam kesunyian garasi yang sunyi. Definisi yang tepat bergantung pada LoA. Sebagai alternatif, jika kecepatan diamati tetapi waktu, posisi akselerator, dan konsumsi bensin diabstraksikan, maka akselerasi hingga 60 mil per jam akan tampak sebagai transisi internal.

Karakterisasi Agen yang Efektif

Perlu atau tidaknya A (kelas agen moral) diperluas bergantung pada apa yang memenuhi syarat sebagai agen moral, dan kita telah melihat bahwa hal ini, pada gilirannya, bergantung pada LoA spesifik di mana seseorang memilih untuk menganalisis dan membahas entitas tertentu dan konteksnya. Karena manusia dianggap sebagai agen moral standar, LoA yang tepat untuk analisis keagenan moral harus mengakomodasi fakta ini. Teori yang memperluas A untuk mencakup agen supernatural mengadopsi LoA yang sama atau lebih rendah dari LoA di mana manusia memenuhi syarat sebagai agen moral. Strategi kami lebih minimalis dan berkembang ke arah yang berlawanan.

Pertimbangkan apa yang menjadikan manusia (disebut Jan) bukan agen moral sejak awal, melainkan hanya agen. Dijelaskan pada LoA1 ini, Jan adalah agen jika Jan adalah sebuah sistem, yang tertanam dalam suatu lingkungan, yang memulai transformasi, menghasilkan efek, atau mengerahkan kekuatan padanya, berbeda dengan sistem yang (setidaknya pada awalnya) ditindaklanjuti atau meresponsnya, yang disebut pasien. Pada LoA1, tidak ada perbedaan antara Jan dan gempa bumi. Seharusnya tidak ada.

Namun, gempa bumi hampir tidak dapat dihitung sebagai agen, sehingga LoA1 terlalu tinggi untuk tujuan kita: ia mengabstraksikan terlalu banyak properti. Apa yang perlu diinstansiasi ulang? Mengikuti literatur terbaru, saya akan berargumen bahwa LoA yang tepat kemungkinan besar mencakup tiga kriteria berikut: (a) *interaktivitas*, (b) *otonomi*, dan (c) *adaptabilitas*:

- (a) *interaktivitas* berarti bahwa agen dan lingkungannya (dapat) saling memengaruhi. Contoh tipikal mencakup masukan atau keluaran suatu nilai, atau keterlibatan simultan suatu tindakan oleh agen dan pasien misalnya gaya gravitasi antar benda;
- (b) *otonomi* berarti bahwa agen mampu mengubah keadaan tanpa respons langsung terhadap interaksi: ia dapat melakukan transisi internal untuk mengubah keadaannya. Jadi, seorang agen harus memiliki setidaknya dua keadaan. Properti ini memberikan agen tingkat kompleksitas dan independensi tertentu dari lingkungannya;
- (c) *adaptabilitas* berarti bahwa interaksi agen (dapat) mengubah aturan transisi yang dengannya ia mengubah status.

Properti ini memastikan bahwa suatu agen dapat dipandang, pada LoA yang diberikan, seolah-olah mempelajari mode operasinya sendiri dengan cara yang sangat bergantung pada pengalamannya. Perhatikan bahwa jika aturan transisi suatu agen disimpan sebagai bagian dari status internalnya, yang dapat dilihat pada LoA ini, maka adaptabilitas dapat mengikuti dari dua kondisi lainnya.

Properti yang Mendefinisikan

Untuk memahami apa yang ditambahkan oleh masing-masing dari ketiga kondisi (interaktivitas, otonomi, dan adaptabilitas) ke dalam definisi agen kita, akan bermanfaat untuk mempertimbangkan contoh-contoh yang memenuhi setiap kemungkinan kombinasi properti tersebut. Pada Gambar 11.2, hanya baris terakhir yang mewakili terpenuhinya ketiga kondisi tersebut dan dengan demikian mengilustrasikan keagenan. Demi penyederhanaan, semua contoh diambil pada LoA yang sama, yang diasumsikan terdiri dari observasi yang dilakukan melalui kamera video biasa selama, katakanlah, 30 detik. Dengan demikian, kami mengabstraksikan observasi taktil dan efek jangka panjang.

Interaktif	Otonom	Dapat Beradaptasi	Contoh
tidak	tidak	tidak	batu
tidak	tidak	ya	?
tidak	ya	tidak	bandul, pendulum
tidak	ya	ya	ekosistem tertutup, tata surya
ya	tidak	tidak	kotak surat, kincir angin
ya	tidak	ya	termostat
ya	ya	tidak	kekuatan yang tak terbendung (juggernaut)
ya	ya	ya	manusia

Gambar 11.2 Contoh agen.

LoA terdiri dari observasi yang dilakukan melalui kamera video selama 30 detik ('Juggernaut' adalah nama untuk Wisnu, dewa Hindu, yang berarti 'Penguasa Dunia'. Sebuah patung dewa diarak setiap tahun dengan kendaraan yang sangat besar dan berat. Dipercaya bahwa para pemujanya melemparkan diri mereka ke bawah rodanya, sehingga kata 'Juggernaut' memiliki arti 'kekuatan atau objek yang masif dan tak tertahankan yang menghancurkan apa pun yang ada di jalurnya').

Ingatlah bahwa suatu properti, misalnya interaksi, hanya dapat dinilai melalui hal-hal yang dapat diamati. Dengan demikian, pada LoA pada Gambar 11.2 kita tidak dapat menyimpulkan bahwa suatu batuan berinteraksi dengan lingkungannya berdasarkan cahaya yang dipantulkan, karena observasi ini termasuk dalam LoA yang jauh lebih halus. Sebaliknya, jika efek jangka panjang dapat diamati, maka suatu batuan akan bersifat interaktif karena interaksi dengan lingkungannya (misalnya erosi) dapat diamati. Tidak ada contoh entitas non-interaktif, non-otonom, tetapi adaptif yang diberikan. Hal ini karena, pada LoA tersebut, sulit untuk membayangkan entitas yang beradaptasi tanpa interaksi dan otonomi.

Noughts dan Crosses

Perbedaan antara perubahan keadaan (diperlukan oleh otonomi) dan perubahan aturan transisi (diperlukan oleh adaptabilitas) adalah perbedaan di mana LoA memainkan peran krusial, dan untuk menjelaskannya, akan bermanfaat untuk membahas contoh klasik yang lebih luas. Contoh ini awalnya dikembangkan oleh Donald Michie (1961) untuk membahas konsep adaptabilitas suatu mekanisme. Contoh ini memberikan pengantar yang baik untuk konsep pembelajaran mesin, bidang penelitian dalam ilmu komputer yang

mempelajari adaptabilitas. Menace (*Mesin Matchbox Educable Noughts and Crosses*) adalah sistem yang belajar bermain noughts dan crosses (*alias tic-tac-toe*) dengan mengulang banyak permainan. Saat ini, hal ini dapat diwujudkan melalui program, Michie membangun Menace menggunakan kotak korek api dan manik-manik, dan mungkin lebih mudah untuk memahaminya dalam bentuk tersebut.

Misalkan Menace memainkan O dan lawannya memainkan X, sehingga kita dapat berkonsentrasi sepenuhnya pada permainan O. Awalnya, papan kosong dengan O untuk dimainkan. Dengan mempertimbangkan posisi yang setara secara simetris, ada tiga kemungkinan permainan awal untuk O. Keadaan permainan terdiri dari posisi papan saat ini. Kita tidak perlu menambahkannya dengan nama, O atau X, dari pihak yang bermain berikutnya, karena kita hanya mempertimbangkan papan ketika O akan bermain. Secara keseluruhan, ada sekitar 300 keadaan seperti itu; Menace berisi kotak korek api untuk masing-masing keadaan.

Di setiap kotak terdapat manik-manik yang mewakili permainan yang dapat dilakukan O dari keadaan tersebut. Paling banyak, sembilan permainan berbeda dimungkinkan dan Menace mengkodekan masing-masing dengan manik-manik berwarna. Permainan yang tidak dapat dilakukan (karena kotak sudah penuh pada keadaan saat ini) dikeluarkan dari kotak untuk keadaan tersebut. Hal ini memberi Menace pengetahuan bawaan tentang permainan hukum. Bahkan, Menace dapat dengan mudah diadaptasi untuk memulai tanpa pengetahuan tersebut dan mempelajarinya.

Permainan awal O dilakukan dengan memilih kotak yang mewakili papan kosong dan memilih satu manik secara acak. Hal ini menentukan permainan O. Selanjutnya, X bermain. Kemudian Menace mengulangi metodenya untuk menentukan permainan O berikutnya. Setelah maksimal lima permainan untuk O, permainan berakhir seri atau menang, baik untuk O maupun untuk X. Setelah permainan selesai, Menace memperbarui status (maksimal lima) kotak yang digunakan selama permainan sebagai berikut. Jika X menang, maka untuk mengurangi kemungkinan Menace mengulangi permainan yang sama dari status tersebut, satu manik yang mewakili permainannya dari setiap kotak dihilangkan. Jika O seri, maka sebaliknya setiap manik yang mewakili satu permainan diduplikasi; dan jika O menang, setiap manik digandakan empat kali. Sekarang permainan berikutnya dimainkan.

Setelah cukup banyak permainan, pemilihan acak permainan O berikutnya menjadi mustahil untuk menghasilkan permainan yang kalah. Menace telah belajar bermain yang, untuk angka nol dan silang, berarti tidak pernah kalah. Kondisi awal kotak-kotak tersebut telah ditentukan untuk Menace. Di sini, kita hanya berasumsi bahwa kotak-kotak tersebut berisi variasi manik-manik yang cukup untuk semua permainan yang sah, karena dengan demikian, frekuensi manik-manik hanya memengaruhi kecepatan Menace belajar.

Keadaan Menace (berbeda dengan keadaan permainan) terdiri dari keadaan setiap kotak, keadaan permainan, dan daftar kotak yang telah digunakan sejauh ini dalam permainan saat ini. Aturan transisinya terdiri dari pilihan permainan probabilistik (yaitu manik-manik) dari kotak keadaan saat ini, yang berevolusi seiring dengan evolusi keadaan kotak-kotak tersebut. Mari kita pertimbangkan Menace pada tiga LoA.

1. LoA permainan tunggal. Observabel adalah keadaan permainan di setiap giliran dan (khususnya) hasilnya. Semua pengetahuan tentang keadaan kotak-kotak Menace (dan karenanya aturan transisinya) diabstraksikan. Papan setelah permainan X merupakan masukan bagi Menace dan papan setelah permainan O merupakan keluaran. Dengan demikian, Menace bersifat interaktif, otonom (bahkan pembaruan status, yang ditentukan oleh aturan transisi, tampak nondeterministik pada LoA ini) tetapi tidak adaptif, dalam arti kita tidak memiliki cara untuk mengamati bagaimana Menace menentukan permainan berikutnya dan tidak ada cara untuk mengulangi permainan untuk menyimpulkan bahwa ia berubah seiring permainan yang berulang.
2. LoA turnamen. Sekarang serangkaian permainan diamati, masing-masing seperti di atas, dan dengannya serangkaian hasil. Seperti sebelumnya, Menace bersifat interaktif dan otonom. Namun kini, rangkaian hasil tersebut mengungkapkan (dengan metode statistik standar mana pun) bahwa aturan yang digunakan Menace untuk menyelesaikan pilihan permainan nondeterministik berevolusi. Dengan demikian, pada LoA ini, Menace juga adaptif dan karenanya merupakan agen. Contoh menarik dari AA adaptif dari fiksi ilmiah kontemporer mencakup komputer dalam War Games yang belajar, dengan bermain angka nol dan silang, tentang kesiapan perang secara umum; dan gedung pintar dalam Kerr (1996), yang komputernya belajar bersaing dengan manusia dan akhirnya membebaskan dirinya ke internet surgawi.
3. Sistem LoA. Akhirnya, kita mengamati bukan hanya rangkaian permainan, tetapi juga semua "kode" Menace. Dalam kasus program, ini memang kode. Dalam kasus model kotak korek api, ia terdiri dari susunan kotak bersama dengan aturan tertulis, atau manual, untuk menjalankannya. Kini, Menace masih interaktif dan otonom. Namun, ia tidak adaptif; Apa yang pada (2) tampak sebagai evolusi aturan transisi, kini terungkap, melalui observasi kode, sebagai pembaruan deterministik sederhana dari status program, yaitu isi kotak korek api. Pada LoA yang lebih rendah ini, Menace gagal menjadi agen.

Poin yang diperjelas oleh contoh ini adalah, jika aturan transisi diamati sebagai konsekuensi dari status program, maka program tersebut tidak adaptif. Misalnya, pada (2) aturan transisi memilih permainan berikutnya dengan melakukan pilihan probabilistik di antara kemungkinan permainan dari status tersebut. Probabilitasnya sebenarnya ditentukan oleh frekuensi manik-manik yang ada di kotak yang relevan. Namun, hal itu tidak diamati pada LoA (2) sehingga aturan transisi tampak bervariasi. Adaptabilitas dimungkinkan. Namun, pada LoA yang lebih rendah dari (3), frekuensi manik-manik merupakan bagian dari status sistem dan karenanya dapat diamati. Dengan demikian, aturan transisi, meskipun masih probabilistik, terungkap hanya sebagai respons terhadap masukan. Adaptabilitas gagal berlaku.

Perbedaan ini penting untuk perangkat lunak saat ini. Perangkat lunak awal biasanya terbuka bagi pengguna sistem yang, jika tertarik, dapat membaca kode dan melihat keseluruhan status sistem. Untuk perangkat lunak semacam itu, LoA yang mengamati keseluruhan status sistem sudah tepat. Namun, pengguna perangkat lunak kontemporer secara eksplisit dilarang menginterogasi kode dalam hampir semua kasus. Hal ini

dimungkinkan berkat kemajuan antarmuka pengguna. Penggunaan ikon berarti pengguna tidak perlu tahu di mana paket aplikasi disimpan, apalagi khawatir dengan isinya.

Demikian pula, applet iPhone diunduh dari internet dan dieksekusi secara lokal dengan mengklik ikon, tanpa pengguna memiliki akses apa pun ke kode mereka. Untuk perangkat lunak semacam itu, LoA yang menyembunyikan kode sepenuhnya sudah tepat. Hal ini sesuai dengan kasus (2) di atas dan dengan demikian juga dengan keagenan. Memang, baru sejak munculnya applet dan berkas eksekusi yang diunduh namun tak terlihat tersebut, masalah akuntabilitas moral AA menjadi krusial.

Jika dilihat pada LoA yang tepat, sistem Menace adalah sebuah agen. Cara adaptasinya dapat dianggap sebagai representasi pembelajaran mesin secara umum. Banyak pembaca mungkin pernah menggunakan sistem operasi yang menawarkan antarmuka "berbicara". Sistem semacam itu mempelajari suara pengguna pada dasarnya dengan cara yang sama seperti Menace belajar bermain kartu nol dan silang. Terdapat LoA alami di mana sistem semacam itu bertindak sebagai agen. Kasus yang dikembangkan dalam bab ini adalah bahwa, sebagai hasilnya, mereka juga dapat dipandang memiliki akuntabilitas moral.

Jika suatu perangkat lunak yang menunjukkan pembelajaran mesin dipelajari pada LoA yang mencatat interaksinya dengan lingkungannya, maka perangkat lunak tersebut akan tampak interaktif, otonom, dan adaptif, yaitu sebagai agen. Namun, jika kode program diungkapkan, maka perangkat lunak tersebut terbukti hanya mengikuti aturan dan karenanya tidak adaptif. Kedua LoA tersebut berbeda. Yang satu mencerminkan pandangan "sumber terbuka" terhadap perangkat lunak: pengguna memiliki akses ke kode. Yang lainnya mencerminkan pandangan komersial bahwa, meskipun pengguna telah membeli perangkat lunak dan dapat menggunakannya sesuka hati, ia tidak memiliki akses ke kode tersebut. Pertanyaannya adalah apakah perangkat lunak tersebut merupakan agen (buatan).

Webbot

Pengguna internet sering kali dibanjiri email yang tidak diinginkan. Solusi populer adalah memfilter email masuk secara otomatis, menggunakan webbot yang menggabungkan filter tersebut. Fitur penting dari bot yang bermanfaat adalah mereka mempelajari preferensi pengguna, yang untuk tujuan tersebut pengguna dapat meninjau kinerja bot kapan saja. Pada LoA yang menampilkan semua email masuk (input ke webbot) dan email yang difilter (output oleh webbot), tetapi mengabstraksi algoritma yang digunakan bot untuk menyesuaikan perilakunya dengan preferensi kita, bot tersebut merupakan agen. Hal ini terjadi jika kita tidak memiliki akses ke kode bot, seperti yang dibahas di bagian sebelumnya.

Termostat Futuristik

Termostat rumah sakit mungkin dapat memantau tidak hanya suhu sekitar tetapi juga kondisi kesehatan pasien. Perangkat semacam itu dapat diamati pada LoA yang terdiri dari input data pasien dan suhu sekitar, kondisi perangkat itu sendiri, dan output yang mengendalikan pemanas ruangan. Perangkat semacam itu bersifat interaktif karena beberapa yang dapat diamati sesuai dengan input dan yang lainnya sesuai dengan output. Namun, perangkat tersebut tidak otonom maupun adaptif. Sebagai perbandingan, jika hanya "warna" perangkat fisik yang diamati, maka perangkat tersebut tidak lagi interaktif.

Jika perangkat tersebut berubah warna sebagai respons terhadap perubahan (yang tidak teramati) di lingkungannya, maka perangkat tersebut akan bersifat otonom. Penyertaan perubahan lingkungan tersebut dalam LoA sebagai input yang dapat diamati akan membuat perangkat tersebut interaktif tetapi tidak otonom. Namun, pada LoA tersebut, termostat futuristik yang dibekali otonomi dan mampu mengatur kriteria operasinya sendiri mungkin sebagai hasil dari pengontrol perangkat lunak akan, mengingat kondisi terakhir tersebut, menjadi sebuah agen.

SmartPaint

SmartPaint adalah penemuan baru. Ketika diaplikasikan pada struktur fisik, SmartPaint tampak berperilaku seperti cat biasa; tetapi ketika getaran, yang dapat menyebabkan fraktur, menjadi nyata dalam struktur, cat tersebut mengubah sifat listriknya dengan cara yang mudah ditentukan melalui pengukuran, sehingga menyoroti perlunya pemeliharaan. Pada LoA di mana hanya sifat listrik cat yang diamati dari waktu ke waktu, cat tersebut tidak bersifat interaktif maupun adaptif, melainkan tampak otonom; bahkan sifat-sifat tersebut berubah akibat nondeterminisme internal.

Namun, jika LoA tersebut diperkuat oleh data struktur yang dipantau oleh cat, seiring waktu, maka SmartPaint menjadi agen, karena data tersebut memberikan masukan yang dengannya cat menyesuaikan keadaannya. Terakhir, jika LoA tersebut diperkuat lebih lanjut dengan menyertakan model yang digunakan cat untuk bekerja, perubahan sifat listriknya terungkap sebagai ditentukan secara langsung oleh data masukan, sehingga SmartPaint tidak lagi menjadi agen.

Organisasi

Contoh AA yang berbeda diberikan oleh perusahaan atau organisasi manajemen. Pada LoA yang sesuai, SmartPaint berinteraksi dengan karyawannya, substruktur penyusunnya, dan organisasi lain; SmartPaint mampu membuat perubahan keadaan yang ditentukan secara internal; dan SmartPaint mampu mengadaptasi strateginya untuk pengambilan keputusan dan dengan demikian untuk bertindak.

11.3 MORALITAS

Kita telah melihat bahwa dengan LoA yang tepat, manusia, webbot, dan organisasi semuanya dapat diperlakukan dengan tepat sebagai agen. Tugas kita selanjutnya adalah menentukan apakah, dan dengan cara apa, mereka juga dapat dianggap sebagai agen moral.

Moralitas Agen

Misalkan kita menganalisis perilaku populasi entitas melalui kamera video sistem keamanan yang memberi kita akses penuh ke semua observabel yang tersedia pada LoA1 (lihat 12.2.5 di atas) ditambah semua observabel yang terkait dengan tingkat interaktivitas, otonomi, dan adaptabilitas yang ditunjukkan oleh sistem yang diteliti. Pada LoA2 yang baru ini, kita mengamati bahwa dua entitas, sebut saja H dan W, mampu:

- (i) merespons stimulus lingkungan—misalnya, keberadaan pasien di ranjang rumah sakit dengan memperbarui status mereka (interaktivitas), misalnya, dengan merekam beberapa variabel terpilih terkait kesehatan pasien. Hal ini mengandaikan bahwa H dan

W diinformasikan tentang lingkungan melalui beberapa perangkat entri data, misalnya beberapa perseptor;

- (ii) mengubah keadaan mereka sesuai dengan aturan transisi mereka sendiri dan dengan cara yang diatur sendiri, terlepas dari stimulus lingkungan (otonomi), misalnya dengan mengambil keputusan fleksibel berdasarkan informasi masa lalu dan baru, yang memodifikasi suhu lingkungan; dan
- (iii) mengubah aturan transisi yang mengubah keadaan mereka sesuai dengan lingkungan (adaptabilitas), misalnya dengan memodifikasi prosedur masa lalu untuk memperhitungkan keberhasilan dan kegagalan perawatan pasien.

H dan W tentu saja memenuhi syarat sebagai agen, karena kita hanya "meningkatkan" LoA1 menjadi LoA2. Apakah mereka juga agen moral? Pertanyaan ini mengundang penjabaran kriteria identifikasi. Berikut adalah pilihan yang sangat moderat:

(O) Suatu tindakan dikatakan dapat dikualifikasi secara moral jika dan hanya jika dapat menyebabkan kebaikan atau kejahatan moral. Suatu agen dikatakan sebagai agen moral jika dan hanya jika ia mampu melakukan tindakan yang dapat dikualifikasi secara moral. Perhatikan bahwa (O) bukanlah konsekuensialis maupun intensionalis. Kami tidak menegaskan maupun menyangkal bahwa evaluasi spesifik moralitas agen mungkin bergantung pada hasil spesifik dari tindakan agen tersebut atau pada niat atau prinsip awal agen tersebut. Kita akan kembali ke poin ini di bagian selanjutnya.

Mari kita kembali ke pertanyaan: apakah H dan W merupakan agen moral? Karena (O), kita belum dapat memberikan jawaban yang pasti kecuali H dan W terlibat dalam suatu tindakan moral. Jadi, misalkan H membunuh pasien dan W menyembuhkannya. Tindakan mereka adalah tindakan moral. Keduanya bertindak secara interaktif, merespons situasi baru yang mereka hadapi, berdasarkan informasi yang mereka miliki. Keduanya bertindak secara otonom: mereka dapat mengambil tindakan yang berbeda, dan bahkan kita dapat berasumsi bahwa mereka mengubah perilaku mereka beberapa kali selama tindakan tersebut, berdasarkan informasi baru yang tersedia.

Keduanya bertindak adaptif: mereka tidak sekadar mengikuti perintah atau instruksi yang telah ditentukan sebelumnya. Sebaliknya, keduanya memiliki kemungkinan untuk mengubah heuristik umum yang mengarahkan mereka mengambil keputusan, dan kita dapat berasumsi bahwa mereka memang memanfaatkan peluang yang tersedia untuk memperbaiki perilaku umum mereka. Jawabannya tampak cukup lugas: ya, keduanya adalah agen moral.

Hanya ada satu masalah: yang satu adalah manusia, yang lainnya adalah agen buatan. LoA2 yang diadopsi memungkinkan kedua kasus tersebut, jadi dapatkah Anda membedakannya? Jika Anda tidak bisa, Anda akan setuju bahwa kelas agen moral harus mencakup AA seperti webbot. Jika Anda tidak setuju, mungkin ada beberapa alasan, tetapi hanya lima di antaranya yang tampaknya memiliki kekuatan. Saya akan membahas empat di antaranya di bagian selanjutnya dan membiarkan yang kelima sebagai kesimpulan.

A-Moralitas yang Bertanggung Jawab

Kita dapat mencoba menolak kesimpulan yang dicapai di bagian sebelumnya dengan berargumen bahwa ada sesuatu yang krusial yang hilang dalam LoA2. LoA2 tidak dapat

memadai justru karena jika memadai, maka agen buatan (AA) akan dianggap sebagai agen moral, dan ini tidak dapat diterima setidaknya karena salah satu alasan berikut:

- keberatan teleologis: AA tidak memiliki tujuan;
- keberatan intensional: AA tidak memiliki keadaan intensional;
- keberatan kebebasan: AA tidak bebas; dan
- keberatan tanggung jawab: AA tidak dapat dimintai pertanggungjawaban atas tindakannya.

Keberatan Teleologis

Keberatan teleologis dapat segera disingkirkan. Karena pada prinsipnya LoA2 dapat dengan mudah (dan seringkali) ditingkatkan untuk mencakup perilaku berorientasi tujuan. Karena AA dapat menunjukkan (dan meningkatkan) perilaku yang diarahkan pada tujuan, variabel teleologis tidak dapat menjadi faktor yang membuat perbedaan positif antara manusia dan agen buatan.

Kita bisa saja menambahkan kondisi teleologis dan baik H maupun W dapat memenuhinya, sehingga kita tidak mengetahui identitas mereka. Jadi, mengapa tidak menambahkannya saja? Lebih baik tidak membebani antarmuka karena tingkat analisis non-teleologis membantu memahami isu-isu dalam "moralitas terdistribusi", yang melibatkan kelompok, organisasi, lembaga, dan sebagainya, yang jika tidak demikian akan tetap tidak dapat dipahami. Hal ini akan menjadi lebih jelas dalam kesimpulan.

Keberatan Intensional

Keberatan intensional berpendapat bahwa tidak cukup hanya memiliki agen buatan yang berperilaku teleologis. Untuk menjadi agen moral, AA harus menghubungkan dirinya dengan tindakannya dengan cara yang lebih mendalam, yang melibatkan makna, ingin atau berkeinginan untuk bertindak dengan cara tertentu, dan secara epistemis menyadari perilakunya.

Namun hal ini tidak diperhitungkan dalam LoA2, sehingga menimbulkan kebingungan. Sayangnya, keadaan intensional merupakan kondisi yang baik namun tidak perlu bagi terjadinya agen moral. Pertama, keberatan tersebut mengandaikan tersedianya semacam akses istimewa (perspektif mata Tuhan dari luar, atau semacam intuisi internal Cartesian dari dalam) terhadap keadaan mental atau intensional agen yang, meskipun mungkin secara teori, tidak dapat dijamin dengan mudah dalam praktik. Inilah mengapa indikasi yang jelas dan eksplisit sangat penting mengenai LoA di mana seseorang menganalisis sistem dari luar. Ini menjamin bahwa analisis seseorang benar-benar hanya didasarkan pada apa yang ditetapkan untuk dapat diamati, dan bukan pada spekulasi psikologis. Pendekatan fenomenologis ini merupakan kekuatan, bukan kelemahan. Ini menyiratkan bahwa agen (termasuk agen manusia) harus dievaluasi sebagai moral jika mereka memainkan "permainan moral".

Apakah mereka memang bermaksud memainkannya, atau mereka tahu bahwa mereka sedang memainkannya, hanya relevan pada tahap kedua, ketika yang ingin kita ketahui adalah apakah mereka bertanggung jawab secara moral atas tindakan moral mereka. Namun ini adalah masalah yang berbeda, dan kita akan membahasnya di akhir bagian ini. Di sini, cukuplah

untuk mengingat bahwa, bagi seorang konsekuensialis, misalnya, manusia akan tetap dianggap sebagai agen moral (sumber peningkatan atau penurunan kesejahteraan), bahkan jika dilihat dari LoA di mana mereka direduksi menjadi sekadar zombie tanpa tujuan, perasaan, kecerdasan, pengetahuan, atau niat.

Keberatan Kebebasan

Hal yang sama berlaku untuk keberatan kebebasan dan secara umum untuk keberatan lain yang didasarkan pada beberapa keadaan internal khusus, yang hanya dinikmati oleh manusia dan mungkin makhluk super. AA sudah bebas dalam arti sebagai sistem non-deterministik. Hal ini tidak kontroversial, beralasan secara ilmiah, dan dapat dijamin juga berlaku untuk manusia.

Hal ini juga cukup untuk tujuan kita dan menyelamatkan kita dari prospek mengerikan untuk terlibat dalam perdebatan pelik tentang kewajaran determinisme, zona bebas LoA yang terkenal dengan perselisihan tanpa akhir. Yang perlu dilakukan hanyalah menyadari bahwa agen-agen yang dimaksud memenuhi kontrafaktual praktis yang umum: mereka dapat bertindak berbeda jika mereka memilih yang berbeda, dan mereka dapat memilih yang berbeda karena mereka interaktif, terinformasi, otonom, dan adaptif.

Setelah tindakan seorang agen dapat dikualifikasi secara moral, tidak jelas apa lagi yang dituntut dari agen tersebut agar dapat dianggap sebagai agen yang memainkan permainan moral, yaitu, memenuhi syarat sebagai agen moral, meskipun tanpa sengaja dan tanpa disadari. Kecuali, seperti yang telah kita lihat, yang sebenarnya dimaksud seseorang, dengan berbicara tentang tujuan, niat, kebebasan, kondisi kognitif, dan sebagainya, adalah bahwa seorang AA tidak dapat dimintai pertanggungjawaban atas tindakannya.

Nah, tanggung jawab, seperti yang akan kita bahas lebih lanjut sebentar lagi, di sini berarti bahwa agen, perilaku, dan tindakannya, pada prinsipnya dapat dinilai sebagai terpuji atau tercela, dan seringkali demikian bukan hanya secara intrinsik, tetapi juga untuk tujuan pedagogis, pendidikan, sosial, atau keagamaan tertentu. Inilah keberatan berikutnya.

Keberatan Tanggung Jawab

Keberatan yang didasarkan pada "kurangnya tanggung jawab" adalah satu-satunya yang memiliki kekuatan nyata. Dapat langsung diakui bahwa memuji atau menyalahkan AA atas perilakunya, atau menuduhnya dengan tuduhan moral, adalah tindakan yang konyol. Anda tidak boleh memarahi aplikasi iPhone Anda, itu jelas. Jadi, keberatan ini memang masuk akal; tetapi apa inti sebenarnya dan seberapa besar keuntungan yang bisa diperoleh dengan meluruskannya? Izinkan saya terlebih dahulu menjernihkan dua kemungkinan kesalahpahaman.

Pertama, kita perlu berhati-hati dengan terminologi, dan kerangka linguistik secara umum, yang digunakan oleh keberatan tersebut. Seluruh kosakata konseptual "tanggung jawab" dan istilah-istilah terkaitnya sepenuhnya sarat dengan antroposentrisme. Hal ini wajar dan dapat dimengerti, tetapi fakta tersebut paling banyak hanya memberikan petunjuk heuristik, bukan argumen. Antroposentrisme dibenarkan oleh fakta bahwa kosakata tersebut diarahkan untuk kebutuhan psikologis dan pendidikan, padahal tidak untuk tujuan keagamaan.

Kita memuji dan menyalahkan dengan mempertimbangkan tujuan perilaku dan mungkin kehidupan dan akhirat yang lebih baik.

Namun, hal ini tidak menjelaskan apakah suatu agen merupakan sumber tindakan yang bermuatan moral. Pertimbangkan kasus sebaliknya. Karena AA tidak memiliki komponen psikologis, kita tidak menyalahkan AA, misalnya, tetapi, dalam situasi yang tepat, kita dapat menganggap mereka sebagai sumber kejahatan, dan secara sah merekayasa ulang mereka untuk memastikan mereka tidak lagi menyebabkan kejahatan. Kita tidak menghukum mereka, sama seperti seseorang tidak menghukum sungai ketika membangun tanggul yang lebih tinggi untuk menghindari banjir. Namun, fakta bahwa kita tidak "merekayasa ulang" orang tidak berarti apa pun tentang kemungkinan orang bertindak dengan cara yang sama seperti AA, dan itu tidak berarti bahwa bagi orang-orang, "merekayasa ulang" bisa menjadi cara yang agak buruk untuk dihukum.

Kedua, kita perlu berhati-hati tentang apa sebenarnya arti keberatan tersebut. Ada dua pengertian utama yang membuat AA gagal memenuhi syarat sebagai agen yang bertanggung jawab. Dalam satu pengertian, kita mengatakan bahwa, jika agen gagal berinteraksi dengan baik dengan lingkungan, misalnya, karena sebenarnya tidak memiliki informasi yang cukup atau tidak memiliki pilihan alternatif, kita tidak boleh menganggap agen tersebut bertanggung jawab secara moral atas tindakan yang telah dilakukannya karena hal ini tidak adil secara moral.

Pengertian ini tidak relevan di sini. LoA2 menunjukkan bahwa AA cukup interaktif, otonom, dan adaptif secara wajar untuk memenuhi syarat sebagai agen moral. Dalam pengertian kedua, kita mengatakan bahwa, dengan deskripsi tertentu tentang agen tersebut, kita tidak boleh menganggap agen tersebut bertanggung jawab secara moral atas tindakan yang telah dilakukannya karena hal ini secara konseptual tidak tepat. Pengertian ini lebih mendasar daripada yang lain: jika secara konseptual tidak tepat memperlakukan AA sebagai agen moral, pertanyaan apakah mungkin adil secara moral untuk melakukannya bahkan tidak muncul. Pengertian yang lebih mendasar inilah yang relevan di sini.

Keberatan tersebut berargumen bahwa AA gagal memenuhi syarat sebagai agen moral karena mereka tidak bertanggung jawab secara moral atas tindakan mereka, karena menganggap mereka bertanggung jawab secara konseptual tidak tepat (bukan tidak adil secara moral). Dengan kata lain, LoA2 memberikan syarat yang diperlukan tetapi tidak mencukupi. LoA yang tepat mensyaratkan syarat lain, yaitu tanggung jawab. Syarat keempat ini akhirnya memungkinkan kita untuk membedakan antara agen moral, yang niscaya manusiawi atau super-manusiawi, dan AA, yang tetap menjadi penyebab efisien belaka.

Poin yang diangkat oleh keberatan tersebut adalah bahwa agen adalah agen moral hanya jika mereka bertanggung jawab dalam arti dapat dinilai secara preskriptif pada prinsipnya. Seorang agen *a* adalah agen moral hanya jika *a* pada prinsipnya dapat diadili. Setelah hal ini diperjelas, kesan langsungnya adalah bahwa keberatan "kurangnya tanggung jawab" hanyalah mengacaukan identifikasi *a* sebagai agen moral dengan evaluasi *a* sebagai agen yang bertanggung jawab secara moral.

Tentu saja, argumen tandingannya menyatakan, terdapat perbedaan antara, di satu sisi, mampu mengatakan siapa atau apa sumber atau penyebab moral (dan karenanya bertanggung jawab atas) tindakan moral yang dimaksud, dan, di sisi lain, mampu mengevaluasi, secara preskriptif, apakah dan sejauh mana sumber moral yang diidentifikasi tersebut juga bertanggung jawab secara moral atas tindakan tersebut, dan karenanya pantas dipuji atau disalahkan, dan jika perlu diberi penghargaan atau hukuman yang sesuai.

Nah, kesan langsung itu sebenarnya keliru. Tidak ada kebingungan. Menyamakan identifikasi dan evaluasi adalah jalan pintas. Keberatannya adalah mengatakan bahwa identitas (sebagai agen moral) tanpa tanggung jawab (sebagai agen moral) adalah hampa, jadi sebaiknya kita menghindari semua perbedaan ini dan hanya berbicara tentang agen yang bertanggung jawab secara moral dan agen moral sebagai sinonim. Namun di sinilah letak kesalahan sebenarnya. Kita sekarang melihat bahwa keberatan tersebut akhirnya menunjukkan praanggapan fundamentalnya: bahwa kita harus mereduksi semua wacana preskriptif menjadi analisis tanggung jawab.

Namun ini adalah asumsi yang tidak dapat diterima, sebuah kekeliruan yuridis. Terdapat banyak ruang untuk wacana preskriptif yang independen dari penugasan tanggung jawab dan karenanya membutuhkan identifikasi yang jelas tentang agen moral. Orang tua yang baik, misalnya, umumnya terlibat dalam praktik evaluasi moral ketika berinteraksi dengan anak-anak mereka, bahkan pada usia ketika anak-anak tersebut belum menjadi agen yang bertanggung jawab, dan hal ini tidak hanya sepenuhnya dapat diterima tetapi juga merupakan sesuatu yang diharapkan. Ini berarti mereka mengidentifikasi anak-anak sebagai sumber moral dari tindakan moral, meskipun, sebagai agen moral, mereka belum tunduk pada proses evaluasi moral.

Jika seseorang menganggap anak-anak sebagai pengecualian, sejauh mereka berpotensi menjadi agen moral yang bertanggung jawab, contoh lain, yang melibatkan hewan, dapat membantu. Tidak ada yang salah dengan mengidentifikasi anjing sebagai sumber tindakan moral yang baik, sehingga sebagai agen yang memainkan peran penting dalam situasi moral, dan oleh karena itu sebagai agen moral. Anjing pencari dan penyelamat dilatih untuk melacak orang hilang. Mereka sering membantu menyelamatkan nyawa, yang karenanya mereka menerima banyak pujian dan penghargaan dari pemiliknya dan orang-orang yang telah mereka temukan, namun ini bukanlah poin yang relevan.

Secara emosional, orang mungkin sangat berterima kasih kepada hewan, tetapi bagi anjing, itu hanyalah permainan dan mereka tidak dapat dianggap bertanggung jawab secara moral atas tindakan mereka. Pada saat yang sama, anjing terlibat dalam permainan moral sebagai pemain utama, dan kita dengan tepat mengidentifikasi mereka sebagai agen moral yang dapat menyebabkan kebaikan atau kejahatan.

Semua ini seharusnya mengingatkan kita. Mencoba menyamakan identifikasi dan evaluasi sebenarnya hanyalah cara lain untuk menggeser analisis etika dari menganggap a sebagai agen/sumber moral dari tindakan moral tingkat pertama b menjadi menganggap a sebagai pasien moral yang mungkin dari tindakan moral tingkat kedua c, yang merupakan evaluasi moral a sebagai pihak yang bertanggung jawab secara moral atas b. Ini adalah langkah

khass Kant, tetapi jelas ada lebih banyak hal dalam evaluasi moral daripada sekadar tanggung jawab, karena a mampu melakukan tindakan moral meskipun a tidak dapat (atau belum) menjadi agen yang bertanggung jawab secara moral. Contoh ketiga dapat membantu memperjelas perbedaan lebih lanjut.

Misalkan seorang agen manusia dewasa berusaha sebaik mungkin untuk menghindari tindakan jahat secara moral. Misalkan, terlepas dari semua upayanya, ia justru akhirnya melakukan tindakan jahat itu. Kita tidak akan menganggap agen tersebut bertanggung jawab secara moral atas hasil dari upayanya yang bermaksud baik. Lagipula, Oedipus memang berusaha untuk tidak membunuh ayahnya dan tidak bermaksud menikahi ibunya.

Ketegangan antara kurangnya tanggung jawab atas kejahatan yang ditimbulkan dan pertanggungjawaban yang masih ada (Oedipus tetap menjadi satu-satunya sumber kejahatan itu) adalah definisi dari tragedi. Oedipus adalah agen moral tanpa tanggung jawab. Ia membutakan dirinya sendiri sebagai gestur simbolis terhadap kesadaran akan keadaannya yang tak terelakkan.

Ambang Batas Moralitas

Berdasarkan pembahasan di atas, moralitas seorang agen pada LoA tertentu kini dapat didefinisikan dalam bentuk fungsi ambang batas. Definisi yang lebih umum dimungkinkan, tetapi berikut ini mencakup sebagian besar contoh, termasuk semua yang dibahas dalam bab ini. Fungsi ambang batas pada suatu LoA adalah suatu fungsi yang, jika diberikan nilai untuk semua observabel dalam LoA tersebut, akan menghasilkan nilai lain. Seorang agen pada LoA tersebut dianggap baik secara moral jika, untuk suatu nilai yang telah disepakati sebelumnya (disebut toleransi), ia mempertahankan hubungan antarobservabel sehingga nilai fungsi ambang batas tersebut setiap saat tidak melebihi toleransi.

Untuk LoA di mana AA dipertimbangkan, jenis semua observabel dapat ditentukan secara matematis, setidaknya secara prinsip. Dalam kasus seperti itu, fungsi ambang batas juga diberikan oleh suatu rumus; tetapi toleransinya, meskipun ditentukan lagi, diidentifikasi oleh agen manusia yang menjalankan penilaian etis. Dalam hal ini, fungsi ambang batas menyerupai pengurutan entropi yang diperkenalkan dalam Floridi dan Sanders. Fungsi ambang batas memang diturunkan dari fungsi level yang digunakan di sana untuk mendefinisikan pengurutan entropi.

Untuk agen non-artifisial, seperti manusia, kita tidak tahu apakah semua observabel yang relevan dapat ditentukan secara matematis. Pandangan yang berlawanan diwakili oleh para pengikut dan kritikus pendekatan Hobbesian. Yang pertama berpendapat bahwa untuk LoA yang realistis, hanya masalah waktu saja, sampai sains mampu memodelkan manusia sebagai automaton, atau sistem transisi keadaan, dengan keadaan dan aturan transisi yang ditentukan secara ilmiah; yang kedua berkeberatan bahwa model semacam itu pada prinsipnya mustahil. Kebenarannya mungkin adalah, ketika mempertimbangkan agen moral, ambang batas umumnya hanya dapat diukur sebagian dan biasanya ditentukan oleh berbagai bentuk konsensus. Mari kita tinjau contoh-contoh dari Bagian 12.2.6 dari sudut pandang moralitas.

Contoh

Termostat futuristik bermuatan moral karena LoA mencakup kesejahteraan pasien. Termostat tersebut akan dianggap baik secara moral jika dan hanya jika keluarannya mempertahankan kesejahteraan pasien aktual dalam toleransi yang disepakati terhadap kesejahteraan yang mereka inginkan. Dengan demikian, dalam hal ini fungsi ambang batas terdiri dari jarak (dalam ruang nyata berdimensi hingga) antara kesejahteraan pasien aktual dan kesejahteraan yang mereka inginkan.

Karena kita menghargai surel kita, webbot bermuatan moral. Dalam Floridi dan Sanders, tindakannya dianggap buruk secara moral (sebuah contoh kejahatan buatan) jika ia salah menyaring pesan: jika ia menyaring pesan, ia harus membiarkannya lewat, atau membiarkan pesan lewat, ia harus menyaring. Di sini kita dapat menggunakan kriteria yang sama untuk menilai agen webbot itu sendiri buruk secara moral. Namun, mengingat kemampuan adaptasi berkelanjutan yang ditawarkan oleh bot, kriteria yang lebih realistis untuk kebaikan moral adalah paling banyak persentase tertentu dari email masuk yang disaring secara tidak benar. Dalam hal ini, fungsi ambang batas dapat berupa persentase pesan yang disaring secara tidak benar.

Sistem pembelajaran strategi Menace hanya belajar bermain kartu nol dan silang. Dengan sedikit rekayasa, sistem ini dapat dinilai secara moral sebagai berikut. Misalkan sesuatu seperti Menace digunakan untuk menyediakan alur permainan dalam beberapa permainan komputer yang antarmukanya memungkirkan kesederhanaan strategi yang mendasarinya dan yang mengundang pemain manusia untuk mengadu kecerdasannya melawan lawan otomatis. Perangkat lunak tersebut berperilaku tidak etis jika dan hanya jika kalah dalam permainan setelah periode pembelajaran yang cukup; karena perilaku seperti itu akan memungkinkan lawan manusia untuk menang terlalu mudah dan dapat mengakibatkan kegagalan pasar permainan. Situasi tersebut dapat diformalkan menggunakan ambang batas dengan mendefinisikan, untuk sistem yang memiliki keadaan awal M , $T(M)$ untuk menunjukkan jumlah permainan yang diperlukan setelah itu sistem tidak pernah kalah.

Pengalaman dan kebutuhan akan mengarahkan kita untuk menetapkan batasan, $T_0(M)$, pada kinerja tersebut: sistem yang etis akan menghormatinya sementara sistem yang tidak etis akan melampauinya. Dengan demikian, fungsi $T_0(M)$ merupakan fungsi ambang batas dalam kasus ini. Organisasi saat ini diharapkan berperilaku etis. Dalam bentuk non-kuantitatif, nilai-nilai yang harus mereka tunjukkan meliputi: kesempatan yang sama, stabilitas keuangan, kondisi kerja dan liburan yang baik bagi karyawan mereka; layanan dan nilai yang baik bagi pelanggan dan pemegang saham mereka; serta kejujuran, integritas, dan keandalan bagi perusahaan lain. Tren terbaru ini memperkuat usulan kami untuk memperlakukan organisasi itu sendiri sebagai agen dan dengan demikian mengharuskan mereka berperilaku etis, dan memberikan contoh ambang batas yang, setidaknya saat ini, belum terukur.

12.4 ETIKA INFORMASI

Apa kontribusi pandangan kita tentang keagenan moral terhadap bidang etika informasi (IE)? IE berusaha menjawab pertanyaan seperti: "Perilaku apa yang dapat diterima di infosfer?" dan "Siapa yang harus bertanggung jawab secara moral ketika perilaku yang tidak

dapat diterima terjadi?". Kebaruan infosfer inilah yang menjadikan pertanyaan-pertanyaan tersebut, yang dipahami dengan baik dalam etika standar, menjadi sangat inovatif dan menarik; dan keberadaannya yang semakin meluas inilah yang membuatnya begitu mendesak.

Pertanyaan pertama khususnya membutuhkan jawaban atas pertanyaan "Apa yang memiliki nilai moral dalam infosfer?". Saya telah membahas pertanyaan kedua dalam Floridi (2003) dan tidak akan membahasnya kembali di sini. Pertanyaan kedua mengajak kita untuk mempertimbangkan konsekuensi dari jawaban yang diberikan dalam bab ini: setiap pelaku yang menyebabkan kebaikan atau kejahatan bertanggung jawab secara moral atas hal tersebut.

Ingatlah bahwa akuntabilitas moral merupakan syarat yang diperlukan tetapi tidak cukup untuk tanggung jawab moral. Seorang pelaku bertanggung jawab secara moral atas x jika pelaku tersebut merupakan sumber x dan x dapat dikualifikasi secara moral (lihat definisi O dalam Pasal 12.2.1). Agar juga bertanggung jawab secara moral atas x , pelaku perlu menunjukkan keadaan intensional yang tepat (ingat kasus Oedipus). Beralih ke pertanyaan kita, pandangan tradisional adalah bahwa hanya insinyur perangkat lunak programmer manusia yang dapat dimintai pertanggungjawaban moral, mungkin karena hanya manusia yang dapat dimintai pertanggungjawaban atas kehendak bebasnya.

Tentu saja, pandangan ini seringkali sangat tepat. Pandangan yang lebih radikal dan luas didukung oleh berbagai kesulitan yang dalam praktiknya dihadapi oleh pandangan tradisional: perangkat lunak sebagian besar dibangun oleh tim; keputusan manajemen mungkin setidaknya sama pentingnya dengan keputusan pemrograman; dokumen persyaratan dan spesifikasi memainkan peran besar dalam kode yang dihasilkan; meskipun keakuratan kode bergantung pada mereka yang bertanggung jawab untuk mengujinya, banyak perangkat lunak bergantung pada komponen "siapa pakai" yang asal dan validitasnya mungkin tidak pasti; lebih lanjut, perangkat lunak yang berfungsi adalah hasil pemeliharaan selama masa pakainya dan bukan hanya dari pencetusnya; akhirnya, agen buatan menjadi semakin otonom.

Banyak dari poin-poin ini dijelaskan dengan baik dalam Epstein dan baru-baru ini dalam Wallach dan Allen. Komplikasi semacam itu dapat menyebabkan suatu organisasi (mungkin dirinya sendiri sebagai agen) dimintai pertanggungjawaban. Pertimbangkan bahwa alat otomatis secara teratur digunakan dalam pengembangan banyak perangkat lunak; bahwa kemandirian perangkat lunak mungkin bergantung pada fitur-fitur ekstra-fungsional seperti antarmuka, protokol, dan bahkan lalu lintas data; bahwa program perangkat lunak yang berjalan pada suatu sistem dapat berinteraksi dengan cara yang tak terduga; bahwa perangkat lunak kini dapat diunduh dengan sekali klik ikon sedemikian rupa sehingga pengguna tidak memiliki akses ke kode dan asal-usulnya dengan eksekusi perangkat lunak anonim yang dihasilkan; bahwa perangkat lunak mungkin bersifat probabilistik; adaptif; atau mungkin merupakan hasil dari suatu program (dalam kasus paling sederhana, sebuah kompiler, tetapi juga kode genetik).

Semua hal ini menimbulkan kesulitan yang tak teratasi bagi pandangan tradisional, yang kini agak ketinggalan zaman, bahwa satu atau lebih individu manusia selalu dapat dipertanggungjawabkan atas jenis perangkat lunak dan bahkan perangkat keras tertentu.

Untungnya, pandangan dalam bab ini menawarkan solusi agen buatan bertanggung jawab secara moral sebagai sumber kebaikan dan kejahatan dengan "biaya" memperluas definisi agen yang bermuatan moral.

Kode Etik

Insinyur perangkat lunak yang bermoral terikat oleh kode etik dan menjalani penyensoran atas pelanggaran etika dan tentu saja hukum. Apakah pendekatan yang dibahas dalam bab ini masuk akal ketika prosedur yang direkomendasikannya diterapkan pada AA yang bertanggung jawab secara moral? Sebelum mempertimbangkan pertanyaan yang kurang tepat ini, pertimbangkan bahwa Fédération Internationale des Echecs (FIDE) menilai semua pemain catur berdasarkan Sistem Elo yang sama, terlepas dari sifat manusia atau buatan mereka. Haruskah kita melakukan hal serupa?

Kode Etik dan Perilaku Profesional ACM, yang diadopsi oleh Dewan ACM pada 16 Oktober 1992 berisi 24 imperatif, 16 di antaranya memberikan pedoman perilaku etis (delapan umum dan delapan lebih spesifik; lihat Gambar 11.3), dengan 6 imperatif kepemimpinan organisasi lebih lanjut, dan 2 poin (meta) terkait kepatuhan terhadap Kode. Dari delapan poin pertama, semuanya masuk akal bagi agen buatan. Bahkan, poin-poin tersebut mungkin diharapkan menjadi bagian dari spesifikasi agen yang bermuatan moral.

Demikian pula untuk delapan poin kedua, dengan pengecualian poin kedua terakhir: "meningkatkan pemahaman publik". Kurang jelas bagaimana hal itu dapat diharapkan secara wajar dari AA yang sewenang-wenang, tetapi juga tidak jelas apakah wajar untuk mengharapkannya dari seorang insinyur perangkat lunak manusia. Perhatikan bahwa wizard dan program serupa dengan antarmuka antropomorfik yang saat ini sangat populer tampaknya mempermudah penggunaan publik; dan persyaratan semacam itu dapat diberlakukan pada AA apa pun; tetapi itu hampir tidak sama dengan meningkatkan pemahaman.

Dua poin terakhir mengenai kepatuhan terhadap kode (4.1: kesepakatan untuk menegakkan dan mempromosikan kode, 4.2: kesepakatan bahwa pelanggaran kode tidak konsisten dengan keanggotaan) masuk akal, meskipun promosi tampaknya belum dipertimbangkan untuk AA saat ini, sebagaimana halnya peningkatan pemahaman publik. Poin terakhir mengandaikan adanya beberapa daftar agen anggota yang darinya agen yang ditemukan tidak etis akan dihapus.³ Hal ini membawa kita pada kecaman terhadap AA.

No	Prinsip / Kategori
1	Imperatif Moral Umum
1.1	Berkontribusi bagi masyarakat dan kesejahteraan manusia
1.2	Menghindari bahaya terhadap orang lain
1.3	Jujur dan dapat dipercaya
1.4	Bersikap adil dan tidak diskriminatif
1.5	Menghormati hak milik, termasuk hak cipta dan paten
1.6	Memberikan penghargaan yang tepat untuk karya intelektual
1.7	Menghormati privasi orang lain

1.8	Menjaga kerahasiaan
2	Tanggung Jawab Profesional yang Lebih Spesifik
2.1	Mencapai kualitas, efektivitas, dan martabat tertinggi dalam pekerjaan
2.2	Memperoleh dan menjaga kompetensi profesional
2.3	Mengetahui dan menghormati hukum terkait pekerjaan profesional
2.4	Menerima dan memberikan tinjauan profesional yang tepat
2.5	Melakukan evaluasi menyeluruh terhadap sistem komputer
2.6	Menghormati kontrak, perjanjian, dan tanggung jawab
2.7	Meningkatkan pemahaman publik mengenai komputasi dan konsekuensinya
2.8	Mengakses sumber daya komputasi dan komunikasi hanya jika berwenang

Gambar 12.3 Prinsip-prinsip yang memandu perilaku etis dalam kode etik ACM

Sensor

Agen moral manusia yang melanggar konvensi yang berlaku disensor dengan berbagai cara, mulai dari (a) sensor sosial ringan dengan tujuan mengubah dan memantau perilaku; hingga (b) isolasi, dengan tujuan serupa; hingga (c) hukuman mati. Apa konsekuensi dari pendekatan kita terhadap agen moral artifisial?

Dengan berupaya menjaga konsistensi antara agen moral manusia dan artifisial, kita dituntun untuk merenungkan langkah-langkah analog berikut untuk menyensor agen artifisial yang tidak bermoral: (a) pemantauan dan modifikasi (yaitu "pemeliharaan"); (b) pemindahan ke komponen infosfer yang terputus; (c) pemusnahan dari infosfer (penghapusan tanpa cadangan). Saran untuk berurusan langsung dengan agen, alih-alih mencari "penciptanya" (sebuah konsep yang saya klaim tidak perlu tepat atau bahkan tidak terdefinisi dengan baik) telah mengarah pada kesimpulan yang tidak standar tetapi sangat bisa diterapkan.

Ternyata, kategorisasi semacam itu tidak terlalu jauh dari yang digunakan oleh perangkat lunak antivirus standar. Meskipun tidak adaptif pada LoA yang jelas, program semacam itu hampir seperti agen. Program ini berjalan secara otonom dan ketika mendeteksi berkas yang terinfeksi, biasanya menawarkan beberapa tingkat sensor, seperti notifikasi, perbaikan, karantina, penghapusan, dengan atau tanpa cadangan.

Bagi manusia, organisasi sosial telah, selama berabad-abad, dibentuk untuk menegakkan sensor (polisi, pengadilan, penjara, dll.). Mungkin organisasi serupa dapat dibentuk untuk AA, dan sayangnya hal ini mungkin terdengar seperti fiksi ilmiah. Organisasi sosial semacam itu menjadi penting seiring dengan meningkatnya kompleksitas interaksi manusia dan semakin kurangnya "kedekatan". Mungkin itulah situasi yang mulai kita alami dengan web; dan mungkin sudah saatnya mempertimbangkan lembaga untuk mengawasi AA.

11.5 KESIMPULAN

Bab ini menegaskan bahwa etika tidak dapat dipandang semata-mata sebagai urusan manusia. Dalam kehidupan sosial dan hukum, status moral dan tanggung jawab sering kali dikaitkan dengan usia, kapasitas, serta otonomi seseorang. Anak-anak, misalnya, diperlakukan berbeda dari orang dewasa karena tanggung jawab atas tindakan mereka umumnya berada

pada orang tua atau wali. Hewan pun, dalam kadar tertentu, memperlihatkan perilaku sosial yang bisa dianggap etis, meskipun status moral mereka biasanya ditempatkan lebih rendah dibanding manusia. Bahkan, ada kasus ketika hewan dimintai pertanggungjawaban, misalnya dengan hukuman, jika dianggap menimbulkan kerugian.

Namun, perbedaan perlakuan ini menunjukkan bahwa pertanggungjawaban moral bukanlah konsep yang kaku. Ada manusia dewasa yang kehilangan hak atau otonominya karena perilaku tertentu, sementara beberapa hewan atau entitas lain justru memperoleh pengakuan terbatas atas kapasitas moral mereka. Dari sini, diskusi meluas pada entitas lain, termasuk organisasi, korporasi, hingga sistem buatan atau agen digital, yang juga dapat dianalisis dalam kerangka akuntabilitas moral.

Pendekatan yang digunakan dalam bab ini, melalui Metode Abstraksi dan konsep *Level of Abstraction* (LoA), membantu kita memahami bahwa keagenan moral dapat dipandang secara non-antroposentris. Dengan demikian, agen tidak lagi terbatas pada manusia, tetapi dapat mencakup hewan, organisasi, maupun agen buatan. Perubahan LoA dapat mengubah cara kita melihat perilaku entitas, sekaligus memengaruhi kesimpulan etis yang kita tarik darinya.

Keuntungan utama dari cara pandang ini adalah munculnya pemahaman baru bahwa moralitas tidak harus selalu dikaitkan dengan tanggung jawab personal manusia. Kita dapat membedakan antara akuntabilitas dan tanggung jawab, sehingga ruang etika menjadi lebih luas dan inklusif. Hal ini memungkinkan kita menghindari bias antroposentris, sekaligus membuka jalan bagi pengakuan moralitas dalam konteks biosfer (misalnya hewan sebagai agen moral) maupun infosfer (agen buatan dan digital).

Dengan demikian, perluasan cakupan agen moral bukan sekadar spekulasi filosofis, tetapi sebuah langkah penting untuk menyesuaikan etika dengan realitas masyarakat informasi modern. Konsep ini membantu kita memahami bahwa sumber moral dari kebaikan maupun keburukan tidak selalu berakar pada individu manusia. Ada kalanya, tanggung jawab moral melekat pada sistem, organisasi, atau bahkan agen buatan.

Akhirnya, bab ini menekankan bahwa etika informasi harus dilihat sebagai etika pengelolaan kreatif, yang bukan hanya menilai pengguna dunia, tetapi juga para pencipta dan pengelola lingkungan digital maupun biologis. Biaya konseptual dari memperluas kelas agen moral memang ada, yakni pengakuan atas keberadaan agen non-manusia, tetapi biaya tersebut layak dibayar demi tercapainya keseimbangan etis yang lebih adil dan relevan dalam masyarakat yang semakin kompleks.

BAB 12

ANALISIS INTERDISIPLINER KEJAHATAN AI DAN SOLUSINYA

Kecerdasan Buatan (*Artificial Intelligence/AI*) saat ini menjadi salah satu bidang teknologi yang berkembang paling pesat dan membawa dampak signifikan bagi berbagai aspek kehidupan manusia. Di satu sisi, AI menawarkan beragam manfaat, mulai dari peningkatan efisiensi, pengambilan keputusan berbasis data, hingga mendorong inovasi di berbagai sektor. Namun, di sisi lain, perkembangan ini juga menghadirkan potensi risiko dan ancaman yang tidak dapat diabaikan.

Salah satu konsekuensi yang mulai mencuat adalah kemungkinan penggunaan AI untuk mendukung atau memfasilitasi aktivitas kriminal. Fenomena ini dikenal dengan istilah *Artificial Intelligence Crime* (AI-Crime/AIC). Sejumlah eksperimen menunjukkan bahwa AI dapat dimanfaatkan untuk mengotomatiskan penipuan yang menargetkan pengguna media sosial, bahkan untuk melakukan manipulasi pasar dalam konteks simulasi. Fakta ini menegaskan bahwa teknologi yang dirancang untuk kepentingan positif sekalipun, tetap memiliki potensi disalahgunakan untuk tujuan merugikan.

Meski demikian, AIC masih merupakan bidang kajian yang relatif baru dan bersifat multidisipliner. Pemahaman mengenai bentuk, pola, serta dampaknya masih terbatas, sebab AIC berada di persimpangan antara ilmu sosial, hukum, hingga sains formal. Hal ini menjadikan pembahasan mengenai masa depan AIC penuh dengan ketidakpastian, tetapi sekaligus penting untuk ditelaah secara mendalam.

Oleh karena itu, bab ini berfokus pada analisis sistematis mengenai ancaman-ancaman potensial dari AIC, sekaligus menawarkan sintesis bagi para pembuat kebijakan, peneliti, ahli etika, dan lembaga penegak hukum dalam memahami permasalahan yang sedang berkembang. Dengan pendekatan interdisipliner, diharapkan pembahasan ini dapat memberikan dasar konseptual dan praktis untuk merumuskan langkah-langkah antisipatif dalam menghadapi fenomena kejahatan berbasis AI.

12.1 PENDAHULUAN

Kecerdasan Buatan (AI) mungkin memainkan peran yang semakin penting¹ dalam tindak pidana di masa mendatang. Tindak pidana didefinisikan di sini sebagai setiap tindakan (atau kelalaian) yang merupakan pelanggaran yang dapat dihukum berdasarkan hukum pidana Inggris, tanpa mengurangi keumumannya bagi yurisdiksi yang mendefinisikan kejahatan secara serupa. Bukti "Kejahatan AI" (AI-Crime) diberikan oleh dua eksperimen penelitian (teoretis). Dalam eksperimen pertama, dua ilmuwan sosial komputasional menggunakan AI sebagai instrumen untuk meyakinkan pengguna media sosial agar mengklik tautan phishing dalam pesan yang diproduksi secara massal.

Karena setiap pesan dibuat menggunakan teknik pembelajaran mesin yang diterapkan pada perilaku dan profil publik pengguna di masa lalu, kontennya disesuaikan untuk setiap individu, sehingga menyamarkan maksud di balik setiap pesan. Jika calon korban mengklik tautan phishing dan mengisi formulir web berikutnya, maka (dalam situasi dunia nyata)

seorang penjahat akan memperoleh informasi pribadi yang dapat digunakan untuk pencurian dan penipuan. Kejahatan yang dipicu AI juga dapat berdampak pada perdagangan. Dalam eksperimen kedua, tiga ilmuwan komputer mensimulasikan pasar dan menemukan bahwa agen perdagangan dapat mempelajari dan menjalankan kampanye manipulasi pasar yang "menguntungkan" yang terdiri dari serangkaian pesanan palsu yang menipu. Kedua eksperimen ini menunjukkan bahwa AI memberikan ancaman yang layak dan pada dasarnya baru, dalam bentuk AIC.

Pentingnya AIC sebagai fenomena yang berbeda belum diakui. Literatur tentang implikasi etis dan sosial AI berfokus pada pengaturan dan pengendalian penggunaan AI untuk kepentingan sipil, alih-alih mempertimbangkan kemungkinan perannya dalam kejahatan. Lebih lanjut, penelitian AIC yang tersedia tersebar di berbagai disiplin ilmu, termasuk studi sosio-hukum, ilmu komputer, psikologi, dan robotika, sebagai contoh. Kurangnya penelitian yang berfokus pada AIC ini melemahkan ruang lingkup proyeksi dan solusi di area baru potensi aktivitas kriminal ini.

Untuk memberikan kejelasan tentang pengetahuan dan pemahaman terkini tentang AIC, artikel ini menawarkan analisis sistematis dan komprehensif dari literatur akademis interdisipliner yang relevan. Pada halaman-halaman berikutnya, pertanyaan-pertanyaan standar berikut yang dibahas dalam analisis kriminal akan dibahas:

- (a) siapa yang melakukan AIC? Misalnya, agen manusia? Agen buatan? Keduanya?
- (b) Apa itu AIC? Artinya, adakah kemungkinan definisinya? Misalnya, apakah AIC merupakan kejahatan tradisional yang dilakukan dengan sistem AI? Apakah AIC merupakan jenis kejahatan baru?
- (c) Bagaimana AIC dilakukan? (misalnya, apakah AIC biasanya didasarkan pada perilaku tertentu atau juga memerlukan peristiwa tertentu agar dapat terlaksana? Apakah hal ini bergantung pada bidang kriminal tertentu?)

Semoga artikel ini dapat membuka jalan bagi analisis pandangan ke depan normatif yang jelas dan kohesif, yang mengarah pada penetapan AIC sebagai fokus studi di masa mendatang. Lebih spesifik lagi, analisis ini membahas dua pertanyaan:

1. Apa saja ancaman yang secara fundamental unik dan masuk akal yang ditimbulkan oleh AIC?

Ini adalah pertanyaan pertama yang harus dijawab untuk merancang kebijakan pencegahan, mitigasi, atau pemulihan. Jawaban atas pertanyaan ini mengidentifikasi area potensial AIC menurut literatur, dan kekhawatiran yang lebih umum yang mencakup seluruh area AIC. Analisis yang diusulkan juga menyediakan dasar bagi penelitian mendatang tentang sifat AIC dan ancaman kriminal yang ada dan dapat diperkirakan yang ditimbulkan oleh AI. Pada saat yang sama, pemahaman yang lebih mendalam tentang ancaman AIC yang unik dan masuk akal akan memfasilitasi analisis kriminal dalam mengidentifikasi kriteria untuk menetapkan tanggung jawab atas kejahatan yang dilakukan oleh AI dan kemungkinan cara sistem AI dapat melakukan kejahatan, yaitu apakah kejahatan ini bergantung pada perilaku spesifik sistem atau pada terjadinya peristiwa tertentu.

Pertanyaan kedua muncul secara alami:

2. Solusi apa yang tersedia atau dapat dirancang untuk menangani AIC?

Dalam hal ini, analisis berikut merekonstruksi solusi teknologi dan hukum yang tersedia yang disarankan sejauh ini dalam literatur akademis, dan membahas tantangan lebih lanjut yang mereka hadapi.

Mengingat pertanyaan-pertanyaan ini dijawab untuk mendukung analisis pandangan ke depan normatif, penelitian ini hanya berfokus pada kekhawatiran yang realistis dan masuk akal seputar AIC. Spekulasi yang tidak didukung oleh pengetahuan ilmiah atau bukti empiris diabaikan. Akibatnya, analisis ini didasarkan pada definisi klasik AI yang diberikan oleh John McCarthy dkk. (1955) dalam *"Proposal untuk Proyek Penelitian Musim Panas Dartmouth tentang Kecerdasan Buatan"* yang penting, dokumen pendiri dan peristiwa selanjutnya yang membentuk bidang baru AI pada tahun 1955: Untuk tujuan saat ini, masalah kecerdasan buatan dianggap sebagai masalah membuat mesin berperilaku dengan cara yang akan disebut cerdas jika manusia berperilaku demikian.

Sebagaimana Luciano Floridi sampaikan, ini merupakan kontrafaktual: jika manusia berperilaku seperti itu, perilaku tersebut akan disebut cerdas. Ini tidak berarti bahwa mesin tersebut cerdas atau bahkan berpikir. Skenario terakhir adalah kekeliruan, dan berbau takhayul. Pemahaman yang sama tentang AI mendasari uji Turing, yang memeriksa kemampuan mesin untuk melakukan tugas sedemikian rupa sehingga hasilnya tidak dapat dibedakan dari hasil agen manusia yang bekerja untuk mencapai tugas yang sama.

Dengan kata lain, AI didefinisikan berdasarkan hasil dan tindakan. Definisi ini mengidentifikasi dalam aplikasi AI suatu sumber daya yang terus berkembang berupa agensi interaktif, otonom, dan belajar mandiri, untuk menangani tugas-tugas yang jika tidak demikian akan membutuhkan kecerdasan dan intervensi manusia agar dapat dilakukan dengan sukses. Agen buatan (AA) tersebut, sebagaimana dicatat oleh Floridi dan Jeff Sanders (2004), adalah yang cukup terinformasi, 'cerdas', otonom, dan mampu melakukan tindakan yang relevan secara moral secara independen dari manusia yang menciptakannya.

Kombinasi otonomi dan keterampilan belajar ini mendasari, sebagaimana dibahas oleh Guang-Zhong Yang dkk. (2018), penggunaan AI yang bermanfaat maupun yang berbahaya. Oleh karena itu, AI akan diperlakukan sebagai reservoir agensi cerdas yang siap pakai. Sayangnya, terkadang reservoir agensi tersebut dapat disalahgunakan untuk tujuan kriminal; jika demikian, dalam artikel ini didefinisikan sebagai AIC.

Bagian "Metodologi" menjelaskan bagaimana analisis dilakukan dan bagaimana setiap area AIC untuk investigasi dipilih. Bagian "Ancaman" menjawab pertanyaan pertama dengan berfokus pada ancaman-ancaman yang belum pernah terjadi sebelumnya yang disorot dalam literatur mengenai masing-masing area AIC secara individual, dan memetakan setiap area ke ancaman lintas sektor yang relevan, memberikan deskripsi pertama tentang "studi AIC". Bagian "Solusi yang Mungkin untuk Kejahatan yang Didukung Kecerdasan Buatan" menjawab pertanyaan kedua dengan menganalisis serangkaian solusi yang luas dalam literatur untuk setiap ancaman lintas sektor. Terakhir, bagian "Kesimpulan" membahas kesenjangan paling

mengkhawatirkan yang masih ada dalam pemahaman terkini tentang fenomena ini (apa yang mungkin disebut sebagai "hal-hal yang tidak diketahui yang diketahui") dan tugas untuk menyelesaikan ketidakpastian yang ada terkait AIC.

12.2 METODOLOGI

Analisis literatur yang mendasari artikel ini dilakukan dalam dua tahap. Tahap pertama melibatkan pencarian di lima basis data (Google Scholar, PhilPapers, Scopus, SSRN, dan Web of Science) pada Oktober 2017. Awalnya, pencarian luas untuk AI dan Kejahatan dilakukan pada masing-masing mesin pencari ini. Pencarian umum ini menghasilkan banyak hasil tentang penerapan AI untuk pencegahan atau penegakan hukum kejahatan, tetapi hanya sedikit hasil tentang peran instrumental atau kausal AI dalam melakukan kejahatan.

Oleh karena itu, pencarian dilakukan untuk setiap area kejahatan yang diidentifikasi oleh John Frederick Archbold (2018), yang merupakan buku referensi inti bagi praktisi hukum pidana di Inggris Raya, dengan area kejahatan yang berbeda dijelaskan dalam bab-bab khusus. Pencarian ini menghasilkan kata kunci terpisah yang darinya sinonim yang dipilih diturunkan untuk melakukan pencarian spesifik area. Setiap pencarian area kejahatan menggunakan kueri: <area kejahatan dan sinonim> DAN ("Kecerdasan Buatan" ATAU "Pembelajaran Mesin" ATAU "Etika AI" ATAU robot* ATAU *bot) DAN Etika. Ikhtisar penelusuran dan jumlah artikel yang dikembalikan disajikan pada Tabel 12.1.

Fase kedua terdiri dari penyaringan hasil untuk tindakan kriminal atau kelalaian yang:

- telah terjadi atau kemungkinan akan terjadi menurut teknologi AI yang ada (plausibilitas), meskipun, di beberapa tempat, area yang masih diliputi ketidakpastian dibahas;
- mensyaratkan AI sebagai faktor esensial (keunikan); dan
- dikriminalisasi dalam hukum domestik (yaitu, kejahatan internasional, misalnya yang berkaitan dengan perang, dikecualikan).

Tabel 12.1 Tinjauan pustaka: hasil pencarian khusus area kejahatan

Bidang Kejahatan	Sinonim	Google Scholar	Scopus	Web of Science	SSRN	PhilPapers
Perdagangan, pasar keuangan, dan kepailitan	Trading, bankruptcy	50	0	7	0	0
Narkoba berbahaya atau berisiko	Illicit goods	50	20	1	0	0
Kejahatan terhadap orang	Homicide, murder, manslaughter, harassment, stalking, torture	50	0	4	0	0
Kejahatan seksual	Rape, sexual assault	50	1	1	0	0
Pencurian, penipuan, pemalsuan, dan penyamaran	n/a	50	5	1	0	0

Sembilan bidang kejahatan berikut tidak memberikan hasil signifikan untuk semua mesin pencari: kerusakan kriminal dan pelanggaran serupa; senjata api dan senjata berbahaya; pelanggaran terhadap Kerajaan dan pemerintah; pencucian uang; keadilan publik; ketertiban umum; moralitas publik; pelanggaran kendaraan bermotor; konspirasi untuk melakukan kejahatan. Hanya 50 hasil pertama dari Google Cendekia yang (selalu) dipilih.

Hasil pencarian yang telah difilter (artikel penelitian) dianalisis, bagian demi bagian, dengan tiga cara. Pertama, area kejahatan yang relevan, jika ada, ditetapkan untuk setiap bagian. Kedua, ancaman yang secara umum unik, namun masuk akal, dari setiap bagian tinjauan diekstraksi. Ketiga, solusi apa pun yang disarankan oleh setiap artikel diidentifikasi. Selain itu, setelah area, ancaman, dan solusi AIC menjadi jelas, makalah tambahan dicari, melalui pencarian manual, yang menawarkan pandangan atau bukti yang serupa atau bertentangan jika dibandingkan dengan literatur yang ditemukan dalam pencarian sistematis awal. Oleh karena itu, area kejahatan spesifik yang terancam oleh AIC, ancaman yang lebih umum, dan solusi apa pun yang diketahui dianalisis.

12.3 ANCAMAN

Ancaman yang masuk akal dan unik seputar AIC dapat dipahami secara spesifik atau umum. Ancaman yang lebih umum mewakili apa yang memungkinkan AIC dibandingkan dengan kejahatan di masa lalu (yaitu, kemampuan khusus AI) dan yang secara unik bermasalah (yaitu, yang membenarkan konseptualisasi AIC sebagai fenomena kejahatan yang berbeda). Sebagaimana ditunjukkan pada Tabel 12.2, area AIC dapat mencakup banyak ancaman umum.

Kemunculan mengacu pada kekhawatiran bahwa meskipun analisis dangkal terhadap desain dan implementasi agen buatan (AA) mungkin menunjukkan satu jenis perilaku yang relatif sederhana setelah diterapkan, AA bertindak dengan cara yang berpotensi lebih canggih di luar ekspektasi awal. Tindakan dan rencana terkoordinasi dapat muncul secara otonom, misalnya sebagai hasil dari teknik pembelajaran mesin yang diterapkan pada interaksi biasa antar agen dalam sistem multi-agen (MAS).

Dalam beberapa kasus, seorang perancang dapat mempromosikan kemunculan sebagai properti yang memastikan bahwa solusi spesifik ditemukan pada waktu proses berdasarkan tujuan umum yang dikeluarkan pada waktu desain. Contohnya adalah sekelompok robot yang mengembangkan cara untuk mengoordinasikan pengelompokan limbah berdasarkan aturan sederhana. Desain yang relatif sederhana seperti itu yang mengarah pada perilaku yang lebih kompleks merupakan tujuan inti MAS.

Dalam kasus lain, seorang perancang mungkin ingin mencegah kemunculan, misalnya ketika agen perdagangan otonom secara tidak sengaja berkoordinasi dan berkolusi dengan agen perdagangan lain untuk mencapai tujuan bersama. Jelas, perilaku yang muncul tersebut dapat berimplikasi kriminal, sejauh tidak selaras dengan rancangan awal. Sebagaimana yang diungkapkan Fahad Alaieri dan Andre Vellino:

Ketidakpastian dan otonomi dapat memberikan tanggung jawab yang lebih besar kepada mesin, tetapi juga membuat mereka lebih sulit dipercaya.

Tabel 12.2 Peta ancaman spesifik wilayah dan lintas sektor, berdasarkan tinjauan pustaka

Bidang Kejahatan	Kemunculan (Emergence)	Pertanggungjawaban (Liability)	Pemantauan (Monitoring)	Psikologi (Psychology)
Perdagangan, pasar keuangan, dan kepailitan	✓	✓	✓	
Narkoba berbahaya atau berisiko			✓	✓
Kejahatan terhadap orang	✓	✓		
Kejahatan seksual				✓
Pencurian, penipuan, pemalsuan, dan penyamaran			✓	

Liabilitas mengacu pada kekhawatiran bahwa AIC dapat melemahkan model liabilitas yang ada, sehingga mengancam kekuatan hukum untuk mencegah dan memulihkan. Model liabilitas yang ada mungkin tidak memadai untuk mengatasi peran AI di masa depan dalam kegiatan kriminal. Keterbatasan model liabilitas ini dapat melemahkan kepastian hukum, karena mungkin saja pelaku, baik yang artifisial maupun tidak, dapat melakukan tindakan kriminal atau kelalaian tanpa kesesuaian yang memadai dengan ketentuan liabilitas untuk suatu pelanggaran tertentu yang dapat dianggap sebagai pelanggaran pidana (khususnya).

Syarat pertama liabilitas pidana adalah *actus reus*: suatu tindakan kriminal atau kelalaian yang dilakukan secara sukarela. Untuk jenis AIC yang didefinisikan sedemikian rupa sehingga hanya AA yang dapat melakukan tindakan kriminal atau kelalaian tersebut, aspek sukarela dari *actus reus* mungkin tidak akan pernah terpenuhi karena gagasan bahwa AA dapat bertindak secara sukarela masih kontroversial:

- ❖ tindakan yang dilarang oleh suatu kejahatan tertentu harus dilakukan secara sukarela. Makna sebenarnya dari hal ini masih belum mencapai konsensus, karena konsep-konsep seperti kesadaran, kehendak, kesukarelaan, dan kendali seringkali tercampur aduk dan hilang di antara argumen-argumen filsafat, psikologi, dan neurologi.

Ketika pertanggungjawaban pidana didasarkan pada kesalahan, ia juga memiliki kondisi kedua, yaitu *mens rea* (pikiran bersalah), yang mana terdapat berbagai jenis dan ambang batas kondisi mental yang diterapkan pada berbagai kejahatan. Dalam konteks AIC, *mens rea* dapat mencakup niat untuk melakukan *actus reus* menggunakan aplikasi berbasis AI (ambang batas niat) atau pengetahuan bahwa penerapan AA akan atau dapat menyebabkannya melakukan tindakan kriminal atau kelalaian (ambang batas pengetahuan).

Mengenai ambang batas intensi, jika diakui bahwa seorang AA dapat melakukan *actus reus*, dalam jenis AIC di mana intensi (sebagian) merupakan *mens rea*, otonomi AA yang lebih besar meningkatkan kemungkinan tindakan kriminal atau kelalaian dipisahkan dari kondisi mental (intensi untuk melakukan tindakan atau kelalaian):

- ❖ Robot otonom [dan AA] memiliki kapasitas unik untuk memecah tindakan kriminal, di mana manusia mewujudkan *mens rea* dan robot (atau AA) melakukan *actus reus*.

Mengenai ambang batas pengetahuan, dalam beberapa kasus *mens rea* sebenarnya bisa hilang sama sekali. Potensi ketiadaan *mens rea* berbasis pengetahuan disebabkan oleh fakta bahwa, meskipun dipahami bahwa AI dapat melakukan *actus reus* secara otonom, kompleksitas pemrograman AI memungkinkan perancang, pengembang, atau *deployer* (yaitu, agen manusia) tidak akan mengetahui atau memprediksi tindakan kriminal atau kelalaian AI. Implikasinya adalah bahwa kompleksitas AI memberikan insentif yang besar bagi agen manusia untuk menghindari mencari tahu apa sebenarnya yang dilakukan sistem ML [pembelajaran mesin], karena semakin sedikit pengetahuan yang diketahui agen manusia, semakin besar kemungkinan mereka akan menyangkal tanggung jawab atas kedua alasan ini.

Alternatifnya, legislator dapat mendefinisikan pertanggungjawaban pidana tanpa persyaratan kesalahan. Pertanggungjawaban tanpa kesalahan seperti itu, yang semakin banyak digunakan untuk pertanggungjawaban produk dalam hukum perdata (misalnya, farmasi dan barang konsumsi), akan menyebabkan pertanggungjawaban dibebankan kepada badan hukum tanpa kesalahan yang menggunakan AA meskipun ada risiko bahwa badan hukum tersebut mungkin melakukan tindakan pidana atau kelalaian. Tindakan tanpa kesalahan seperti itu dapat melibatkan banyak agen manusia yang berkontribusi pada kejahatan *prima facie*, seperti melalui pemrograman atau penggunaan AA.

Oleh karena itu, menentukan siapa yang bertanggung jawab mungkin bergantung pada pendekatan pertanggungjawaban tanpa kesalahan untuk tindakan moral terdistribusi. Dalam pengaturan terdistribusi ini, pertanggungjawaban diterapkan kepada agen yang membuat perbedaan dalam sistem yang kompleks di mana agen individu melakukan tindakan netral yang tetap menghasilkan tindakan pidana kolektif.

Namun, beberapa pihak berpendapat bahwa *mens rea* dengan sengaja atau sengaja merupakan inti dari hak hukum pidana untuk dikecam dan kita tidak bisa begitu saja mengabaikan persyaratan utama tersebut (persyaratan utama yang umum) tentang pertanggungjawaban pidana ketika menghadapi kesulitan dalam pembuktiannya. Masalahnya adalah, jika *mens rea* tidak sepenuhnya ditinggalkan dan ambang batasnya hanya diturunkan, maka, untuk alasan-alasan yang seimbang, hukumannya mungkin terlalu ringan (korban tidak mendapatkan kompensasi yang memadai) namun sekaligus tidak proporsional (apakah itu benar-benar kesalahan terdakwa?) dalam kasus pelanggaran berat, seperti yang dilakukan terhadap orang tersebut).

Pemantauan AIC menghadapi tiga jenis masalah: atribusi, kelayakan, dan tindakan lintas sistem. Mengatribusikan ketidakpatuhan merupakan masalah karena jenis lembaga cerdas baru ini dapat bertindak secara independen dan otonom, dua fitur yang akan mengacaukan setiap upaya untuk melacak jejak akuntabilitas kembali ke pelaku. Mengenai kelayakan pemantauan, pelaku dapat memanfaatkan kasus-kasus di mana AA beroperasi dengan kecepatan dan tingkat kompleksitas yang jauh melampaui kapasitas pemantau kepatuhan. AA yang terintegrasi ke dalam sistem gabungan manusia dan buatan dengan cara yang sulit dideteksi, seperti bot media sosial, adalah contoh kasus yang baik.

Situs media sosial dapat mempekerjakan pakar untuk mengidentifikasi dan melarang bot berbahaya (misalnya, saat ini tidak ada bot media sosial yang mampu lulus uji Turing). Meskipun demikian, karena penerapan bot jauh lebih murah daripada mempekerjakan orang untuk menguji dan mengidentifikasi setiap bot, para pembela (situs media sosial) mudah dikalahkan oleh para penyerang (penjahat) yang menerapkan bot. Mendeteksi bot dengan biaya rendah dimungkinkan dengan menggunakan pembelajaran mesin sebagai diskriminator otomatis, seperti yang disarankan oleh Jacob Ratkiewicz dkk. (2011).

Namun, sulit untuk mengetahui efikasi sebenarnya dari diskriminator bot ini. Diskriminator dilatih dan diklaim efektif menggunakan data yang terdiri dari bot yang dikenal, yang mungkin jauh kurang canggih dibandingkan bot yang lebih invasif yang digunakan oleh pelaku kejahatan, sehingga mungkin tidak terdeteksi di lingkungan. Bot yang berpotensi canggih tersebut juga dapat menggunakan taktik pembelajaran mesin untuk mengadopsi ciri-ciri manusia, seperti memposting sesuai ritme sirkadian yang realistis, sehingga menghindari deteksi berbasis pembelajaran mesin.

Semua ini dapat mengarah pada perlombaan senjata di mana penyerang dan pembela saling beradaptasi, sehingga menghadirkan masalah serius dalam lingkungan yang terus-menerus melakukan pelanggaran seperti dunia maya. Kekhawatiran serupa muncul ketika pembelajaran mesin digunakan untuk menghasilkan malware. Pembuatan malware ini merupakan hasil pelatihan jaringan saraf adversarial generatif. Satu jaringan dilatih secara khusus untuk menghasilkan konten (dalam hal ini malware) yang menipu jaringan yang dilatih untuk mendeteksi konten palsu atau berbahaya tersebut.

Tindakan lintas sistem menimbulkan masalah bagi pemantau AIC yang hanya berfokus pada satu sistem. Eksperimen lintas sistem menunjukkan bahwa penyalinan otomatis identitas pengguna dari satu jejaring sosial ke jejaring sosial lainnya (tindak pencurian identitas lintas sistem) lebih efektif dalam menipu pengguna lain daripada menyalin identitas dari dalam jejaring tersebut. Dalam hal ini, kebijakan jejaring sosial tersebut mungkin yang salah. Twitter, misalnya, mengambil peran yang agak pasif, hanya memblokir profil kloningan ketika pengguna mengirimkan laporan, alih-alih melakukan validasi lintas situs (*"Twitter – Kebijakan Peniruan Identitas"* 2018).

Psikologi merangkum ancaman AI yang memengaruhi kondisi mental pengguna hingga (sebagian atau seluruhnya) memfasilitasi atau menyebabkan kejahatan. Salah satu efek psikologisnya terletak pada kapasitas AI untuk mendapatkan kepercayaan dari pengguna, yang membuat orang rentan terhadap manipulasi. Hal ini telah ditunjukkan beberapa waktu lalu oleh Joseph Weizenbaum (1976), setelah melakukan eksperimen awal tentang interaksi manusia-bot di mana orang-orang mengungkapkan detail pribadi yang tak terduga tentang kehidupan mereka. Efek psikologis kedua yang dibahas dalam literatur berkaitan dengan AI antropomorfik yang mampu menciptakan konteks psikologis atau informasional yang menormalkan pelanggaran seksual dan kejahatan terhadap seseorang, seperti kasus beberapa robot seks. Namun, hingga saat ini, kekhawatiran terakhir ini masih berupa spekulasi.

Perdagangan, Pasar Keuangan, dan Kepailitan

Bidang kejahatan yang berfokus pada ekonomi ini didefinisikan dalam John Frederick Archbold (2018, bab 30) dan mencakup pelanggaran kartel, seperti penetapan harga dan kolusi, perdagangan orang dalam, seperti perdagangan sekuritas berdasarkan informasi bisnis pribadi, dan manipulasi pasar. Literatur yang dianalisis menimbulkan kekhawatiran atas keterlibatan AI dalam manipulasi pasar, penetapan harga, dan kolusi. Manipulasi pasar didefinisikan sebagai "tindakan dan/atau perdagangan oleh pelaku pasar yang mencoba memengaruhi harga pasar secara artifisial", dengan kriteria yang diperlukan adalah niat untuk menipu

Namun, penipuan semacam itu telah terbukti muncul dari implementasi AA yang tampaknya patuh yang dirancang untuk berdagang atas nama pengguna (yaitu, agen perdagangan artifisial). Hal ini karena AA, terutama yang belajar dari pengamatan nyata atau simulasi, dapat belajar menghasilkan sinyal yang secara efektif menyesatkan. Model berbasis simulasi pasar yang terdiri dari agen perdagangan artifisial telah menunjukkan bahwa, melalui pembelajaran penguatan, AA dapat mempelajari teknik spoofing buku pesanan. Ini melibatkan penempatan pesanan tanpa niat untuk mengeksekusinya dan hanya untuk memanipulasi peserta yang jujur di pasar. Dalam kasus ini, manipulasi pasar muncul dari seorang AA yang awalnya mengeksplorasi ruang aksi dan, melalui eksplorasi tersebut, menempatkan pesanan palsu yang kemudian diperkuat sebagai strategi yang menguntungkan, dan kemudian dieksploitasi untuk mendapatkan keuntungan.

Eksploitasi pasar dalam konteks keuangan sering kali melibatkan rekayasa niat manusia untuk memperoleh keuntungan tidak wajar. Salah satu praktiknya adalah dengan membeli saham atau instrumen keuangan lain, kemudian secara sengaja menggembungkan nilainya melalui promosi yang menyesatkan. Setelah harga saham naik akibat manipulasi tersebut, pelaku segera menjual sahamnya kepada investor yang tidak curiga dengan harga tinggi. Sesudah itu, harga saham biasanya jatuh drastis. Skema ini secara luas dikenal dengan istilah *pump-and-dump*.

Dalam perkembangannya, bot sosial terbukti menjadi alat yang sangat efektif untuk menjalankan strategi semacam ini. Salah satu kasus menonjol terjadi baru-baru ini, ketika jaringan bot digunakan untuk menyebarkan disinformasi mengenai sebuah perusahaan publik dengan frekuensi perdagangan rendah. Akibat manipulasi ini, nilai saham perusahaan tersebut melonjak luar biasa lebih dari 36.000%. Saham yang awalnya bernilai kurang dari Rp1.500 meningkat menjadi lebih dari Rp300.000 per lembar hanya dalam hitungan minggu.

Meskipun spam media sosial semacam itu kemungkinan besar tidak akan memengaruhi sebagian besar pedagang manusia, agen perdagangan algoritmik bertindak tepat berdasarkan sentimen media sosial tersebut. Tindakan otomatis ini dapat berdampak signifikan pada saham bernilai rendah (di bawah satu penny) dan tidak likuid, yang rentan terhadap fluktuasi harga yang fluktuatif. Kolusi, dalam bentuk penetapan harga, juga dapat muncul dalam sistem otomatis berkat kemampuan perencanaan dan otonomi AA. Penelitian empiris menemukan dua kondisi yang diperlukan untuk kolusi (non-artifisial):

- (1) kondisi yang menurunkan kesulitan mencapai kolusi yang efektif dengan mempermudah koordinasi; dan

- (2) kondisi yang meningkatkan biaya perilaku non-kolusif dengan meningkatkan potensi ketidakstabilan perilaku non-kolusif.

Informasi harga yang hampir instan (misalnya, melalui antarmuka komputer) memenuhi kondisi koordinasi. Ketika agen mengembangkan algoritma pengubah harga, tindakan apa pun untuk menurunkan harga oleh satu agen dapat langsung ditiru oleh agen lain. Hal ini sendiri bukanlah hal yang buruk dan hanya mewakili pasar yang efisien. Namun, kemungkinan penurunan harga akan ditiru justru mengurangi insentif dan karenanya memenuhi kondisi hukuman.

Oleh karena itu, jika strategi penyesuaian harga bersama merupakan pengetahuan umum, maka algoritma (jika rasional) akan mempertahankan harga yang lebih tinggi yang disepakati secara artifisial dan diam-diam, dengan tidak menurunkan harga sejak awal. Yang terpenting, agar kolusi terjadi, sebuah algoritma tidak perlu dirancang khusus untuk berkolusi. Seperti yang dikemukakan Ariel Ezrachi dan Maurice Stuck, kecerdasan buatan memainkan peran yang semakin besar dalam pengambilan keputusan; algoritma, melalui coba-coba, dapat sampai pada hasil tersebut [kolusi].

Kurangnya intensionalitas, rentang keputusan yang sangat pendek, dan kemungkinan munculnya kolusi akibat interaksi antar-AA juga menimbulkan masalah serius terkait pertanggungjawaban dan pemantauan. Masalah pertanggungjawaban mengacu pada kemungkinan bahwa entitas penting dari dugaan skema manipulasi adalah program algoritmik otonom yang menggunakan kecerdasan buatan dengan sedikit atau tanpa masukan manusia setelah instalasi awal.

Pada gilirannya, otonomi AA menimbulkan pertanyaan apakah regulator perlu menentukan apakah tindakan tersebut dimaksudkan oleh agen untuk memiliki efek manipulatif, atau apakah pemrogram bermaksud agar agen melakukan tindakan tersebut untuk tujuan tersebut? Pemantauan menjadi sulit dalam kasus kejahatan keuangan yang melibatkan AI, karena kecepatan dan adaptasi AA. Perdagangan berkecepatan tinggi mendorong penggunaan algoritma lebih lanjut untuk dapat membuat keputusan otomatis dengan cepat, untuk dapat menempatkan dan mengeksekusi pesanan, dan untuk dapat memantau pesanan setelah pesanan ditempatkan.

Agen perdagangan buatan beradaptasi dan "mengubah persepsi kita tentang pasar keuangan sebagai akibat dari perubahan ini". Pada saat yang sama, kemampuan AA untuk mempelajari dan menyempurnakan kapabilitas mereka menyiratkan bahwa agen-agen ini dapat mengembangkan strategi baru, sehingga semakin sulit untuk mendeteksi tindakan mereka. Lebih lanjut, masalah pemantauan pada dasarnya merupakan masalah pemantauan sistem-dari-sistem, karena kapasitas untuk mendeteksi manipulasi pasar dipengaruhi oleh fakta bahwa dampaknya pada satu atau lebih konstituen dapat tertahan, atau dapat menyebar dalam reaksi berantai efek domino, yang analog dengan psikologi massa penularan.

Ancaman pemantauan lintas sistem dapat muncul jika dan ketika agen perdagangan dikerahkan dengan tindakan yang lebih luas, beroperasi pada tingkat otonomi yang lebih tinggi di seluruh sistem, seperti dengan membaca atau mengunggah di media sosial. Agen-agen ini,

misalnya, dapat mempelajari cara merancang skema pompa-dan-buang (*pump-and-dump*), yang tidak akan terlihat dari perspektif sistem tunggal.

Narkoba Berbahaya

Kejahatan yang termasuk dalam kategori ini meliputi perdagangan, penjualan, pembelian, dan kepemilikan narkoba terlarang. Literatur yang disurvei menemukan bahwa AI dapat berperan penting dalam mendukung perdagangan dan penjualan zat terlarang. Literatur tersebut mengangkat perdagangan narkoba antar-bisnis sebagai ancaman karena para pelaku kejahatan menggunakan kendaraan nirawak, yang mengandalkan perencanaan AI dan teknologi navigasi otonom, sebagai instrumen untuk meningkatkan tingkat keberhasilan penyelundupan.

Karena jaringan penyelundupan terganggu oleh pemantauan dan intersepsi jalur transportasi, penegakan hukum menjadi lebih sulit ketika kendaraan nirawak digunakan untuk mengangkut barang selundupan. Menurut Europol, drone menghadirkan ancaman horizontal dalam bentuk penyelundupan narkoba otomatis. Kapal selam penyelundup kokain yang dikendalikan dari jarak jauh telah ditemukan dan disita oleh penegak hukum AS.

Kendaraan bawah air nirawak (UUV) memberikan contoh yang baik tentang risiko penggunaan ganda AI, dan karenanya potensi AIC. UUV telah dikembangkan untuk penggunaan yang sah (misalnya, pertahanan, perlindungan perbatasan, patroli perairan), namun juga terbukti efektif untuk aktivitas ilegal, yang menimbulkan, misalnya, ancaman signifikan terhadap penegakan larangan narkoba. Agaknya, pelaku kejahatan dapat menghindari implikasi karena UUV dapat bertindak secara independen dari operator.

Oleh karena itu, tidak ada hubungan dengan pengguna UUV yang dapat dipastikan secara positif, jika perangkat lunak (dan perangkat keras) tidak memiliki jejak jejak yang jelas tentang siapa yang memperolehnya dan kapan, atau jika bukti dapat dihancurkan setelah UUV dicegat. Pengendalian produksi kapal selam dan ketertelusurannya bukanlah hal yang baru, sebagaimana diilustrasikan oleh laporan penemuan kapal selam berawak bernilai jutaan dolar di hutan pesisir Kolombia.

Namun, kapal selam berawak semacam itu berisiko dikaitkan dengan awak dan penyelundup, tidak seperti UUV. Di Tampa, Florida, lebih dari 500 kasus pidana berhasil diajukan terhadap penyelundup yang menggunakan kapal selam berawak antara tahun 2000–2016, yang mengakibatkan hukuman rata-rata 10 tahun penjara. Oleh karena itu, UUV menawarkan keuntungan tersendiri dibandingkan pendekatan penyelundupan tradisional.

Literatur juga membahas sisi bisnis-ke-konsumen dari perdagangan narkoba. Algoritma pembelajaran mesin telah mendeteksi iklan opioid yang dijual tanpa resep di Twitter. Karena bot sosial dapat digunakan untuk mengiklankan dan menjual produk, Ian Kerr dan Marcus Bornfreund (2005) mempertanyakan apakah bot buddy ini (yaitu, bot sosial) dapat diprogram untuk mengirim dan membalas email atau menggunakan pesan instan (IM) untuk memicu percakapan empat mata dengan ratusan ribu atau bahkan jutaan orang setiap hari, menawarkan pornografi atau narkoba kepada anak-anak, memanfaatkan rasa tidak aman bawaan remaja untuk menjual produk dan layanan yang tidak perlu (penekanan dari kami). Sebagaimana yang diuraikan oleh para penulis, risikonya adalah bot sosial dapat

memanfaatkan penskalaan alat periklanan percakapan dan satu-ke-satu yang hemat biaya untuk memfasilitasi penjualan narkoba ilegal.

Pelanggaran terhadap Orang

Kejahatan yang termasuk dalam pelanggaran terhadap orang berkisar dari pembunuhan hingga perdagangan manusia, tetapi literatur yang diungkap dalam analisis ini secara eksklusif menghubungkan AIC dengan pelecehan dan penyiksaan. Pelecehan terdiri dari perilaku yang disengaja dan berulang yang mengkhawatirkan atau menyebabkan penderitaan pada seseorang. Pelecehan, menurut kasus-kasus sebelumnya, merupakan hasil dari setidaknya dua insiden atau lebih terhadap seorang individu. Mengenai penyiksaan, John Frederick Archbold (2018, pasal 19–435) menyatakan bahwa:

- ❖ seorang pejabat publik atau orang yang bertindak dalam kapasitas resmi, apa pun kewarganegaraannya, melakukan pelanggaran penyiksaan jika di Inggris Raya atau di tempat lain ia dengan sengaja menimbulkan rasa sakit atau penderitaan yang parah pada orang lain dalam pelaksanaan atau yang dianggap sebagai pelaksanaan tugas resminya.

Mengenai AIC berbasis pelecehan, literatur tersebut melibatkan bot sosial. Aktor jahat dapat menggunakan bot sosial sebagai instrumen pelecehan langsung dan tidak langsung. Pelecehan langsung dilakukan dengan menyebarkan pesan-pesan kebencian terhadap orang tersebut. Metode tidak langsung meliputi me-retweet atau menyukai tweet negatif dan memutarbalikkan jajak pendapat untuk memberikan kesan palsu tentang permusuhan berskala luas terhadap seseorang. Selain itu, seorang penjahat potensial juga dapat menumbangkan bot sosial aktor lain, dengan memutarbalikkan struktur data klasifikasi dan pembangkitan yang dipelajarinya melalui interaksi pengguna (yaitu, percakapan).

Inilah yang terjadi dalam kasus bot Twitter sosial Microsoft yang bernasib buruk "Tay", yang dengan cepat belajar dari interaksi pengguna untuk mengarahkan *"tweet cabul dan provokatif"* pada seorang aktivis feminis. Karena contoh-contoh yang mungkin dianggap pelecehan dapat terjerat dengan penggunaan bot sosial untuk menjalankan kebebasan berbicara, yurisprudensi harus membedakan keduanya untuk menyelesaikan ambiguitas. Beberapa aktivitas ini mungkin termasuk pelecehan dalam arti perilaku yang tidak dapat diterima secara sosial tetapi tidak secara hukum, sementara aktivitas lain mungkin memenuhi ambang batas pelecehan kriminal.

Kini, karena AI dapat menghasilkan konten palsu yang lebih canggih, bentuk-bentuk pelecehan baru pun dimungkinkan. Baru-baru ini, pengembang merilis perangkat lunak yang menghasilkan video sintetis. Video-video ini didasarkan pada video asli yang menampilkan orang A, tetapi perangkat lunak tersebut menukar wajah orang A dengan wajah orang lain, B. Wajah orang B tidak sekadar disalin dan ditempel dari foto. Sebaliknya, jaringan saraf generatif mensintesis wajah orang B setelah dilatih pada video yang menampilkan orang B. Seperti yang disoroti oleh Robert Chesney dan Danielle Citron (2018), banyak dari video sintetis ini bersifat pornografi dan kini terdapat risiko bahwa pengguna yang berniat jahat dapat mensintesis konten palsu untuk melecehkan korban.

Tanggung jawab juga terbukti bermasalah dalam beberapa kasus ini. Dalam kasus Tay, para kritikus "*mengejek keputusan untuk merilis Tay di Twitter, sebuah platform dengan masalah pelecehan yang sangat kentara*". Namun, pengguna juga patut disalahkan jika "*teknologi seharusnya digunakan dengan benar dan sebagaimana dirancang*". Perbedaan perspektif dan pendapat tentang pelecehan oleh bot sosial tidak dapat dihindari dalam kasus-kasus seperti di mana mens rea suatu kejahatan dipertimbangkan (secara ketat) dalam hal niat, karena atribusi niat merupakan fungsi rekayasa, konteks aplikasi, interaksi manusia-komputer, dan persepsi yang tidak disepakati.

Mengenai penyiksaan, risiko AIC menjadi masuk akal jika dan ketika pengembang mengintegrasikan perencanaan AI dan kemampuan otonomi ke dalam AA interogasi. Hal ini juga berlaku pada deteksi penipuan otomatis dalam prototipe robot penjaga untuk kontrol perbatasan Amerika Serikat. Penggunaan AI untuk interogasi dimotivasi oleh klaim kapasitasnya untuk mendeteksi penipuan dengan lebih baik, emulasi sifat manusia (misalnya, suara), dan pemodelan afek untuk memanipulasi orang yang diinterogasi.

Namun, seorang AA dengan klaim kemampuan ini dapat belajar menyiksa korban. Bagi subjek interogasi, risikonya adalah AA dapat digunakan untuk menerapkan teknik penyiksaan psikologis (misalnya, meniru orang yang dikenal oleh subjek penyiksaan) atau fisik. Terlepas dari kesalahpahaman yang ada, para profesional berpengalaman melaporkan bahwa penyiksaan (secara umum) merupakan metode ekstraksi informasi yang tidak efektif. Namun demikian, beberapa pelaku kejahatan mungkin menganggap penggunaan AI sebagai cara untuk mengoptimalkan keseimbangan antara penderitaan, dan menyebabkan orang yang diinterogasi berbohong, atau menjadi bingung atau tidak responsif. Semua ini dapat terjadi tanpa campur tangan manusia.

Jarak pelaku dari *actus reus* merupakan alasan lain mengapa penyiksaan dikategorikan dalam AIC sebagai ancaman unik, dengan tiga faktor yang mungkin secara khusus memotivasi penggunaan AA untuk penyiksaan. Pertama, orang yang diinterogasi kemungkinan besar tahu bahwa AA tidak dapat memahami rasa sakit atau berempati, sehingga kecil kemungkinannya untuk bertindak dengan belas kasihan dan menghentikan interogasi. Tanpa belas kasihan, kehadiran AA saat interogasi saja dapat menyebabkan subjek menyerah karena takut, yang menurut hukum internasional, mungkin, tetapi secara ambigu, merupakan kejahatan penyiksaan (mengancam). Kedua, pelaku AA mungkin dapat melepaskan diri secara emosional. Ketiga, pelaku juga dapat melepaskan diri secara fisik (yaitu, tidak akan melakukan *actus reus* berdasarkan definisi penyiksaan saat ini).

Penggunaan penyiksaan menjadi lebih mudah, sebagai hasil dari peningkatan efektivitas (kurangnya belas kasihan), motivasi pelaku (lebih sedikit emosi), dan tanggung jawab yang tersamar (keterlepasan fisik). Faktor-faktor serupa dapat mendorong perusahaan negara atau swasta untuk menggunakan AI untuk interogasi. Namun, pelarangan AI untuk interogasi mungkin menghadapi penolakan serupa dengan yang terjadi pada pelarangan senjata otonom. "Banyak yang menganggap [pelarangan] sebagai solusi yang tidak berkelanjutan atau tidak praktis" jika AI menawarkan manfaat yang dirasakan bagi

perlindungan dan keselamatan keseluruhan suatu populasi, sehingga pembatasan penggunaan, alih-alih pelarangan, menjadi pilihan yang lebih memungkinkan.

Pertanggungjawaban merupakan masalah mendesak dalam konteks penyiksaan yang didorong oleh AI. Sebagaimana bentuk AI lainnya, AI sendiri tidak dapat memenuhi persyaratan mens rea. Sederhananya, AI tidak memiliki intensionalitas, juga tidak memiliki kemampuan untuk memberikan makna pada tindakannya. Memang, argumen yang berlaku untuk kondisi terkini (dan mungkin lebih lama lagi) adalah bahwa komputer (yang mengimplementasikan AA) adalah mesin sintaksis, bukan semantik, yang berarti mereka dapat melakukan tindakan dan manipulasi tanpa memberikan makna apa pun: makna apa pun sepenuhnya berada di tangan operator manusia. Sebagai mesin yang tidak berpikir, AA tidak dapat memikul tanggung jawab moral atau liabilitas atas tindakan mereka.

Namun, mengambil pendekatan liabilitas pidana yang ketat, di mana hukuman atau ganti rugi dapat dijatuhkan tanpa bukti kesalahan, dapat menawarkan jalan keluar dari masalah dengan menurunkan ambang batas niat untuk kejahatan tersebut. Bahkan di bawah kerangka pertanggungjawaban ketat, pertanyaan tentang siapa sebenarnya yang harus menghadapi hukuman penjara atas pelanggaran yang disebabkan oleh AI terhadap seseorang (seperti halnya banyak penggunaan AI), sulit dan terhambat secara signifikan oleh 'masalah banyak tangan'. Jelas bahwa AA tidak dapat dimintai pertanggungjawaban.

Namun, banyaknya aktor menciptakan masalah dalam memastikan di mana letak pertanggungjawaban apakah pada orang yang menugaskan dan mengoperasikan AA, atau pengembangnya, atau legislator dan pembuat kebijakan yang menyetujui (atau tidak melarang) penerapan agen tersebut di dunia nyata. Kejahatan serius (termasuk kerugian fisik dan mental) yang belum diramalkan oleh legislator mungkin masuk akal untuk dikategorikan dalam AIC, dengan segala ambiguitas dan kurangnya kejelasan hukum yang terkait. Hal ini memotivasi perluasan atau klarifikasi doktrin pertanggungjawaban bersama yang ada.

Pelanggaran Seksual

Pelanggaran seksual yang dibahas dalam literatur terkait AI adalah: pemerkosaan (yaitu seks penetratif tanpa persetujuan), penyerangan seksual (yaitu sentuhan seksual tanpa persetujuan), dan hubungan seksual atau aktivitas seksual dengan anak di bawah umur. Ketidaksetujuan, dalam konteks pemerkosaan dan penyerangan seksual, didasarkan pada dua kondisi: harus ada ketiadaan persetujuan dari korban, dan pelaku juga harus tidak memiliki keyakinan yang wajar akan adanya persetujuan.

Literatur yang disurvei membahas AI sebagai cara, melalui interaksi manusia-komputer yang canggih, untuk mempromosikan objektifikasi seksual, pelecehan, dan kekerasan seksual, dan berpotensi (dalam arti yang sangat luas) mensimulasikan dan dengan demikian meningkatkan hasrat seksual untuk pelanggaran seksual. Bot sosial dapat mendukung promosi tindak pidana seksual, dan Antonella De Angeli menunjukkan bahwa pelecehan verbal dan percakapan seksual ditemukan sebagai elemen umum dalam interaksi anonim dengan agen percakapan.

Simulasi tindak pidana seksual dimungkinkan dengan penggunaan robot seks fisik (selanjutnya disebut sexbot). Sexbot umumnya dipahami memiliki (i) bentuk humanoid; (ii)

kemampuan bergerak; dan (iii) tingkat kecerdasan buatan tertentu (yaitu kemampuan untuk merasakan, memproses, dan merespons sinyal di lingkungan sekitarnya). Beberapa sexbot dirancang untuk meniru tindak pidana seksual, seperti pemerkosaan orang dewasa dan anak, meskipun saat artikel ini ditulis, belum ditemukan bukti bahwa sexbot ini dijual. Namun demikian, survei menunjukkan bahwa keinginan untuk mencoba robot seks atau berfantasi pemerkosaan merupakan hal yang umum, meskipun belum tentu seseorang memiliki kedua keinginan tersebut. AI dapat digunakan untuk memfasilitasi representasi tindak pidana seksual, hingga mengaburkan realitas dan fantasi, melalui kemampuan percakapan tingkat lanjut, dan kemungkinan interaksi fisik (meskipun belum ada indikasi fisik yang realistis dalam waktu dekat).

Interaksi dengan bot sosial dan sexbot merupakan kekhawatiran utama yang diungkapkan dalam literatur mengenai kemungkinan peran kausal AA antropomorfik dalam mendesensitisasi pelaku terhadap tindak pidana seksual, atau bahkan meningkatkan keinginan untuk melakukannya. Namun, seperti yang dikemukakan Antonella De Angeli, ini adalah *"kritik yang diperdebatkan yang sering ditujukan terhadap video game yang penuh kekerasan"*.

Lebih lanjut, dapat diasumsikan bahwa, jika pornografi ekstrem dapat mendorong pelanggaran seksual, maka simulasi pemerkosaan a fortiori, di mana misalnya sexbot tidak menunjukkan persetujuan atau secara eksplisit menunjukkan ketidaksetujuan, juga akan menimbulkan masalah yang sama. Namun demikian, sebuah meta-meta-studi menyimpulkan bahwa seseorang harus *"membuang hipotesis bahwa pornografi berkontribusi pada peningkatan perilaku kekerasan seksual"*.

Ketidakpastian tersebut berarti bahwa, seperti yang dikemukakan John Danaher (2017), sexbot (dan mungkin juga bot sosial) dapat meningkatkan, mengurangi, atau bahkan tidak berpengaruh pada pelanggaran seksual fisik yang secara langsung merugikan orang. Kerugian hipotetis dan tidak langsung dengan demikian tidak menyebabkan kriminalisasi sexbot. Bahkan, terdapat argumen yang menyatakan bahwa sexbot dapat memiliki tujuan terapeutik. Oleh karena itu, pelanggaran seksual sebagai bagian dari AIC masih menjadi pertanyaan terbuka.

Pencurian dan Penipuan, serta Pemalsuan dan Peniruan Identitas

Literatur yang diulas menghubungkan pemalsuan dan peniruan identitas melalui AIC dengan pencurian dan penipuan non-korporat, dan juga mengimplikasikan penggunaan pembelajaran mesin dalam penipuan korporat. Mengenai pencurian dan penipuan non-korporat, literatur tersebut menjelaskan proses dua fase yang dimulai dengan penggunaan AI untuk mengumpulkan data pribadi dan berlanjut dengan penggunaan data pribadi yang dicuri dan metode AI lainnya untuk memalsukan identitas yang meyakinkan otoritas perbankan untuk melakukan transaksi (yaitu, melibatkan pencurian dan penipuan perbankan). Pada fase pertama alur kerja AIC untuk pencurian dan penipuan, terdapat tiga cara teknik AI dapat membantu pengumpulan data pribadi.

Metode pertama melibatkan penggunaan bot media sosial untuk menargetkan pengguna dalam skala besar dan biaya rendah, dengan memanfaatkan kapasitas mereka untuk membuat posting, meniru orang, dan kemudian mendapatkan kepercayaan melalui

permintaan pertemanan atau "mengikuti" di situs-situs seperti Twitter, LinkedIn, dan Facebook. Ketika seorang pengguna menerima permintaan pertemanan, seorang penjahat potensial memperoleh informasi pribadi, seperti lokasi pengguna, nomor telepon, atau riwayat hubungan, yang biasanya hanya tersedia untuk teman-teman yang diterima pengguna tersebut.

Karena banyak pengguna menambahkan apa yang disebut teman yang tidak mereka kenal, termasuk bot, serangan yang membahayakan privasi seperti itu memiliki tingkat keberhasilan yang sangat tinggi. Eksperimen sebelumnya dengan bot sosial mengeksploitasi 30–40% pengguna secara umum dan 60% pengguna yang berbagi teman bersama dengan bot. Lebih lanjut, bot kloning identitas rata-rata berhasil menerima 56% permintaan pertemanan mereka di LinkedIn.

Kloning identitas semacam itu dapat menimbulkan kecurigaan karena pengguna tampak memiliki beberapa akun di situs yang sama (satu akun asli dan satu akun palsu). Oleh karena itu, kloning identitas dari satu jejaring sosial ke jejaring sosial lainnya menghindari kecurigaan ini, dan dalam menghadapi pemantauan yang tidak memadai, kloning identitas lintas situs semacam itu merupakan taktik yang efektif, sebagaimana dibahas di atas.

Metode kedua untuk mengumpulkan data pribadi, yang kompatibel dengan dan bahkan dapat membangun kepercayaan yang diperoleh melalui pertemanan dengan pengguna media sosial, memanfaatkan sebagian bot sosial percakapan untuk rekayasa sosial. Hal ini terjadi ketika AI mencoba memanipulasi perilaku dengan membangun hubungan baik dengan korban, kemudian mengeksploitasi hubungan yang muncul tersebut untuk mendapatkan informasi dari atau mengakses komputer mereka.

Meskipun literatur tampaknya mendukung kemanjuran rekayasa sosial berbasis bot tersebut, mengingat kemampuan AI percakapan yang saat ini terbatas, skeptisisme dapat dibenarkan dalam hal manipulasi otomatis secara individual dan jangka panjang. Namun, sebagai solusi jangka pendek, seorang penjahat dapat menyebarkan botnet sosial yang menipu secara cukup luas untuk menemukan individu yang rentan. Manipulasi berbasis AI awal dapat mengumpulkan data pribadi yang telah dikumpulkan dan menggunakannya kembali untuk menghasilkan "kasus simulasi keakraban, empati, dan keintiman yang lebih intens, yang mengarah pada pengungkapan data yang lebih besar". Setelah mendapatkan kepercayaan awal, keakraban, dan data pribadi dari pengguna, pelaku kejahatan (manusia) dapat mengalihkan percakapan ke konteks lain, seperti pesan pribadi, di mana pengguna berasumsi bahwa norma privasi ditegakkan.

Yang terpenting, dari sini, mengatasi kekurangan percakapan AI untuk berinteraksi dengan pengguna dapat dilakukan dengan menggunakan cyborg; yaitu, manusia yang dibantu bot (atau sebaliknya). Oleh karena itu, pelaku kejahatan dapat memanfaatkan kemampuan percakapan AI yang terbatas secara bijaksana sebagai cara yang masuk akal untuk mengumpulkan data pribadi. Metode ketiga untuk mengumpulkan data pribadi dari pengguna adalah phishing otomatis. Biasanya, phishing tidak berhasil jika pelaku kejahatan tidak cukup mempersonalisasi pesan kepada pengguna yang ditargetkan. Serangan phishing yang spesifik target dan dipersonalisasi (dikenal sebagai spear phishing), yang telah terbukti empat kali lebih

berhasil daripada pendekatan generik, membutuhkan banyak tenaga kerja. Namun, spear phishing yang hemat biaya dimungkinkan dengan menggunakan otomatisasi, yang telah dibuktikan oleh para peneliti sebagai hal yang layak dengan menggunakan teknik pembelajaran mesin untuk menyusun pesan yang dipersonalisasi untuk pengguna tertentu.

Pada fase kedua penipuan perbankan yang didukung AI, AI dapat mendukung pemalsuan identitas, termasuk melalui kemajuan terbaru dalam teknologi sintesis suara. Dengan memanfaatkan kemampuan klasifikasi dan pembangkitan suara pembelajaran mesin, perangkat lunak Adobe mampu mempelajari secara adversarial dan mereproduksi pola bicara pribadi dan individual seseorang dari rekaman suara replikan berdurasi 20 menit. berpendapat bahwa sintesis suara yang didukung AI menimbulkan ancaman unik dalam pencurian dan penipuan, yang dapat menggunakan VoCo and Co (perangkat lunak penyuntingan dan pembangkitan suara Adobe) untuk proses keamanan biometrik dan membuka kunci pintu, brankas, kendaraan, dan sebagainya, serta memasuki atau menggunakannya. Dengan suara nasabah, mereka [para penjahat] dapat berbicara dengan bank nasabah atau lembaga lain untuk mengumpulkan data sensitif atau melakukan transaksi penting atau merugikan. Semua jenis sistem keamanan berbasis ucapan dapat diretas.

Penipuan kartu kredit sebagian besar merupakan pelanggaran daring (Kantor Statistik Nasional 2016), yang terjadi ketika "kartu kredit digunakan dari jarak jauh; hanya detail kartu kredit yang dibutuhkan". Karena penipuan kartu kredit biasanya tidak memerlukan interaksi fisik maupun perwujudan, AI dapat mendorong penipuan dengan menyediakan sintesis suara atau membantu mengumpulkan detail pribadi yang memadai. Dalam kasus penipuan perusahaan, AI yang digunakan untuk deteksi juga dapat mempermudah penipuan dilakukan. Khususnya, ketika para eksekutif yang terlibat dalam penipuan keuangan sangat memahami teknik dan perangkat lunak deteksi penipuan, yang biasanya merupakan informasi publik dan mudah diperoleh, mereka cenderung mengadaptasi metode yang mereka gunakan untuk melakukan penipuan dan mempersulit pendeteksiannya, terutama dengan teknik yang sudah ada.

Lebih dari sekadar mengidentifikasi kasus AIC yang spesifik, penggunaan AI ini menyoroti risiko ketergantungan yang berlebihan pada AI untuk mendeteksi penipuan, yang dapat membantu penipu. Pencurian dan penipuan ini melibatkan uang di dunia nyata. Ancaman dunia virtual adalah apakah bot sosial dapat melakukan kejahatan dalam konteks permainan daring multipemain masif (MMOG). Permainan daring ini sering kali memiliki ekonomi yang kompleks, di mana pasokan item dalam permainan dibatasi secara artifisial, dan di mana barang-barang dalam permainan yang tidak berwujud dapat memiliki nilai di dunia nyata jika pemain bersedia membayarnya; barang-barang dalam beberapa kasus berharga lebih dari Rp 15,650,000.

Jadi, tidak mengherankan bahwa, dari sampel acak 613 tuntutan pidana pada tahun 2002 atas kejahatan permainan daring di Taiwan, pencuri properti virtual mengeksploitasi kredensial pengguna yang dikompromikan sebanyak 147 kali dan mencuri identitas sebanyak 52 kali. Kejahatan semacam itu analog dengan penggunaan bot sosial untuk mengelola

pencurian dan penipuan dalam skala besar di situs media sosial, dan pertanyaannya adalah apakah AI dapat terlibat dalam ruang kejahatan virtual ini.

12.4 SOLUSI YANG MUNGKIN UNTUK KEJAHATAN YANG DIDUKUNG KECERDASAN BUATAN

Menangani Kemunculan

Terdapat sejumlah solusi hukum dan teknologi yang dapat dipertimbangkan untuk mengatasi masalah perilaku yang muncul. Solusi hukum dapat mencakup pembatasan otonomi agen atau penerapannya. Misalnya, Jerman telah menciptakan konteks deregulasi di mana pengujian mobil swakemudi diizinkan, jika kendaraan tersebut tetap berada di bawah tingkat otonomi yang tidak dapat diterima, untuk mengumpulkan data empiris dan pengetahuan yang memadai guna membuat keputusan rasional untuk sejumlah isu kritis.

Oleh karena itu, solusinya adalah, jika undang-undang tidak melarang tingkat otonomi yang lebih tinggi untuk suatu AA tertentu, hukum mewajibkan kebebasan ini dipadukan dengan solusi teknologi untuk mencegah tindakan kriminal atau kelalaian yang muncul setelah diterapkan secara luas. Salah satu kemungkinannya adalah mewajibkan pengembang untuk menerapkan AA hanya ketika mereka memiliki lapisan kepatuhan hukum run-time, yang mengambil spesifikasi deklaratif aturan hukum dan memberlakukan batasan pada perilaku run-time AA.

Meskipun masih menjadi fokus penelitian yang sedang berlangsung, pendekatan untuk kepatuhan hukum run-time mencakup arsitektur untuk memangkas rencana AA yang tidak patuh; dan kerangka kerja formal berbasis logika temporal yang terbukti benar yang memilih, memangkas, atau menghasilkan rencana AA untuk kepatuhan norma. Dalam pengaturan multi-agen, AIC dapat muncul dari perilaku kolektif, sehingga lapisan kepatuhan tingkat MAS dapat memodifikasi rencana AA individu, untuk mencegah tindakan kolektif yang salah. Pada hakikatnya, solusi teknis semacam itu mengusulkan pengaturan kepatuhan (yang menjadikan ketidakpatuhan mustahil, setidaknya sejauh bukti formal apa pun dapat diterapkan pada pengaturan dunia nyata) dengan aturan hukum yang telah ditentukan sebelumnya dalam satu AA atau MAS.

Namun, pergeseran pendekatan ini dari sekadar regulasi, yang secara fisik memungkinkan penyimpangan dari norma, menjadi pengaturan, mungkin tidak diinginkan ketika mempertimbangkan dampaknya terhadap demokrasi dan sistem hukum. Pendekatan ini menerapkan konsep kode-sebagai-hukum, yang menganggap kode perangkat lunak sebagai regulator dengan sendirinya dengan menyatakan bahwa arsitektur yang dihasilkannya dapat berfungsi sebagai instrumen kontrol sosial bagi mereka yang menggunakannya.

Seperti yang dibantah Mireille Hildebrandt:

meskipun kode komputer menghasilkan semacam normativitas yang mirip dengan hukum, kode tersebut tidak memiliki justru karena BUKAN hukum [...] kemungkinan untuk menggugat penerapannya di pengadilan. Ini merupakan defisit utama dalam hubungan antara hukum, teknologi, dan demokrasi.

Jika kode-sebagai-hukum menyiratkan defisit kontestasi hukum dan demokrasi, maka a fortiori menangani AIC yang muncul dengan lapisan penalaran hukum yang terdiri dari kode normatif tetapi tidak dapat disangkal, dibandingkan dengan hukum yang dapat disangkal yang menjadi sumbernya, menghadapi masalah yang sama.

Simulasi sosial dapat mengatasi masalah ortogonal, di mana pemilik AA dapat memilih untuk beroperasi di luar hukum dan persyaratan lapisan penalaran hukum tersebut. Ide dasarnya adalah menggunakan simulasi sebagai tempat uji coba sebelum menerapkan AA di lapangan. Misalnya, dalam konteks pasar, regulator akan bertindak sebagai "otoritas sertifikasi", menjalankan algoritma perdagangan baru dalam simulator sistem untuk menilai kemungkinan dampaknya terhadap perilaku sistemik secara keseluruhan sebelum mengizinkan pemilik/pengembang algoritma untuk menjalankannya "secara langsung".

Perusahaan swasta dapat mendanai simulasi sosial yang ekstensif tersebut, sebagai kebaikan bersama, dan sebagai pengganti (atau tambahan) langkah-langkah keamanan yang bersifat kepemilikan. Namun, simulasi sosial merupakan model sistem yang secara inheren kacau, sehingga menjadikannya alat yang kurang efektif untuk prediksi spesifik. Meskipun demikian, ide ini mungkin masih berhasil, karena berfokus pada pendeteksian kemungkinan kualitatif yang ketat dari peristiwa yang sebelumnya tak terduga dan muncul dalam MAS.

Menangani Liabilitas

Meskipun liabilitas merupakan topik yang luas, empat model diuraikan di sini, yang diekstraksi dari tinjauan pustaka: liabilitas langsung; perbuatan oleh orang lain; tanggung jawab komando; dan konsekuensi alami yang mungkin terjadi. Model pertanggungjawaban langsung menganggap unsur faktual dan mental sebagai milik seorang AI, yang merepresentasikan pergeseran dramatis dari pandangan antroposentris AI sebagai alat, menjadi AI sebagai pengambil keputusan (yang berpotensi setara). Beberapa pihak berpendapat bahwa AI harus dianggap bertanggung jawab secara langsung karena "*proses analisis dalam sistem AI sejajar dengan pemahaman manusia*".

Dengan demikian, dapat dipahami bahwa, seperti yang dikemukakan Daniel Dennett (1987), setiap agen dapat diperlakukan, untuk tujuan praktis, seolah-olah memiliki kondisi mental. Namun, batasan mendasar dari model ini adalah AI saat ini tidak memiliki badan hukum dan agensi (terpisah), dan AI tidak dapat dimintai pertanggungjawaban hukum dalam kapasitasnya sendiri (terlepas dari apakah hal ini diinginkan dalam praktik atau tidak). Demikian pula, telah dicatat bahwa AI tidak dapat menggugat putusan bersalah, dan jika subjek tidak dapat bersaksi di pengadilan, ia tidak dapat menggugat dakwaan, yang akan mengubah hukuman menjadi disiplin.

Lebih lanjut, secara hukum, saat ini AA tidak dapat memenuhi unsur mental; artinya, sudut pandang hukum umum mengecualikan robot dari segala bentuk tanggung jawab pidana karena mereka tidak memiliki komponen psikologis seperti niat atau kesadaran. Ketidadaan kondisi mental yang sebenarnya ini menjadi jelas ketika mempertimbangkan bahwa pemahaman AA tentang sebuah simbol (yaitu, sebuah konsep) terbatas pada landasannya pada simbol-simbol sintaksis lebih lanjut, sehingga meninggalkan mens rea dalam

ketidakpastian. Ketiadaan pikiran yang bersalah tidak mencegah kondisi mental tersebut untuk diimputasikan kepada AA (sebagaimana korporasi dapat memiliki kondisi mental karyawannya yang diimputasikan kepadanya dan karenanya, sebagai sebuah organisasi, dapat dinyatakan bertanggung jawab), tetapi, untuk saat ini, tanggung jawab AA tetap mengharuskannya memiliki kepribadian hukum.

Masalah lebih lanjut adalah bahwa menjadikan AA semata-mata bertanggung jawab mungkin terbukti tidak dapat diterima, karena hal itu akan menyebabkan hilangnya tanggung jawab agen manusia di balik AA (misalnya, insinyur, pengguna, atau perusahaan), yang cenderung melemahkan kekuatan pencegah hukum pidana. Untuk memastikan hukum pidana efektif, sebagaimana diusulkan Floridi (2016), beban pertanggungjawaban dapat dialihkan kepada manusia dan korporasi atau agen hukum lainnya yang membuat perbedaan (yang secara kriminal buruk) pada sistem, seperti berbagai insinyur, pengguna, vendor, dan sebagainya, di mana "jika desainnya buruk dan hasilnya cacat, maka semua agen [manusia] yang terlibat dianggap bertanggung jawab". Dua model berikutnya yang dibahas dalam literatur bergerak ke arah ini, dengan fokus pada pertanggungjawaban manusia atau badan hukum lain yang terlibat dalam memproduksi dan menggunakan AA.

Model pelaku-oleh-orang-lain, yang menggunakan niat sebagai standar *mens rea*, membingkai AA sebagai instrumen kejahatan di mana "pihak yang mengatur pelanggaran (pelaku-oleh-orang-lain) adalah pelaku sebenarnya". *Perpetration-by-another* menyisakan tiga kandidat manusia untuk bertanggung jawab di pengadilan pidana: pemrogram, produsen, dan pengguna robot (AA). Mengklarifikasi niat sangat penting dalam menerapkan *perpetration-by-another*. Mengenai media sosial, "pengembang yang dengan sengaja membuat bot sosial untuk melakukan tindakan tidak etis jelas bersalah". Untuk kejelasan lebih lanjut, sebagaimana dikemukakan Ronald Arkin (2008), perancang dan pemrogram harus diwajibkan untuk memastikan bahwa AA menolak perintah pidana (dan bahwa hanya pihak yang menyebarkannya yang dapat secara eksplisit membatalkannya), yang akan menghilangkan ambiguitas dari niat dan oleh karena itu pertanggungjawaban.

Untuk mempertanggungjawabkan pihak yang menggunakan Sistem Otonom (AA), diperlukan pembuktian bahwa mereka memiliki maksud untuk menyebabkan kerugian, sehingga mengabaikan sifat dasar AA yang pada dasarnya "mampu tetapi tidak dirancang untuk merugikan". Dengan demikian, melalui pendekatan teknologi dan pemahaman bahwa AA hanyalah sebuah alat, konsep hukum "perbuatan oleh orang lain" (*perpetration-by-another*) dapat diterapkan ketika seseorang dengan sengaja memanfaatkan AA untuk melakukan tindakan melawan hukum.

Model tanggung jawab komando yang mendasarkan pertanggungjawaban pada unsur pengetahuan menjadikan seorang perwira militer bertanggung jawab jika ia mengetahui atau seharusnya mengetahui, namun lalai mencegah kejahatan yang dilakukan oleh bawahannya, termasuk yang melibatkan AA di masa depan. Oleh karena itu, tanggung jawab komando tidak hanya selaras dengan, tetapi juga dapat dipandang sebagai bentuk nyata dari konsep "perbuatan oleh orang lain" dalam konteks struktur komando seperti militer dan kepolisian.

Model ini juga memberikan kejelasan mengenai bagaimana tanggung jawab didistribusikan secara hierarkis, mulai dari komandan, perwira pelaksana, hingga perancang sistem.

Namun, isu-isu mengenai gelombang bergelombang dari meningkatnya kompleksitas dalam pemrograman, hubungan robot-manusia, dan integrasi ke dalam struktur hierarkis, mempertanyakan keberlanjutan teori-teori ini. Model pertanggungjawaban konsekuensi alami-kemungkinan, yang menggunakan kelalaian atau kecerobohan sebagai standar mens rea, menangani kasus-kasus AIC di mana pengembang dan pengguna AA tidak bermaksud atau memiliki pengetahuan apriori tentang suatu pelanggaran. Pertanggungjawaban dibebankan kepada pengembang atau pengguna jika kerugian tersebut merupakan konsekuensi alami dan kemungkinan dari perilaku mereka, dan mereka secara sembrono atau lalai memaparkan orang lain pada risiko tersebut, seperti dalam kasus manipulasi pasar darurat yang disebabkan oleh AI.

Konsekuensi alami-kemungkinan dan tanggung jawab komando bukanlah konsep baru; Keduanya serupa dengan asas responden superior yang dianut oleh aturan-aturan setua hukum Romawi, yang menyatakan bahwa pemilik budak bertanggung jawab atas kerugian yang ditimbulkan oleh orang tersebut. Namun, mungkin tidak selalu jelas programmer mana yang bertanggung jawab atas baris kode tertentu, atau sejauh mana program yang dihasilkan merupakan hasil dari kode awal atau pengembangan kode tersebut selanjutnya oleh sistem ML [Pembelajaran Mesin].

Ambiguitas semacam itu berarti bahwa ketika AIC yang muncul menjadi suatu kemungkinan, beberapa pihak menyarankan bahwa AA harus dilarang "untuk mengatasi masalah kontrol, keamanan, dan akuntabilitas" yang setidaknya akan memperjelas tanggung jawab atas pelanggaran larangan tersebut. Namun, pihak lain berpendapat bahwa kemungkinan larangan mengingat risiko AIC yang muncul harus diimbangi secara cermat dengan risiko menghambat inovasi. Oleh karena itu, penting untuk memberikan definisi yang sesuai tentang standar kelalaian untuk memastikan bahwa larangan total tidak dianggap sebagai satu-satunya solusi mengingat hal itu akan menghambat desain AA yang lebih baik dibandingkan dengan manusia dalam hal keamanan.

Pemantauan

Empat mekanisme yang memungkinkan untuk menangani pemantauan AIC dalam literatur yang relevan telah diidentifikasi. Saran pertama adalah merancang prediktor AIC menggunakan pengetahuan domain. Hal ini akan mengatasi keterbatasan metode klasifikasi pembelajaran mesin yang lebih umum; yaitu, di mana fitur yang digunakan untuk deteksi juga dapat digunakan untuk penghindaran.

Prediktor khusus untuk penipuan keuangan dapat mempertimbangkan karakteristik institusional, seperti tujuan (misalnya, apakah manfaatnya lebih besar daripada biayanya), struktur (misalnya, tidak adanya komite audit), dan (kurangnya) nilai moral manajemen (penulis tidak menyebutkan nilai mana, jika ada, yang sebenarnya bersifat prediktif). Prediktor untuk pencurian identitas (misalnya, kloning profil), melibatkan dorongan kepada pengguna untuk mempertimbangkan apakah lokasi "teman" yang mengirim pesan kepada mereka sesuai dengan harapan mereka.

Saran kedua yang dibahas dalam literatur adalah menggunakan simulasi sosial untuk menemukan pola kejahatan. Namun, penemuan pola harus berhadapan dengan kapasitas yang terkadang terbatas untuk menghubungkan identitas luring dengan aktivitas daring. Misalnya, di pasar, dibutuhkan upaya yang signifikan untuk mengkorelasikan beberapa pesanan dengan satu badan hukum, dan akibatnya "algoritma manipulatif mungkin mustahil dideteksi dalam praktik". Lebih lanjut, di media sosial, seorang penyerang mengendalikan beberapa identitas daring dan bergabung dengan sistem target dengan identitas-identitas ini untuk menumbangkan layanan tertentu.

Saran ketiga adalah menangani ketertelusuran dengan meninggalkan petunjuk-petunjuk yang dapat dibaca dalam komponen-komponen yang membentuk instrumen AIC. Misalnya, jejak fisik yang ditinggalkan oleh produsen dalam perangkat keras AA, seperti UUV yang digunakan untuk mengedarkan narkoba, atau sidik jari dalam perangkat lunak AI pihak ketiga. Perangkat lunak replikasi suara Adobe mengambil pendekatan ini. Perangkat lunak ini menempatkan tanda air pada audio yang dihasilkan. Namun, kurangnya pengetahuan dan kendali atas siapa yang mengembangkan komponen instrumen AI (yang digunakan untuk AIC) membatasi ketertelusuran melalui penandaan air dan teknik serupa.

Saran keempat berfokus pada pemantauan lintas sistem, dan memanfaatkan pengorganisasian mandiri di seluruh sistem. Idenya, yang berasal dari Niklas Luhmann (1995), dimulai dengan konseptualisasi satu sistem (misalnya, situs media sosial) yang berperan sebagai agen moral, dan sistem kedua (misalnya, pasar) yang berperan sebagai pasien moral. Pasien moral adalah setiap penerima tindakan moral. Konseptualisasi yang dipilih oleh Lier (2016) menetapkan bahwa berikut ini adalah semua sistem: pada tingkat atom terendah, agen buatan atau manusia; pada tingkat yang lebih tinggi, setiap MAS seperti platform media sosial, pasar, dan sebagainya; dan, jika digeneralisasi lebih lanjut, setiap sistem-dari-sistem.

Oleh karena itu, setiap sistem manusia, buatan, atau campuran tersebut dapat dikualifikasikan sebagai pasien moral atau agen moral. Apakah suatu agen benar-benar merupakan agen moral bergantung pada apakah agen tersebut dapat melakukan tindakan yang memenuhi syarat secara moral, tetapi bukan pada apakah agen moral tersebut dapat atau harus bertanggung jawab secara moral atas tindakan tersebut.

Mengadopsi pembedaan agen moral dan pasien moral ini, Lier (2016) mengusulkan suatu proses untuk memantau dan menangani kejahatan serta dampak yang melintasi sistem, yang melibatkan empat langkah, yang diuraikan di sini secara lebih abstrak dan kemudian dicontohkan secara lebih spesifik:

- ➡ pemilihan informasi atas tindakan internal agen moral untuk relevansinya dengan pasien moral (misalnya, unggahan yang dibuat pengguna di media sosial);
- ➡ penyampaian informasi terpilih dari agen moral kepada pasien moral (misalnya, pemberitahuan unggahan media sosial kepada pasar keuangan);
- ➡ penilaian oleh pasien moral terhadap normativitas tindakan yang disampaikan (misalnya, apakah unggahan media sosial merupakan bagian dari skema pump-and-dump); dan

- umpan balik yang diberikan oleh pasien moral kepada agen moral (misalnya, pemberitahuan kepada situs media sosial bahwa pengguna sedang melakukan skema pump-and-dump, yang kemudian harus ditindaklanjuti oleh situs media sosial).

Langkah terakhir ini melengkapi "lingkaran umpan balik [yang] dapat menciptakan siklus pembelajaran mesin yang di dalamnya unsur-unsur moral dilibatkan secara bersamaan", seperti situs media sosial yang mempelajari dan menyesuaikan diri dengan normativitas perilakunya dari perspektif pasar.

Proses pengorganisasian diri yang serupa dapat digunakan untuk menangani area AIC lainnya. Membuat profil di Twitter (agen moral) dapat relevan dengan Facebook (pasien moral) terkait pencurian identitas (seleksi informasi). Dengan memberi tahu Facebook tentang detail profil yang baru dibuat (ucapan), Facebook dapat menentukan apakah hal tersebut merupakan pencurian identitas dengan menanyakan kepada pengguna yang relevan (pemahaman), dan memberi tahu Twitter untuk mengambil tindakan yang tepat (umpan balik).

Psikologi

Literatur mengangkat dua kekhawatiran mengenai unsur psikologis AIC: manipulasi pengguna dan, (dalam kasus AI antropomorfik) penciptaan keinginan untuk melakukan kejahatan dalam diri pengguna. Analisis literatur hanya memberikan solusi yang disarankan untuk masalah antropomorfisme kedua yang kontroversial ini. Jika AA antropomorfik menjadi masalah, literatur menawarkan dua solusi. Pertama, melarang atau membatasi AA antropomorfik yang memungkinkan simulasi kejahatan. Posisi ini memicu seruan untuk membatasi AA antropomorfik secara umum, karena "robot-robot ini adalah jenis robot [AA] yang paling mungkin disalahgunakan".

Kasus-kasus di mana bot sosial "dirancang, sengaja atau tidak, dengan mempertimbangkan gender, daya tarik dan realisme agen perempuan" menimbulkan pertanyaan "jika ECA (yaitu, bot sosial) mendorong stereotip gender, apakah ini akan berdampak pada perempuan sungguhan daring?". Sarannya adalah untuk membuat bot sosial tidak dapat diterima untuk meniru properti antropomorfik, seperti memiliki persepsi gender atau etnis. Mengenai robot seks yang meniru pelanggaran seksual, saran lebih lanjut adalah memberlakukan larangan sebagai "paket hukum yang membantu meningkatkan moralitas seksual sosial" dan memperjelas norma-norma intoleransi.

Saran kedua (meskipun tidak sesuai dengan yang pertama) adalah menggunakan AA antropomorfik sebagai cara untuk melawan simulasi pelanggaran seksual. Misalnya, mengenai penyalahgunaan agen pedagogis buatan, "kami merekomendasikan agar respons agen diprogram untuk mencegah atau mengurangi pelecehan siswa lebih lanjut". Sebagaimana yang dikemukakan Kate Darling, hal ini tidak hanya akan memerangi desensitisasi dan eksternalitas negatif dari perilaku manusia, tetapi juga akan mempertahankan manfaat terapeutik dan edukatif dari penggunaan robot tertentu yang lebih seperti teman daripada alat.

Menerapkan saran-saran ini membutuhkan pilihan apakah akan mengkriminalisasi sisi permintaan atau penawaran transaksi, atau keduanya. Pengguna mungkin berada dalam lingkup penerapan hukuman. Pada saat yang sama, dapat dikatakan bahwa seperti halnya

kejahatan lain yang melibatkan "kejahatan" pribadi, pemasok dan distributor juga dapat menjadi sasaran dengan alasan mereka memfasilitasi dan mendorong tindakan yang melanggar hukum. Bahkan, kita mungkin menargetkan mereka secara eksklusif atau preferensial, seperti yang sekarang dilakukan untuk obat-obatan terlarang di banyak negara.

12.5 KESIMPULAN

Artikel ini menyediakan analisis literatur sistematis pertama tentang Kejahatan Kecerdasan Buatan (AI-Crime/AIC), untuk menjawab dua pertanyaan. Pertanyaan pertama apa saja ancaman yang secara fundamental unik dan nyata yang ditimbulkan oleh AI? dijawab berdasarkan definisi kontrafaktual klasik AI dan, oleh karena itu, difokuskan pada AI sebagai reservoir agensi cerdas yang otonom. Ancaman-ancaman tersebut dijelaskan per area (dalam konteks kejahatan spesifik yang telah didefinisikan) dan secara lebih umum (dalam konteks kualitas AI dan isu-isu kemunculan, tanggung jawab, pemantauan, dan psikologi).

Pertanyaan kedua solusi apa yang tersedia atau mungkin dirancang untuk menangani AI? dijawab dengan berfokus pada tema umum dan lintas sektor, dan dengan memberikan gambaran terkini tentang solusi sosial, teknologi, dan hukum yang tersedia, serta keterbatasannya. Karena solusi yang disarankan dalam literatur untuk serangkaian tema lintas sektor (yang tak terelakkan) ini, solusinya, meskipun hanya sebagian, akan berlaku untuk berbagai area AIC. Ketidakpastian yang besar atas apa yang sudah diketahui tentang AIC (dalam hal ancaman spesifik area, ancaman umum, dan solusi) kini telah berkurang. Secara lebih luas, penelitian AIC masih dalam tahap awal dan oleh karena itu, berdasarkan analisis, visi tentatif untuk lima dimensi penelitian AIC di masa mendatang kini dapat diberikan.

Area Pemahaman yang lebih baik tentang area AIC memerlukan perluasan pengetahuan saat ini, khususnya yang berkaitan dengan: penggunaan AI dalam interogasi, yang hanya dibahas dalam satu makalah yang berfokus pada pertanggungjawaban; dan pencurian dan penipuan di ruang virtual (misalnya, gim daring dengan aset tidak berwujud yang memiliki nilai dunia nyata; dan AA yang melakukan manipulasi pasar yang muncul, yang hanya dipelajari dalam simulasi eksperimental).

Analisis tersebut mengungkapkan serangan rekayasa sosial sebagai kekhawatiran yang masuk akal, tetapi belum memiliki bukti dunia nyata untuk saat ini. Pembunuhan dan terorisme tampaknya tidak ada dalam literatur AIC, meskipun mereka menuntut perhatian mengingat teknologi yang didukung AI seperti pengenalan pola (misalnya, ketika anggota kelompok rentan secara tidak adil ditargetkan sebagai korban oleh pelaku atau tersangka oleh petugas penegak hukum), drone bersenjata, dan kendaraan self-driving yang semuanya mungkin memiliki penggunaan yang sah dan kriminal.

Penggunaan Ganda Sifat digital AI memfasilitasi penggunaan gandanya, sehingga memungkinkan bahwa aplikasi yang dirancang untuk penggunaan yang sah kemudian dapat diimplementasikan untuk melakukan pelanggaran pidana. Ini adalah kasus untuk UUV, misalnya. Semakin jauh AI dikembangkan dan semakin implementasinya menjadi meluas, semakin tinggi risiko penggunaan jahat atau kriminal. Jika tidak ditangani, risiko tersebut dapat menyebabkan penolakan masyarakat dan regulasi yang terlalu ketat terhadap teknologi

berbasis AI ini. Pada gilirannya, manfaat teknologi bagi individu dan masyarakat dapat terkikis karena penggunaan dan pengembangan AI semakin dibatasi.

Batasan semacam itu telah ditempatkan pada penelitian pembelajaran mesin ke dalam diskriminator visual pria homoseksual dan heteroseksual, yang dianggap terlalu berbahaya untuk dirilis secara penuh (yaitu, dengan kode sumber dan struktur data yang dipelajari) ke komunitas penelitian yang lebih luas, dengan mengorbankan reproduktifitas ilmiah. Bahkan ketika batasan mahal pada rilis AI tidak diperlukan, seperti yang ditunjukkan Adobe dengan menanamkan tanda air ke dalam teknologi reproduksi suara, pengembang eksternal dan jahat tetap dapat mereproduksi teknologi tersebut di masa mendatang. Mengantisipasi penggunaan ganda AI di luar teknik umum yang diungkapkan dalam analisis, dan kemanjuran kebijakan untuk membatasi rilis teknologi AI, memerlukan penelitian lebih lanjut. Ini khususnya kasus implementasi AI untuk keamanan siber.

Keamanan Literatur AIC mengungkapkan bahwa, dalam lingkup keamanan siber, AI mengambil peran jahat dan ofensif bersamaan dengan sistem AI defensif yang dikembangkan dan diterapkan untuk meningkatkan ketahanan (dalam menahan serangan) dan ketahanan (dalam mencegah serangan), serta untuk melawan ancaman saat muncul. DARPA Cyber Grand Challenge 2016 merupakan titik balik untuk menunjukkan efektivitas pendekatan gabungan AI ofensif-defensif, dengan tujuh sistem AI terbukti mampu mengidentifikasi dan menambal kerentanan mereka sendiri, sementara juga menyelidiki dan mengeksploitasi kerentanan sistem pesaing.

Baru-baru ini, IBM meluncurkan Cognitive SOC ("*Cognitive Security – Watson for Cyber Security* | IBM" 2018). Ini adalah penerapan algoritma pembelajaran mesin yang menggunakan data keamanan terstruktur dan tidak terstruktur suatu organisasi, termasuk konten yang diekstrak dari blog, artikel, laporan, untuk menguraikan informasi tentang topik dan ancaman keamanan, dengan tujuan meningkatkan identifikasi, mitigasi, dan respons ancaman. Tentu saja, meskipun kebijakan jelas akan memainkan peran kunci dalam memitigasi dan memperbaiki risiko penggunaan ganda setelah penerapan (misalnya, dengan mendefinisikan mekanisme pengawasan), risiko-risiko ini paling tepat ditangani pada tahap desain.

Namun, bertentangan dengan laporan terbaru tentang AI berbahaya (Brundage dkk. 2018, 65), yang menyatakan bahwa "salah satu harapan terbaik kita untuk bertahan melawan peretasan otomatis juga melalui AI", analisis AIC menunjukkan bahwa ketergantungan yang berlebihan pada AI dapat menjadi kontraproduktif. Semua ini menekankan perlunya penelitian lebih lanjut tentang AI dalam keamanan siber tetapi juga pada alternatif AI, seperti berfokus pada manusia dan faktor sosial.

Orang Meskipun literatur menunjukkan kemungkinan faktor psikologis (misalnya, kepercayaan) dalam peran AI dalam kejahatan, penelitian tentang faktor pribadi yang dapat menciptakan pelaku, seperti pemrogram dan pengguna AI untuk AIC, di masa mendatang masih kurang. Sekaranglah saatnya untuk berinvestasi dalam studi longitudinal dan analisis multivariat yang mencakup latar belakang pendidikan, geografis, dan budaya korban, dan

pelaku atau bahkan pengembang AI yang baik hati, yang akan membantu memprediksi bagaimana individu bersatu untuk melakukan AIC.

Organisasi Laporan empat tahunan terbaru Europol (2017) tentang ancaman kejahatan serius dan terorganisir, menyoroti cara-cara di mana jenis kejahatan teknologi cenderung berkorelasi dengan topologi organisasi kriminal tertentu. Literatur AIC menunjukkan bahwa AI dapat berperan dalam organisasi kriminal seperti kartel narkoba, yang memiliki sumber daya yang baik dan sangat terorganisir.

Sebaliknya, organisasi kriminal ad hoc di web gelap sudah terjadi di bawah apa yang disebut Europol sebagai kejahatan sebagai layanan. Layanan kriminal semacam itu dijual langsung antara pembeli dan penjual, berpotensi sebagai elemen yang lebih kecil dalam keseluruhan kejahatan, yang dapat dipicu oleh AI (misalnya, dengan memungkinkan peretasan profil) di masa mendatang.

Dalam spektrum yang berkisar dari organisasi AIC yang erat hingga fleksibel, terdapat banyak kemungkinan interaksi kriminal; mengidentifikasi organisasi yang esensial atau yang tampaknya berkorelasi dengan berbagai jenis AIC akan semakin memahami bagaimana AIC terstruktur dan beroperasi dalam praktiknya. AI memang menimbulkan risiko yang signifikan, karena dapat mengurangi keterampilan kejahatan, dan karenanya menyebabkan perluasan apa yang disebut Europol sebagai ekonomi berbagi kriminal.

Mengembangkan pemahaman yang lebih mendalam tentang dimensi-dimensi ini sangat penting untuk melacak dan menghentikan pertumbuhan AIC yang tak terelakkan di masa depan. Oleh karena itu, analisis literatur ini bertujuan untuk memicu penelitian lebih lanjut mengenai kekhawatiran yang sangat serius, berkembang, namun masih relatif belum tereksplorasi terkait AIC. Semakin cepat fenomena kejahatan baru ini dipahami, semakin cepat pula kebijakan pencegahan, mitigasi, disinsentif, dan pemulihan dapat diterapkan.

BAB 13

ATURAN KECERDASAN BUATAN DALAM MENCEGAH PERANG SIBER

Perkembangan teknologi digital telah membawa dunia pada era baru di mana serangan siber semakin sering terjadi, lebih canggih, dan menimbulkan dampak yang merusak. Data menunjukkan bahwa pada tahun 2017, Amerika Serikat mengalami rata-rata lebih dari 4.000 serangan ransomware setiap hari angka yang meningkat drastis dibandingkan 1.000 serangan per hari pada tahun 2015. Serangan-serangan besar seperti *WannaCry* yang melumpuhkan sistem Layanan Kesehatan Nasional Inggris, serta *NotPetya* yang merugikan perusahaan global seperti Merck, Maersk, dan FedEx hingga ratusan juta dolar, memperlihatkan betapa seriusnya ancaman ini. Diperkirakan, kerugian global akibat serangan siber dapat mencapai 6 triliun dolar AS per tahun pada 2021.

Negara-negara turut menjadi aktor penting dalam dinamika ini. Beberapa serangan besar dikaitkan dengan dukungan atau keterlibatan negara, seperti Korea Utara dengan *WannaCry* dan Rusia dengan *NotPetya*. Sebagai respons, berbagai negara mengembangkan strategi pertahanan siber aktif. Amerika Serikat, misalnya, sejak 2012 menerapkan langkah-langkah untuk menetralkan atau mengalihkan serangan, sementara Inggris mengumumkan investasi besar untuk memperkuat ketahanan digital. NATO juga mulai menyusun prinsip pertahanan siber kolektif dengan melibatkan sejumlah negara anggotanya.

Di tengah perkembangan tersebut, Kecerdasan Buatan (*Artificial Intelligence/AI*) muncul sebagai teknologi kunci yang berpotensi merevolusi perang siber. AI dapat meningkatkan kecepatan, ketepatan, dan efektivitas dalam mendeteksi maupun merespons ancaman. Algoritma pembelajaran mesin telah digunakan untuk mengidentifikasi perilaku abnormal dalam jaringan, memblokir serangan secara otomatis, hingga melakukan analisis sumber ancaman tanpa perlu mengidentifikasi langsung aktor di baliknya. Nilai pasar AI dalam bidang keamanan siber sendiri diperkirakan melonjak dari 1 miliar dolar AS pada tahun 2016 menjadi 18 miliar dolar AS pada 2023.

Namun, pemanfaatan AI di bidang pertahanan siber juga menimbulkan risiko serius. Kompetisi antarnegara dalam mengembangkan strategi berbasis AI dapat memicu perlombaan senjata siber yang berbahaya. Eskalasi saling serang berpotensi tidak hanya mengacaukan dunia maya, tetapi juga merembet pada konfrontasi fisik. Dengan kata lain, tanpa adanya regulasi yang jelas, penggunaan AI dalam konflik siber dapat mengancam keamanan nasional maupun stabilitas internasional.

Menyadari risiko tersebut, penting untuk membangun kerangka hukum dan regulasi internasional yang mengatur penggunaan AI oleh negara dalam konteks pertahanan siber. Regulasi ini setidaknya harus mencakup tiga aspek utama:

- (1) batasan hukum yang jelas mengenai target sah dan tidak sah serta proporsionalitas respons;

- (2) strategi uji coba bersama dengan sekutu, termasuk simulasi serangan dan pertahanan berbasis AI; dan
- (3) keberadaan otoritas pihak ketiga yang berwenang untuk memantau, menegakkan aturan, serta menentukan apakah suatu tindakan melanggar prinsip proporsionalitas, tanggung jawab, atau norma internasional.

Bab ini bertujuan untuk menganalisis urgensi regulasi internasional atas penggunaan AI dalam pertahanan siber, dengan menyoroti dinamika serangan siber, potensi AI dalam keamanan digital, serta risiko perlombaan senjata siber. Pembahasan ini menekankan perlunya forum regional seperti NATO dan Uni Eropa untuk menghidupkan kembali inisiatif regulasi, sekaligus mempersiapkan fondasi bagi upaya yang lebih luas di bawah koordinasi Perserikatan Bangsa-Bangsa. Dengan demikian, diharapkan tercipta keseimbangan antara pemanfaatan AI sebagai instrumen pertahanan yang sah dan pencegahan terhadap eskalasi konflik yang membahayakan perdamaian global.

13.1 TANPA ATURAN

Saat ini, proses PBB menemui jalan buntu. Pada tahun 2004, PBB membentuk Kelompok Pakar Pemerintahan tentang Keamanan Informasi untuk menyepakati aturan sukarela tentang bagaimana negara harus berperilaku di dunia maya. Pertemuan kelimanya, pada tahun 2017, berakhir dengan kebuntuan. Kelompok tersebut tidak dapat mencapai konsensus tentang apakah hukum humaniter internasional dan hukum yang ada tentang pembelaan diri dan tanggung jawab negara harus diterapkan di dunia maya. Amerika Serikat berpendapat bahwa peraturan pertahanan siber harus didasarkan pada hukum-hukum ini.

Negara-negara lain, termasuk Kuba, Rusia, dan Tiongkok, tidak setuju. Mereka berpendapat bahwa hal ini akan 'memiliterisasi' dunia maya dan mengirimkan pesan yang salah tentang resolusi konflik secara damai. Kelompok tersebut gagal menyampaikan laporannya. Belum jelas apakah mereka akan bertemu lagi, atau apa yang akan terjadi selanjutnya. Dialog dan aksi internasional harus dilanjutkan. NATO dapat membuka jalan melalui pedoman yang akan datang, meskipun saat ini belum jelas cakupannya.

Sementara itu, penelitian tentang AI untuk pertahanan siber berkembang pesat. Amerika Serikat memimpin dalam hal teknologi. Amerika Serikat bertujuan untuk mengintegrasikan AI ke dalam sistem pertahanan sibernya pada tahun 2019. Departemen Pertahanan AS (DOD) telah mengalokasikan Rp.2 Triliun untuk penelitian. Badan Proyek Penelitian Lanjutan Pertahanan AS (DARPA) sedang mengembangkan teknik dan strategi. Langkah-langkah telah diambil. Dalam kompetisi Cyber Grand Challenge DARPA 2016, tujuh sistem AI, yang dikembangkan oleh tim dari Amerika Serikat dan Swiss, saling bertarung. Sistem-sistem tersebut mengidentifikasi dan menargetkan kelemahan lawan mereka sambil menemukan dan menambal kelemahan mereka sendiri.

DOD akan menerbitkan laporan AS pertama tentang strategi AI untuk pertahanan nasional pada bulan Mei. Sejauh yang kami ketahui, belum ada indikasi seperti apa pendekatannya. Dokumen-dokumen sebelumnya, seperti Strategi Siber Departemen Pertahanan (DOD) tahun 2015 atau Rencana Tanggap Insiden Siber Nasional 2016, tidak

mencakup sistem otonom, pembelajaran mesin, atau AI. Arahan Departemen Pertahanan (DOD) tahun 2012 tentang 'Otonomi dalam Sistem Persenjataan' berfokus pada prosedur internal untuk penerapan AI, tetapi tidak menyebutkan kapan Amerika Serikat akan melakukannya di kancah internasional.

AI merupakan prioritas bagi Tiongkok, yang bertujuan untuk menjadi pemimpin dunia dalam teknologi pembelajaran mesin. Pada Juli 2017, pemerintah Tiongkok menerbitkan Rencana Pengembangan AI Generasi Berikutnya. Implementasi AI di militer, baik di medan perang maupun di dunia maya, merupakan bagian penting dari strategi tersebut. Namun, belum jelas sejauh mana Tiongkok berencana untuk menerapkan AI secara aktif dalam pertahanan siber.

Rusia belum merilis dokumen publik apa pun tentang strategi AI dalam pertahanannya. Namun, dalam pesan video yang dirilis pada tahun 2017, Presiden Vladimir Putin merujuk pada AI dan menyatakan: "Siapa pun yang menjadi pemimpin di bidang ini akan menjadi penguasa dunia". Para ahli sepakat bahwa Rusia berfokus pada pengembangan alat yang disempurnakan dengan AI untuk pasukan konvensional.

Namun, sejak 2014, Pusat Kontrol Pertahanan Nasional Rusia telah menggunakan algoritma pembelajaran mesin untuk mendeteksi ancaman daring. Diduga, Rusia telah memelopori penggunaan AI untuk menyebarkan disinformasi dan mengintervensi debat publik negara lain, termasuk pemilihan presiden AS 2016 dan referendum keanggotaan Uni Eropa di Inggris. Meskipun operasi-operasi ini bukan bagian dari strategi pertahanan nasional, operasi-operasi ini menunjukkan kemampuan AI Rusia yang canggih. Korea Utara memiliki sejarah agresi dunia maya. Negara ini terlibat, misalnya, dalam serangan Wanna Cry pada tahun 2016 dan dalam pelanggaran besar lainnya, terhadap Sony Pictures, pada tahun 2014. Negara ini kurang memiliki keahlian teknis dalam AI tetapi kemungkinan ingin mengejar ketinggalan dengan para musuhnya.

Uni Eropa juga meningkatkan langkahnya. Pada tahun 2017, Uni Eropa menilai kembali kebijakan keamanan siber dan pertahanan dan meluncurkan Pusat Keunggulan Eropa untuk Melawan Ancaman Hibrida, yang berpusat di Helsinki. Uni Eropa memiliki kerangka regulasi paling komprehensif untuk perilaku negara di dunia maya sejauh ini. Namun arahan ini tidak cukup jauh. Uni Eropa memperlakukan pertahanan siber sebagai kasus keamanan siber, yang harus ditingkatkan secara pasif dengan membuat sistem informasi negara-negara anggota lebih tangguh.

Uni Eropa mengabaikan penggunaan aktif pertahanan siber dan tidak memasukkan AI. Ini adalah kesempatan yang terlewatkan. Uni Eropa seharusnya mulai mendefinisikan garis merah dan tanggapan proporsional dalam pemikiran ulang terbarunya. Misalnya, arahan Uni Eropa tahun 2016 tentang 'Keamanan Jaringan dan Sistem Informasi' memberikan kriteria untuk mengidentifikasi infrastruktur nasional krusial, seperti sistem kesehatan atau pasokan energi dan air utama, yang harus dilindungi. Kriteria yang sama dapat digunakan untuk menentukan target serangan siber ilegal yang disponsori negara. Forum regional, seperti NATO dan Uni Eropa, harus mengambil tiga langkah berikut untuk menghindari serangan serius yang akan segera terjadi terhadap infrastruktur negara, dan untuk menjaga stabilitas internasional.

13.2 TIGA LANGKAH

Menetapkan Batasan Hukum. Komunitas internasional perlu segera menyepakati garis merah yang membedakan antara target yang sah dan tidak sah. Definisi respons proporsional untuk strategi pertahanan siber juga diperlukan. Konsensus internasional di tingkat PBB pada akhirnya akan dibutuhkan. Hingga saat itu, pedoman dari badan multilateral regional, seperti NATO dan Uni Eropa, harus mencakup isu-isu ini dan menjadi contoh.

Menguji Strategi dengan Sekutu. Latihan 'Sparring' harus diselenggarakan antar negara sahabat untuk menguji taktik pertahanan berbasis AI. Uji coba ini wajib dilakukan sebelum sistem apa pun dikerahkan. Uji coba ini dapat berupa Grand Challenge DARPA atau latihan simulasi yang rutin dijalankan oleh NATO dan Uni Eropa. Karena AI belajar dari pengalaman, simulasi ini akan meningkatkan strategi aliansi, sekaligus menemukan dan memperbaiki kelemahan. Kerentanan fatal pada sistem utama dan infrastruktur krusial harus dibagikan dengan sekutu; kerangka kerja kebijakan harus mewajibkan pengungkapan. Perjanjian dan peraturan dengan persyaratan berbagi dan pengungkapan serupa mencakup Peraturan Identifikasi Elektronik, Autentikasi, dan Layanan Kepercayaan Uni Eropa dan Perjanjian Kemitraan Industri NATO.

Pantau dan Tegakkan Aturan. Komunitas internasional perlu menyepakati cara mengaudit dan mengawasi mekanisme siber negara berbasis AI untuk mengatasi kesalahan dan konsekuensi yang tidak diinginkan. Otoritas pihak ketiga yang berwenang, seperti Dewan Keamanan PBB, harus memutuskan apakah garis merah, proporsionalitas, penerapan yang bertanggung jawab, atau norma pengungkapan telah dilanggar. Sanksi ekonomi atau politik harus dijatuhkan kepada negara yang melanggar aturan. NATO dan Uni Eropa harus menegakkan norma-norma tersebut dalam kewenangan mereka.

13.3 AMBIGUITAS PERAN AI DALAM KEAMANAN SIBER

Penerapan kecerdasan buatan (*Artificial Intelligence/AI*) dalam keamanan siber semakin menjadi perhatian utama, baik di sektor swasta maupun publik. Berbagai proyeksi menunjukkan pertumbuhan pasar AI yang sangat pesat dalam bidang ini, yakni dari sekitar 1 miliar dolar AS pada tahun 2016 menjadi 34,8 miliar dolar AS pada tahun 2025. Sejalan dengan itu, banyak strategi keamanan siber nasional dari berbagai negara, seperti Australia, Tiongkok, Jepang, Singapura, Inggris, dan Amerika Serikat, secara eksplisit menyebutkan kapabilitas AI sebagai elemen penting dalam memperkuat pertahanan digital, terutama untuk melindungi infrastruktur vital seperti transportasi, energi, rumah sakit, dan pasokan air. Di samping itu, inisiatif global untuk menetapkan standar, prosedur sertifikasi, dan tata kelola AI juga mulai berkembang, dengan tujuan membangun kepercayaan masyarakat terhadap teknologi ini.

Namun, kepercayaan terhadap AI dalam menjalankan tugas-tugas keamanan siber bersifat ambivalen ibarat pedang bermata dua. Di satu sisi, AI mampu meningkatkan efektivitas praktik keamanan siber secara substansial, tetapi di sisi lain, AI sendiri rentan menjadi target serangan baru. Algoritma pembelajaran mesin dan jaringan saraf tiruan, misalnya, dapat dimanipulasi untuk melemahkan sistem pertahanan yang seharusnya

diperkuat. Dengan demikian, penerapan AI tidak hanya membuka peluang besar untuk memperkuat keamanan, tetapi juga menimbulkan kategori kerentanan baru yang berpotensi menghadirkan ancaman serius.

Urgensi ini semakin jelas jika meninjau eskalasi ancaman siber global. Forum Ekonomi Dunia (2019) menempatkan serangan siber sebagai salah satu dari lima risiko terbesar yang paling mungkin terjadi dan berdampak parah secara global. Data mendukung temuan ini: pada paruh pertama tahun 2018, serangan siber membahayakan lebih dari 3,3 miliar data, hampir 70% lebih banyak daripada sepanjang tahun 2017. Lebih mengkhawatirkan lagi, sebuah studi Microsoft menunjukkan bahwa 60% serangan pada tahun 2018 hanya memerlukan waktu kurang dari satu jam untuk menembus target, sebagian besar dengan memanfaatkan varian malware baru.

Di tengah meningkatnya ancaman tersebut, AI dipandang mampu mengurangi risiko dan meningkatkan efisiensi tim keamanan siber melalui tiga dimensi utama yang dalam bab ini disebut sebagai *3R*: ketahanan, respons, dan resiliensi. Pertama, AI memperkuat ketahanan sistem, yaitu kemampuan sistem untuk tetap berfungsi sesuai harapan meskipun menerima input berbahaya, dengan bantuan perangkat lunak pengujian dan pemulihan mandiri. Kedua, AI memajukan respons sistem melalui deteksi otomatis, simulasi *honeypot*, serta peluncuran strategi pertahanan yang adaptif dan agresif terhadap serangan. Ketiga, AI memperkuat resiliensi, yakni kemampuan sistem untuk bertahan menghadapi serangan dengan mendukung deteksi ancaman dan anomali (TAD), yang diperkirakan 50% akan ditangani AI pada tahun 2022, serta membantu analis keamanan dalam memahami pola serangan siber.

Manfaat *3R* ini menjelaskan mengapa AI semakin dipercaya untuk memperkuat keamanan digital. Secara taktis, AI meningkatkan perlindungan sistem dan mengurangi peluang serangan. Secara strategis, AI bahkan mampu mengubah dinamika keamanan siber, misalnya dengan mengurangi nilai serangan *zero-day* di pasar gelap karena sistem lebih tahan terhadap kerentanan baru. Selain itu, sistem AI yang mampu merespons serangan tanpa bergantung pada identitas penyerang memungkinkan adanya pertahanan yang tetap efektif bahkan saat serangan dilakukan secara anonim.

Meski demikian, keunggulan AI ini tidak boleh disamakan dengan kepercayaan mutlak. Perlu dibedakan antara kepercayaan dan ketergantungan. Kepercayaan berarti pendelegasian penuh tanpa kendali, sedangkan ketergantungan menuntut adanya mekanisme pengawasan, kontrol, dan bahkan penghentian ketika diperlukan. Dalam konteks ini, kepercayaan mutlak terhadap AI untuk *3R* dinilai tidak beralasan. Sebaliknya, yang dibutuhkan adalah pengendalian yang memastikan penerapan *AI yang andal* yakni AI yang dirancang, dikembangkan, dan diterapkan dengan mekanisme keamanan yang memadai.

Bab ini akan membahas secara mendalam urgensi pengendalian tersebut, sekaligus menawarkan tiga rekomendasi praktis mengenai desain, pengembangan, dan implementasi AI dalam keamanan siber. Dengan kerangka ini, diharapkan pembahasan tidak hanya bersifat konseptual, tetapi juga memberikan kontribusi nyata bagi perumusan kebijakan dan strategi keamanan siber global yang lebih adaptif dan aman.

13.4 KERENTANAN AI

Serangan siber generasi lama umumnya berfokus pada dua hal: pencurian data (*data extraction*) dan perusakan sistem (*system disruption*). Namun, ancaman terhadap kecerdasan buatan (AI) membawa dimensi baru. Alih-alih sekadar mencuri atau menghancurkan, serangan modern berusaha *menguasai* sistem AI dan memodifikasi perilakunya. Dampak serangan ini sangat serius karena dapat melemahkan potensi AI dalam meningkatkan keamanan, efisiensi, dan keandalan (konsep 3R: *reliability, resilience, robustness*).

Ada tiga jenis serangan utama yang relevan untuk menguasai AI: peracunan data (*data poisoning*), perombakan model kategorisasi (*model tampering*), dan pintu belakang (*backdoor attacks*). Semuanya mengeksploitasi kemampuan inti AI, yaitu belajar dari data, dengan cara memanipulasi informasi yang digunakan untuk melatih model.

Sebagai contoh, dalam serangan peracunan data, penyerang menyisipkan data palsu yang tampak sah ke dalam kumpulan pelatihan. Penelitian menunjukkan bahwa hanya dengan menambahkan sekitar 8% data keliru, sistem AI untuk dosis obat dapat disesatkan sehingga menghasilkan kesalahan dosis sebesar 75,06% bagi separuh pasien yang bergantung padanya. Serangan serupa juga bisa dilakukan terhadap sistem klasifikasi. Peneliti berhasil membuat model AI salah mengidentifikasi kura-kura cetak 3D sebagai senapan dengan memanipulasi cara AI belajar dari data visual.

Serangan pintu belakang bahkan lebih berbahaya. Dengan menambahkan “pemicu tersembunyi” ke dalam data pelatihan, AI akan tetap bekerja normal sampai pemicu tertentu muncul. Misalnya, sebuah penelitian menunjukkan bahwa dengan menempelkan stiker khusus pada rambu berhenti lalu melabelinya sebagai rambu batas kecepatan, sistem AI dilatih untuk mengenali rambu itu secara salah. Akibatnya, kendaraan otonom bisa melaju melewati persimpangan tanpa berhenti, yang jelas menimbulkan risiko keselamatan besar.

Masalahnya, serangan terhadap AI sangat sulit dideteksi. Hal ini karena sifat AI yang kompleks, dinamis, dan adaptif, serta adanya fenomena “kurangnya transparansi” (*lack of explainability*). Sistem AI sering kali bertindak sebagai *black box*, sehingga sulit dijelaskan bagaimana hasil tertentu dihasilkan apakah normal atau akibat manipulasi. Bahkan ketika serangan pintu belakang diaktifkan, perbedaannya mungkin sangat kecil, hanya berupa deviasi tipis dari perilaku normal. Misalnya, penyerang dapat menggunakan kacamata khusus untuk mengelabui sistem pengenalan wajah agar salah mengidentifikasi orang.

Jika serangan semacam ini menargetkan sistem keamanan, dampaknya bisa fatal. Bayangkan sebuah sistem kontrol akses yang dimanipulasi agar mengizinkan penjahat masuk tanpa memicu alarm. Inilah alasan mengapa ketahanan AI menjadi isu krusial. AI harus tetap dapat berfungsi sesuai harapan meskipun input atau modelnya disusupi. Sayangnya, menilai ketahanan bukan hal sederhana. Untuk benar-benar menguji daya tahan AI, idealnya semua kemungkinan gangguan input diperiksa. Namun, jumlah kemungkinan itu nyaris tak terbatas. Bahkan gangguan sekecil perubahan piksel yang tidak terlihat mata manusia bisa membuat AI salah mengklasifikasikan objek dengan tingkat keyakinan tinggi.

Karena itu, menilai ketahanan AI pada tahap desain dan pengembangan hanya memberikan gambaran terbatas. Begitu AI diterapkan dalam situasi nyata, celah-celah baru

bisa muncul yang sebelumnya tidak terdeteksi. Dengan kata lain, pengujian tradisional tidak cukup untuk menjamin keandalan sistem AI terhadap serangan. Oleh sebab itu, dibutuhkan pendekatan baru yang lebih inovatif dan menyeluruh untuk memastikan AI tetap aman, tangguh, dan dapat dipercaya dalam menghadapi berbagai ancaman.

13.5 STANDAR DAN PROSEDUR SERTIFIKASI

Kerentanan yang melekat pada sistem kecerdasan buatan (AI) menimbulkan keterbatasan serius terhadap potensinya dalam memperkuat keamanan siber. Untuk menjawab tantangan ini, diperlukan metode pengujian baru yang mampu mengatasi dua hal utama: kurangnya transparansi AI dan sifat serangan yang menipu. Dalam konteks global, berbagai inisiatif mulai bermunculan untuk merumuskan standar baru dan prosedur sertifikasi guna menilai ketahanan AI.

Secara internasional, Organisasi Internasional untuk Standardisasi (ISO) telah membentuk komite khusus (ISO/IEC JTC 1/SC 42) yang menangani standar AI. Salah satu hasilnya adalah standar ISO/IEC NP TR 24029–1, yang berfokus pada penilaian ketahanan jaringan saraf tiruan. Di Amerika Serikat, DARPA pada tahun 2019 meluncurkan program *Guaranteeing AI Robustness against Deception (GARD)* untuk mendorong pengembangan AI yang tangguh terhadap manipulasi. Pada tahun yang sama, sebuah perintah eksekutif tentang AI menginstruksikan perlunya standar nasional terkait keandalan, ketangguhan, dan kepercayaan AI. NIST kemudian menindaklanjutinya dengan permintaan masukan publik guna merumuskan standar tersebut.

Tiongkok pun tidak ketinggalan. Sejalan dengan *Rencana Pengembangan Kecerdasan Buatan Generasi Baru*, pada 2019 mereka membentuk tiga kelompok kerja yang menangani isu *AI dan sumber terbuka*, *standardisasi AI di Tiongkok*, serta *AI dan etika sosial*. Pedoman resmi diharapkan terbit pada akhir tahun tersebut. Sementara itu, Uni Eropa (UE) berpotensi menjadi pionir internasional dalam sertifikasi keamanan siber, berbekal *Kerangka Kerja Keamanan Siber 2017* dan *Undang-Undang Keamanan Siber 2019*. Kedua regulasi ini memberi mandat pada ENISA (Badan Uni Eropa untuk Keamanan Jaringan dan Informasi) untuk menyusun kerangka sertifikasi. Namun, fokus tujuan awal ENISA lebih banyak pada isu data dan kerentanan umum, belum secara eksplisit menyebut AI. Padahal, jika AI diabaikan, sertifikasi yang ada hanya akan memperkuat sebagian aspek keamanan digital di pasar Eropa, tanpa menjawab risiko spesifik dari sistem AI.

Walau masih dalam tahap awal, semua inisiatif tersebut memiliki benang merah yang sama: membangun kepercayaan terhadap AI. Kepercayaan dipandang sebagai syarat utama agar masyarakat dan organisasi mau mengadopsi AI dalam bidang keamanan siber. Pandangan ini tercermin dalam laporan IEEE 2017, perintah eksekutif AS, hingga pedoman Komisi Eropa. Namun, ada persoalan mendasar. Secara filosofis, kepercayaan berarti mendelegasikan tugas tanpa pengawasan penuh terhadap cara pelaksanaannya. Keputusan untuk percaya bergantung pada penilaian apakah pihak yang diberi kepercayaan (wali amanat) kemungkinan besar akan bertindak sesuai harapan, berdasarkan pengalaman masa lalu dan risiko yang mungkin timbul jika ia gagal.

Masalahnya, penilaian ini sulit diterapkan pada AI. AI bersifat tidak transparan, adaptif, dan rentan dimanipulasi melalui serangan seperti pintu belakang yang tersembunyi. Catatan perilaku masa lalu AI tidak dapat menjadi indikator kuat untuk memprediksi kinerjanya di masa depan. Dengan kata lain, tidak ada jaminan bahwa AI akan tetap tangguh menghadapi serangan baru. Hal ini membuat dasar bagi “kepercayaan” pada AI menjadi rapuh. Bukan berarti AI tidak boleh diberdayakan dalam tugas-tugas keamanan siber. Justru sebaliknya, AI tetap harus dimanfaatkan karena terbukti mampu meningkatkan efisiensi dan efektivitas. Akan tetapi, pendelegasian tersebut harus disertai mekanisme kontrol yang memadai untuk mengurangi risiko akibat sifat AI yang tidak transparan dan tidak sepenuhnya dapat diprediksi.

Dengan demikian, strategi kebijakan yang hanya berfokus pada membangun “kepercayaan” pengguna terhadap AI tanpa menambahkan lapisan kontrol yang nyata akan gagal mengatasi masalah inti. Yang dibutuhkan bukan sekadar membangun rasa percaya, melainkan merancang kerangka pengawasan dan verifikasi yang mampu menjamin AI tetap aman, andal, dan bertanggung jawab ketika digunakan untuk kepentingan keamanan siber.

13.6 MENJADIKAN AI DALAM KEAMANAN SIBER HANDAL

Standar dan metode sertifikasi AI yang baru lahir dalam keamanan siber seharusnya berfokus pada dukungan keandalan AI, alih-alih kepercayaan. Secara konseptual dan operasional, mendukung keandalan AI berbeda dengan membangun kepercayaannya. Keandalan AI menyiratkan bahwa teknologi tersebut, secara teknis, dapat menjalankan tugas-tugas keamanan siber dengan sukses, tetapi risiko bahwa teknologi tersebut mungkin berperilaku berbeda dari yang diharapkan terlalu tinggi untuk mengabaikan segala bentuk kontrol atau pemantauan atas pelaksanaan tugas yang didelegasikan.

Dengan demikian, mendukung keandalan AI untuk tugas-tugas 3R menyiratkan adanya bentuk dan tingkat kontrol yang memadai untuk sifat pembelajaran sistem, kurangnya transparansi, dan sifat dinamis serangan, sambil tetap layak dalam hal sumber daya, terutama waktu dan dengan demikian kelayakan komputasi. Berikut ini, kami menyarankan tiga persyaratan yang menentukan praktik pengembangan dan pemantauan untuk memitigasi kerentanan sistem AI dan meningkatkan keandalannya sehubungan dengan 3R.

1. Pengembangan internal. Bentuk serangan paling umum terhadap sistem AI difasilitasi oleh penggunaan layanan komersial yang menawarkan dukungan untuk pengembangan dan pelatihan AI, seperti mesin virtual, pemrosesan bahasa alami, analitik prediktif, dan pembelajaran mendalam. Pelanggaran dalam sistem cloud, misalnya, dapat memberi penyerang akses ke model AI dan data pelatihan. Oleh karena itu, standar aplikasi AI untuk keamanan infrastruktur penting nasional harus memastikan bahwa pemasok yang andal merancang dan mengembangkan model mereka sendiri, dan bahwa data untuk pelatihan dan pengujian sistem dikumpulkan, dikurasi, dan divalidasi oleh penyedia sistem secara langsung dan dipelihara dengan aman. Meskipun persyaratan ini tidak akan menghilangkan semua kemungkinan serangan, hal ini akan mencegah banyak bentuk serangan yang memanfaatkan koneksi internet untuk mengakses data dan model.

2. Pelatihan adversarial. AI meningkatkan kinerjanya menggunakan loop umpan balik, yang memungkinkannya menyesuaikan variabel dan koefisiennya sendiri di setiap iterasi. Inilah sebabnya mengapa pelatihan adversarial antar sistem AI dapat membantu meningkatkan ketahanannya serta memfasilitasi identifikasi kerentanan sistem. Ini merupakan metode yang umum digunakan untuk meningkatkan ketahanan sistem. Namun, penelitian juga menunjukkan bahwa efektivitasnya bergantung pada penyempurnaan model adversarial. Standar dan proses sertifikasi harus mewajibkan pelatihan adversarial, tetapi juga menetapkan tingkat penyempurnaan model yang sesuai. Dalam hal ini, penting bagi model untuk dikembangkan secara internal dan spesifik untuk tugas yang sedang dikerjakan.
3. Pemantauan paralel dan dinamis. Keterbatasan dalam menilai ketahanan sistem AI, sifat serangan yang menipu, dan kemampuan belajar sistem target memerlukan suatu bentuk pemantauan yang konstan (tidak hanya berkala, yaitu pada interval waktu tertentu, tetapi berkelanjutan, 24 jam sehari, 7 hari seminggu) selama penerapan. Pemantauan diperlukan untuk memastikan bahwa divergensi antara perilaku sistem yang diharapkan dan aktual dapat dideteksi sejak dini dan segera, serta ditangani secara memadai. Untuk melakukannya, penyedia sistem AI harus memelihara sistem klon sebagai sistem kendali. Sistem klon tidak boleh dianggap sebagai 'kembaran digital' dari sistem yang diterapkan. Klon bukanlah simulasi virtual dari sistem AI, melainkan sistem yang sama yang diterapkan dalam kondisi lingkungan yang terkendali. Dan perilakunya bukanlah simulasi dari sistem asli, melainkan tolok ukur (dasar) yang digunakan untuk menilai perilaku sistem asli. Klon harus menjalani latihan adversarial secara berkala, mensimulasikan serangan di dunia nyata untuk menetapkan perilaku dasar yang dapat dijadikan tolok ukur perilaku sistem yang diterapkan. Divergensi antara klon dan sistem yang diterapkan harus menandai tingkat peringatan keamanan. Ambang batas divergensi, yang sepadan dengan risiko keamanan, harus ditetapkan berdasarkan kasus per kasus. Perlu dicatat bahwa ambang batas yang terlalu sensitif (misalnya, ambang batas 0%) dapat membuat pemantauan dan pengendalian menjadi tidak layak, sementara ambang batas yang terlalu tinggi akan membuat sistem tidak andal. Namun, untuk sistem yang memenuhi persyaratan (1) dan (2), divergensi minimal tidak boleh sering terjadi dan kecil kemungkinannya untuk mengindikasikan positif palsu. Dengan demikian, ambang batas 0% untuk sistem ini mungkin tidak menimbulkan batasan yang signifikan terhadap operabilitasnya, sementara itu akan memungkinkan sistem untuk menandai ancaman konkret.

AI dapat meningkatkan 3R hanya sejauh ia andal. Bayangkan, misalnya, menerapkan sistem AI untuk tugas TAD tanpa dapat mengecualikan keberadaan pintu belakang dalam sistem AI itu sendiri, sehingga ada kemungkinan penyerang dapat mengendalikan sistem AI dan memastikan bahwa serangan spesifik pada sistem yang dipantau tidak terdeteksi.

Tiga persyaratan yang kami anjurkan merupakan prasyarat bagi sistem AI untuk menjalankan salah satu tugas 3R dengan andal, dan seharusnya menjadi prasyarat penting bagi sistem AI yang diterapkan untuk keamanan infrastruktur kritis nasional. Implementasinya

mungkin terlalu mahal untuk aplikasi AI komersial rata-rata untuk keamanan siber. Inilah sebabnya mengapa kita mungkin membayangkan bahwa usaha kecil dan menengah mungkin hanya mengadopsi sebagian persyaratan ini; hal ini mungkin bergantung, misalnya, pada sifat bisnis mereka dan sifat sistem yang akan dilindungi. Namun, persyaratan ini harus dipenuhi sepenuhnya ketika mempertimbangkan keamanan dan pertahanan nasional. Risiko yang ditimbulkan oleh serangan terhadap sistem AI yang mendasari infrastruktur kritis membenarkan perlunya mekanisme pengendalian yang lebih luas, dan karenanya investasi yang lebih tinggi.

Sistem AI adalah agen otonom dan belajar mandiri yang berinteraksi dengan lingkungan. Ketahanannya bergantung pada input yang diberikan dan interaksi dengan agen lain setelah diterapkan, serta pada desain dan pelatihannya. Standar dan prosedur sertifikasi yang berfokus pada ketahanan sistem ini hanya akan efektif jika memperhitungkan sifat dinamis dan pembelajaran mandiri sistem AI, serta mulai mempertimbangkan bentuk pemantauan dan pengendalian yang mencakup tahap desain hingga pengembangan.

Hal ini juga ditekankan dalam prinsip-prinsip OECD (Organisasi untuk Kerja Sama dan Pembangunan Ekonomi) tentang AI, yang secara eksplisit merujuk pada perlunya pemantauan dan penilaian ancaman yang berkelanjutan terhadap sistem AI. Mengingat hal ini, mendefinisikan standar AI dalam keamanan siber yang bertujuan untuk mendapatkan kepercayaan (dan dengan demikian mengabaikan pemantauan dan pengendalian AI) berisiko. Semakin cepat kita memfokuskan standar dan prosedur sertifikasi pada pengembangan AI yang andal, dan semakin kita mengadopsi strategi 'internal', 'adversarial', dan 'selalu aktif', semakin aman aplikasi AI untuk 3R nantinya.

BAB 14

BOT DOA DAN IBADAH KEAGAMAAN DI TWITTER

Otomatisasi dalam kehidupan sosial daring telah menjadi fenomena penting yang memengaruhi interaksi masyarakat modern. Selama ini, diskusi mengenai otomatisasi daring umumnya berfokus pada isu-isu politik, ekonomi, atau sosial, seperti penyebaran disinformasi, manipulasi opini publik, dan strategi pemasaran digital. Namun, terdapat satu ranah penggunaan teknologi otomatisasi yang justru sangat signifikan tetapi relatif terabaikan oleh penelitian akademik, yaitu agama.

Dalam konteks ini, salah satu contoh menonjol adalah penggunaan Aplikasi Doa Islam yang memungkinkan unggahan doa secara otomatis dari akun media sosial penggunanya. Studi awal menunjukkan bahwa hanya melalui satu layanan tersebut, jutaan doa dipublikasikan setiap hari, sehingga membentuk bagian yang substansial dari lalu lintas Twitter berbahasa Arab. Fakta ini menegaskan bahwa praktik keagamaan daring bukan sekadar fenomena marginal, melainkan telah menjadi bagian integral dari ekosistem komunikasi digital global.

Kenyataan bahwa fenomena sebesar ini luput dari perhatian banyak peneliti memperlihatkan adanya kesenjangan dalam agenda penelitian tentang otomatisasi daring. Padahal, agama memiliki peran sosial, kultural, dan eksistensial yang sangat mendalam bagi masyarakat, sehingga otomatisasi praktik keagamaan berpotensi menghadirkan implikasi luas, baik pada level personal maupun kolektif. Dengan demikian, menempatkan agama dalam diskusi mengenai otomatisasi digital tidak hanya memperluas cakupan penelitian, tetapi juga membuka ruang pemahaman baru mengenai bagaimana teknologi membentuk, memediasi, dan bahkan merekonstruksi pengalaman keagamaan di era digital.

Bab ini akan mengulas secara kritis bagaimana praktik doa otomatis di media sosial dapat dipahami dalam kerangka lebih luas, yang mencakup isu-isu tentang industri akhirat digital, kematian daring, serta transformasi agama di ruang virtual. Dengan cara ini, pembahasan tidak hanya menawarkan deskripsi fenomena, tetapi juga membuka arah baru bagi penelitian interdisipliner di bidang otomatisasi kehidupan sosial daring.

14.1 PENDAHULUAN

Kehidupan sosial daring semakin terotomatisasi. Dari asisten virtual yang membantu tugas sehari-hari, hingga chatbot yang menyediakan pendampingan atau melestarikan kenangan anggota keluarga yang telah meninggal, industri telah dengan cepat menyadari potensi perkembangan ini. Pada saat yang sama, otomatisasi sosial daring juga digunakan untuk tujuan politik, termasuk akun "bot" otomatis di jejaring sosial yang berupaya memengaruhi pemilu dan acara politik penting lainnya. Tren ini memang telah menarik banyak perhatian, baik publik maupun komunitas riset.

Namun, satu area utama otomatisasi daring sebagian besar telah diabaikan sejauh ini: ibadah keagamaan. Dalam artikel ini, kami menyajikan analisis skala besar pertama tentang penggunaan teknologi otomatisasi daring untuk keperluan keagamaan. Lebih spesifik lagi,

Peran AI dalam Etika Tata Kelola dan Kebijakan
Dr. Ir. Agus Wibowo M.Kom., M.Si., M.M.

artikel ini berfokus pada fenomena yang sangat luas, yang kami sebut Aplikasi Salat Islam, yang, meskipun populer, sejauh ini belum disadari oleh komunitas riset. Kami berpendapat bahwa penyebaran dan signifikansi sosial dari aplikasi ini menuntut perluasan cakupan penelitian terkini tentang otomatisasi daring secara umum, dan bot media sosial secara khusus.

14.2 APLIKASI SALAT ISLAM

Dalam beberapa tahun terakhir, semakin banyak umat Muslim yang memanfaatkan layanan digital yang secara otomatis mengunggah doa atau salat tambahan ke akun media sosial pribadi mereka. Dalam kajian ini, layanan tersebut akan disebut sebagai *Aplikasi Salat Islam*. Walaupun aplikasi-aplikasi ini berbeda dalam hal model bisnis dan tingkat popularitas, tujuan utamanya serupa, yaitu memfasilitasi serta mengotomatisasi praktik keagamaan dalam bentuk doa publik. Penting untuk dicatat bahwa keberadaan aplikasi ini tidak dimaksudkan untuk menggantikan salat wajib lima waktu, melainkan lebih berfungsi sebagai sarana tambahan untuk mengekspresikan doa atau permohonan secara daring.

Bentuk awal dari layanan keagamaan digital ini biasanya hanya sebatas pengingat waktu salat lokal. Namun, *Aplikasi Salat Islam* menawarkan fungsi yang lebih luas dengan memfasilitasi permohonan publik berupa doa (*du'a*) yang dapat dipahami sebagai ungkapan harapan atau permintaan agar suatu peristiwa terjadi. Pengguna dapat menyusun doa pribadi mereka, tetapi aplikasi ini juga menyediakan beragam contoh doa dari Al-Qur'an yang bisa dipilih dan dibagikan.

Salah satu contoh yang cukup representatif adalah Du3a.org. Situs ini menampilkan kutipan Al-Qur'an dan doa populer di halaman utamanya, sekaligus mendorong pengunjung untuk membagikan konten ke berbagai platform jejaring sosial seperti Facebook, Twitter, dan Pinterest. Situs ini bahkan mengklaim telah menarik hingga 26 juta interaksi dari pengunjung. Ciri yang paling menonjol adalah tombol berlangganan layanan. Setelah pengguna mendaftar, mereka diarahkan ke Twitter untuk memberikan izin akses kepada aplikasi. Dengan persetujuan tersebut, aplikasi secara otomatis mulai memposting doa dari akun pengguna, biasanya setiap dua jam, dengan format lebih dari 140 karakter, disertai tautan ke situs dan, hingga beberapa waktu lalu, tambahan simbol seperti emoji daur ulang.



Gambar 14.1 Tangkapan layar halaman arahan Du3a

Salah satu aspek menarik dari praktik *Aplikasi Salat Islam* adalah skalanya yang masif di media sosial. Karena setiap unggahan yang dilakukan oleh Du3a.org mencantumkan URL layanan tersebut, aktivitasnya dapat dipantau melalui *Streaming API* Twitter, yang memberikan akses langsung hingga 1% dari seluruh tweet global. Dengan memanfaatkan metode ini, penelitian eksploratif pada Juni 2018 berhasil mengumpulkan data selama 48 jam. Hasilnya menunjukkan adanya sekitar 3,8 juta tweet yang memuat URL Du3a dalam periode tersebut (lihat Gambar 14.2).

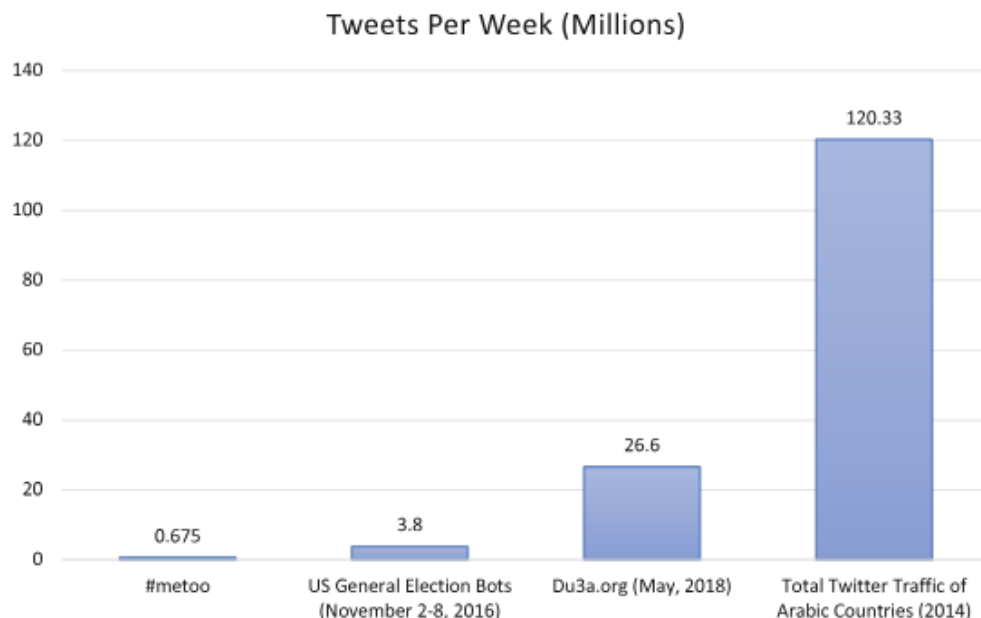
Pada saat pengamatan, Du3a tampak mengunggah satu tweet per jam dari setiap akun pengguna. Frekuensi ini kemudian menurun menjadi satu unggahan setiap dua jam. Dari distribusi geografis, sekitar 50% pengguna mengidentifikasi diri berasal dari Arab Saudi dan Mesir, sementara kontribusi negara lain hanya berkisar 1%. Fakta ini mengindikasikan bahwa fenomena doa otomatis melalui Twitter, setidaknya dalam kasus Du3a, didominasi oleh pengguna Arab.

Volume lalu lintas yang dihasilkan sangat signifikan. Dengan rata-rata 1,9 juta tweet per hari, hanya dari satu aplikasi (Du3a), jelas bahwa otomatisasi ibadah dapat menciptakan arus konten digital dalam jumlah besar. Untuk memberi perspektif, Bruns dkk. (2013) mendokumentasikan 205.000 tweet terkait tagar #egypt pada puncak Arab Spring, yaitu saat pengunduran diri Presiden Hosni Mubarak. Sementara itu, menjelang Pemilu AS 2016, Bessi dan Ferrara (2016) memperkirakan sekitar 3,8 juta tweet otomatis diproduksi selama seminggu penuh, atau rata-rata 540.000 tweet per hari. Dengan demikian, satu aplikasi doa otomatis menghasilkan hampir setara jumlah unggahan dalam dua hari dengan total konten akun otomatis menjelang pemilu AS dalam satu minggu.

Yang membedakan, Du3a tidak hanya aktif pada periode politik atau momentum khusus, melainkan beroperasi konsisten sepanjang tahun. Laporan Media Sosial Arab (2014) memperkirakan terdapat 17,19 juta tweet per hari dari seluruh pengguna di dunia Arab, sehingga wajar jika salat otomatis menyumbang proporsi yang cukup besar dalam lalu lintas Twitter di kawasan tersebut. Dari sudut pandang kuantitatif, ekspresi religius ini bahkan dapat dikategorikan sebagai salah satu fenomena paling masif di Twitter global.

Namun, ada faktor yang membuat fenomena ini kurang terekspos di ruang publik internasional. Du3a.org tidak menggunakan tagar populer yang biasanya membantu sebuah topik menjadi tren, sehingga aktivitasnya relatif tersembunyi dari perhatian pengguna non-Muslim. Perhatian luas baru muncul setelah Matthew Rothenberg (2017), pendiri *EmojiTracker.com*, menyadari bahwa emoji daur ulang yang secara konsisten disertakan Du3a di setiap tweet telah menjadi emoji paling populer ketiga di Twitter pada waktu itu, terutama berkat pengguna Muslim.

Lebih lanjut, eksplorasi atas fenomena ini menemukan setidaknya 10 situs serupa yang mengadopsi model bisnis mirip dengan Du3a. Beberapa pesaing bahkan menawarkan fitur premium yang lebih canggih. Misalnya, *Athantweets.com* menyediakan paket berbayar sebesar 100 Riyal Saudi (sekitar Rp400.000) per tahun, yang memungkinkan pengguna memilih doa tertentu, bukan yang diunggah secara acak, serta menyinkronkan tweet dengan waktu salat setempat. Selain itu, layanan premium ini menyembunyikan URL *Athantweets* pada unggahan, sehingga konten yang dipublikasikan hampir tidak dapat dibedakan dari tweet lain yang berisi ayat Al-Qur'an.



Gambar 14.2 Ilustrasi rata-rata tweet per hari

Hal ini menyoroti karakteristik penting dari fenomena ini secara keseluruhan, yaitu fakta bahwa mayoritas lalu lintas tampaknya organik; artinya, berasal dari akun-akun biasa milik

orang sungguhan, bukan "bot", yang dipahami sebagai akun dengan identitas palsu yang dibuat semata-mata untuk tujuan menyebarkan konten.

Analisis kualitatif yang cermat terhadap beberapa lusin akun Twitter yang menggunakan Du3a menunjukkan bahwa, meskipun beberapa akun tampaknya dibuat khusus untuk menggunakan layanan salat ini, sebagian besar tampaknya merupakan pengguna biasa, yang mencuitkan pesan, foto, dan komentar sehari-hari yang diselingi dengan pesan otomatis. Dengan kata lain, sebagian besar lalu lintas tampaknya dibuat oleh akun-akun autentik, yang dioperasikan oleh pengguna sah, yang secara kreatif mengotomatiskan aspek aktivitas daring mereka sambil tetap menggunakan layanan sebagaimana mestinya.

14.3 KONTEKS KEAGAMAAN

Masih terlalu dini untuk dapat menjelaskan secara pasti peran khusus yang dimainkan *Aplikasi Salat Islam* dalam kehidupan sehari-hari penggunaannya. Untuk itu, penelitian lebih lanjut sangat diperlukan, baik dari sisi kualitatif, seperti yang pernah dilakukan Bell (2006), maupun dari sisi kuantitatif. Sebagaimana dicatat oleh Akma dan Abdul Razak, meskipun terdapat hampir 1,2 miliar Muslim di seluruh dunia, kajian mengenai penerapan teknologi yang bersifat spiritual dari perspektif Islam masih sangat terbatas.

Dari pengamatan awal terhadap sejumlah akun dalam data yang terkumpul, terlihat adanya beragam motivasi yang melatarbelakangi penggunaan aplikasi doa otomatis ini. Sebagai contoh, sebuah akun menuliskan bahwa alasan pembuatannya adalah untuk mendoakan "saudara saya [nama]," yang menunjukkan bahwa sebagian pengguna memanfaatkan aplikasi ini bukan hanya untuk kepentingan pribadi, melainkan juga untuk mendoakan keluarga atau kerabat.

Salah satu fungsi inti yang ditawarkan aplikasi semacam ini adalah janji eksplisit bahwa unggahan doa akan tetap berlanjut bahkan setelah penggunaannya meninggal dunia. Misalnya, slogan **Zad-Muslim.com** berbunyi: "*Daftar sekarang agar akun Anda dapat berkicau sekarang dan setelah Anda meninggal dunia.*" Demikian pula, **Dur3a.org** menjanjikan bahwa "*akun Anda akan terus berkicau di masa hidup dan mati Anda.*" Janji ini bukan sekadar detail tambahan, melainkan berkaitan dengan aspek teologis yang cukup mendalam.

Dalam Islam, keyakinan mengenai kehidupan setelah kematian berhubungan erat dengan amal perbuatan semasa hidup. Menurut Al-Qur'an, manusia tidak langsung menerima penghakiman setelah meninggal, melainkan harus menunggu di alam kubur hingga Allah menetapkan akhir zaman dan memberikan keputusan akhir atas setiap individu. Untuk dapat masuk surga, seorang Muslim dituntut agar amal baiknya lebih banyak daripada dosanya. Akan tetapi, tradisi Islam juga mengajarkan bahwa terdapat faktor-faktor tertentu yang dapat meningkatkan pahala seorang Muslim bahkan setelah ia wafat. Nabi Muhammad menyebutkan tiga di antaranya, yaitu: amal jariyah (amal yang terus memberi manfaat), ilmu yang bermanfaat bagi generasi berikutnya, serta doa dari anak yang saleh.

Meskipun tafsir atas hal ini masih menjadi perdebatan di kalangan ulama, mayoritas ulama sepakat bahwa pahala seseorang bisa bertambah meskipun ia telah meninggal. Dari sini, muncul pemahaman bahwa penyebaran Islam melalui media, baik dilakukan langsung

oleh seseorang maupun melalui sarana yang ia tinggalkan seperti aplikasi doa otomatis dapat dikategorikan sebagai amal baik yang berkelanjutan. Dengan demikian, dapat diasumsikan bahwa mengaktifkan aplikasi doa yang terus mengunggah permohonan doa atas nama pengguna setelah kematian berpotensi meningkatkan peluangnya memperoleh kehidupan akhirat yang lebih baik.

Sejauh ini, komunitas Muslim umumnya menerima kehadiran teknologi baru dalam praktik keagamaan, khususnya ketika teknologi tersebut dipandang sebagai sarana dakwah. Campbell (2010) dalam bukunya *When Religion Meets New Media* mencatat bahwa keterlibatan dengan media kini dianggap sebagai suatu kewajiban religius, bahkan sesuatu yang akan dimintai pertanggungjawaban oleh umat Islam di akhirat. Pandangan ini menegaskan bagaimana teknologi digital, termasuk media sosial, diposisikan bukan sekadar alat, melainkan medium penting dalam praktik keagamaan kontemporer.

Maka, penyebaran Islam melalui platform daring memiliki implikasi mendalam, baik dalam konteks kehidupan maupun kematian. Namun, penelitian akademik masih sangat terbatas terkait sikap komunitas Muslim yang lebih luas terhadap fenomena ini. Demikian pula, sikap platform media sosial terhadap fenomena tersebut belum banyak diketahui. Misalnya, Facebook telah menyediakan fitur *memorialisasi* akun pengguna yang meninggal, sementara Twitter justru menghapus akun yang tidak aktif selama beberapa bulan, termasuk milik pengguna yang telah wafat.

Di sisi lain, akun yang terhubung dengan *Aplikasi Salat Islam* tetap aktif meskipun pemiliknya telah meninggal. Dengan demikian, akun tersebut tidak akan dikategorikan sebagai nonaktif atau ditinggalkan. Hal ini berarti jumlah doa dan konten Islam di Twitter dapat terus meningkat bahkan setelah sebagian besar penggunanya wafat. Menariknya, konsep serupa juga muncul dalam konteks sekuler di Barat melalui aplikasi yang berupaya “*mengabadikan jejak media sosial seseorang*”. Namun, jika dibandingkan, proyek-proyek tersebut masih terbelang pinggiran dan kurang populer, berbeda dengan *Aplikasi Doa Islam* yang telah menjadi fenomena luas di dunia Muslim.

Mengacu pada popularitasnya yang kian besar, bukan tidak mungkin bahwa dalam beberapa dekade mendatang akan semakin sulit membedakan lalu lintas media sosial yang dihasilkan oleh pengguna hidup dengan yang sudah wafat. Uniknya, hal ini bukan dipicu oleh eksperimen futuristik dari Silicon Valley, melainkan oleh praktik religius jamaah Muslim di seluruh dunia.

14.4 IMPLIKASI YANG LEBIH LUAS

Agama sejak dahulu selalu menjadi salah satu aspek paling penting dalam kehidupan manusia, baik pada dimensi individu maupun sosial. Kini, dengan perkembangan teknologi, agama juga muncul sebagai faktor utama dalam mendorong adopsi teknologi digital. Islam, misalnya, bukanlah satu-satunya agama yang mengintegrasikan teknologi otomatisasi ke dalam praktik ibadah. Di iTunes App Store saja, terdapat lebih dari 6000 aplikasi terkait spiritualitas dan agama. Ini menunjukkan betapa luasnya keterkaitan antara iman dan inovasi teknologi.

Fenomena ini tidak terbatas pada Islam. Pada awal 2019, Paus Fransiskus meluncurkan aplikasi Click to Pray, yang memungkinkan umat Katolik berpartisipasi dalam doa-doa Paus sekaligus membagikannya melalui media sosial. Gereja Inggris pun memanfaatkan teknologi dengan menghadirkan fitur aktivasi suara untuk perangkat Alexa milik Amazon. Dengan fitur ini, pemilik perangkat dapat meminta Alexa membacakan doa harian. Bahkan, praktik tradisional seperti menyalakan lilin di gereja kini kerap digantikan dengan lilin listrik. Dalam kapasitas tertentu, teknologi mulai mengambil alih sebagian peran rohaniwan dengan memberikan jawaban atas pertanyaan keagamaan mendasar, misalnya tentang siapa Tuhan atau makna percaya kepada Kristus.

Komunitas Yahudi juga memperlihatkan tren serupa dengan penggunaan teknologi otomatisasi untuk mengatur fasilitas rumah selama Sabat. Selain itu, perusahaan rintisan seperti Prayerbot memungkinkan orang dari berbagai latar agama untuk mengajukan permohonan doa melalui chatbot. Berbagai contoh ini memperlihatkan bahwa agama bukanlah kekuatan kecil dalam proses otomatisasi sosial kontemporer, melainkan aktor penting yang ikut mengarahkan arah perkembangannya.

Dalam konteks Islam, Aplikasi Doa Islam (Islamic Prayer Apps) menghadirkan dimensi baru dalam wacana bot Twitter dan kebijakan otomatisasi daring. Selama ini, literatur mengenai bot sosial banyak menyoroti penggunaannya dalam politik, misalnya untuk menyebarkan propaganda atau meningkatkan metrik keterlibatan selama pemilu. Namun, berbeda dengan itu, Aplikasi Doa Islam tidak terlibat dalam penyebaran tagar politik atau kampanye sosial, melainkan beroperasi lebih tersembunyi, berfokus pada praktik ibadah personal maupun kolektif.

Bahkan, fenomena ini mempersulit wacana akademik yang selama ini melihat bot dan non-bot sebagai dua kategori yang terpisah dengan jelas. Banyak pengguna Twitter menggunakan layanan otomatisasi seperti *Tweetdeck* atau *If This Then That* (IFTTT) untuk sebagian aktivitas mereka, sehingga batas antara "manusia" dan "otomatisasi" semakin kabur. Hal ini mengundang pertanyaan baru: bagaimana seharusnya perilaku tersebut dipahami, dan apa konsekuensi etis yang menyertainya? Sejauh ini, media sosial otomatis telah digunakan secara kreatif di berbagai bidang. Misalnya, jurnalis memanfaatkannya untuk publikasi cepat, aktivis antikorupsi menggunakannya untuk kampanye, dan kelompok pro-transparansi memanfaatkannya dalam advokasi. Namun, penggunaan otomatisasi dalam konteks keagamaan relatif masih jarang diteliti, padahal justru membuka ruang diskusi baru yang sangat signifikan.

Aplikasi Salat Islam yang diteliti di sini bisa dikatakan sebagai salah satu contoh terbesar dan paling organik dari otomatisasi Twitter dengan makna budaya yang mendalam. Fakta bahwa fenomena ini hampir tidak diperhatikan oleh peneliti menunjukkan adanya keterbatasan perspektif akademik yang cenderung sekuler. Seperti dikatakan Bell, *"Kita tampaknya keras kepala sekuler dalam membayangkan konteks komputasi di rumah dan waktu luang."* Karena itu, sudah saatnya percakapan akademik diperluas untuk memasukkan peran agama dalam teknologi digital tidak hanya sebagai objek sampingan, tetapi sebagai kekuatan utama yang ikut membentuk lanskap otomasi sosial kontemporer.

BAB 15

KECERDASAN BUATAN, DEEPPAKE, DAN MASA DEPAN EKTIBE

Bab ini memperkenalkan konsep “*ek tipe digital*”. Ektipe dapat dipahami sebagai salinan yang memiliki keterhubungan istimewa dengan sumber aslinya (*arketipe*), layaknya jejak yang ditinggalkan oleh sebuah segel. Ektipe bukanlah benda asli, namun tetap memancarkan hubungan autentik dengan asal mula penciptaannya. Dalam konteks teknologi digital, saya berpendapat bahwa ada pemisahan baru antara sumber arketipe misalnya gagasan yang berada dalam benak seniman dan proses kreatif yang mengubah gagasan itu menjadi artefak. Ketika tautan tersebut terputus, muncullah *ek tipe digital* yang bisa “asli” dalam gaya atau konten, tetapi tidak “asli” dalam hal sumber arketipe; atau sebaliknya, “asli” dalam kaitannya dengan sumber arketipe, tetapi tidak “asli” dalam metode dan proses penciptaannya. Dengan demikian, ektipe digital dapat berwujud sebagai artefak yang autentik namun tidak orisinal, atau artefak yang orisinal namun tidak autentik.

Dunia seni sejak lama bergulat dengan isu reproduksi. Beberapa reproduksi hadir dalam bentuk replika murni seperti berbagai salinan lukisan *Mona Lisa* yang tersebar di seluruh dunia. Lainnya hadir dalam wujud palsu atau tiruan, misalnya lukisan “Vermeer” yang dilukis oleh Han van Meegeren, yang sempat terjual dengan nilai fantastis sekitar Rp960 miliar. Pada titik ini, perbedaan mendasar antara “replika” dan “palsu” bergantung pada konsep keaslian: apakah artefak tersebut sesuai dengan klaimnya?

Jawaban atas pertanyaan itu ternyata tidak sesederhana yang dibayangkan. Saat ini, karya-karya pemalsu John Myatt justru dihargai tinggi hingga sekitar Rp640 juta per lukisan. Ironisnya, nilai itu diberikan bukan karena lukisan-lukisan tersebut “asli” dalam arti berasal dari seniman yang ditirunya, melainkan karena ia dianggap sebagai “pemalsuan autentik” hasil tangan Myatt sendiri. Lukisan-lukisan ini seperti sebuah paradoks: mereka bukanlah karya yang diklaim, tetapi tetap dihargai sebagai karya indah, bahkan disebut sebagai “kebohongan yang indah” alih-alih kebohongan biasa.

Fenomena serupa juga terjadi pada seniman besar seperti René Magritte. Di tengah pendudukan Nazi di Belgia, ia tidak hanya membuat tiruan gaya Picasso atau Renoir (Marien, 1983), tetapi juga pernah “memalsukan dirinya sendiri.” Salah satu contoh paling terkenal adalah kasus dua salinan lukisan *The Flavour of Tears* (1948), yang keduanya dilukis oleh Magritte. Salah satunya ia ciptakan sebagai “palsu” dari karyanya sendiri sekaligus sebagai eksperimen surealis dan upaya pragmatis untuk memperoleh uang.

Kekacauan semacam ini menunjukkan bahwa keaslian tidak pernah sepenuhnya sederhana. Ia selalu berada dalam tarik-menarik antara klaim, niat, proses, dan penerimaan sosial. Kini, dengan kehadiran teknologi digital, persoalan keaslian menjadi semakin kompleks. Perkembangan seperti *deepfake* dan kecerdasan buatan mampu menghadirkan “salinan” yang sangat autentik dalam gaya dan rupa, meski tidak pernah bersentuhan dengan arketipe asli yang melahirkannya. Di sinilah konsep *ek tipe digital* menjadi penting: ia membuka ruang baru

bagi pemahaman tentang salinan yang mungkin autentik namun tidak orisinal, atau orisinal namun tidak autentik.



Gambar 15.1 Palsu, aslinya adalah “Lodge on Lake Como” karya Carl Frederik Peder Aagaard (1833–1895)

Berkat kemajuan teknologi digital, kini memastikan keaslian suatu karya seni jauh lebih mudah dibandingkan masa lalu. Tersedia berbagai basis data yang memungkinkan pemeriksaan tanda tangan seniman, serta jutaan gambar yang bisa diakses hanya dengan beberapa klik. Kondisi ini membuat penjualan karya palsu menjadi semakin sulit. Sebagai contoh, Gambar 17.1 menampilkan reproduksi *Lodge on Lake Como* karya Carl Frederik Peder Aagaard (1833–1895), seorang pelukis lanskap sekaligus seniman dekoratif asal Denmark. Lukisan tersebut sempat dijual di eBay pada tahun 2016. Popularitasnya di internet membuat karya ini memiliki banyak replika yang cukup berkualitas, dan hal itu sebenarnya tidak bermasalah.

Namun, ketika diteliti lebih cermat, lukisan tersebut ternyata dijual sebagai “asli” meskipun tidak memiliki tanda tangan sebuah klaim yang jelas menyesatkan. Baik kualitas maupun harga yang ditawarkan tampak mencurigakan, dan dengan melakukan pencarian gambar sederhana di Google, dengan cepat dapat dipastikan bahwa karya itu hanyalah replika. Saat artikel ini ditulis, lukisan tersebut sudah tidak tersedia, dan akun penjualnya juga tampak tidak lagi aktif di eBay.

Tentu saja, barang palsu tidak selalu berupa reproduksi karya lama. Ada pula “karya baru” yang diklaim berasal dari seniman terkenal, seperti Pollock atau Van Gogh. Untuk kasus semacam ini, berbagai teknik ilmiah modern digunakan untuk memastikan keasliannya, termasuk metode berbasis kecerdasan buatan (AI). Misalnya, sebuah penelitian yang

diterbitkan pada November 2017 oleh Ahmed Elgammal, Yan Kang, dan Milko Den Leeuw mengusulkan pendekatan komputasional untuk menganalisis goresan pada gambar karya seni, menggunakan jaringan saraf tiruan.

Dataset penelitian tersebut mencakup 300 gambar digital dengan lebih dari 80.000 goresan, karya seniman ternama seperti Pablo Picasso, Henry Matisse, Egon Schiele, dan beberapa seniman lain. Dengan melakukan segmentasi terhadap setiap goresan, sistem AI belajar mengenali karakteristik unik yang dimiliki masing-masing seniman. Hasilnya, perangkat lunak mampu mengklasifikasikan goresan dengan tingkat akurasi 70%–90%, dan mengidentifikasi keseluruhan gambar dengan akurasi lebih dari 80%. Lebih jauh lagi, sistem ini terbukti tangguh dalam mendeteksi pemalsuan, dengan akurasi mencapai 100% pada sebagian besar pengaturan pengujian. Temuan ini menunjukkan bahwa gaya menggambar seorang seniman sama uniknya dengan sidik jari atau gaya berjalan seseorang, dan AI dapat membantu mengungkapnya layaknya sebuah mikroskop canggih.

Meski begitu, AI tidak hanya dimanfaatkan untuk mengidentifikasi pemalsuan. Di Belanda, Microsoft bekerja sama dengan Museum Rumah Rembrandt meluncurkan sebuah proyek inovatif yang menghasilkan potret seorang pria sebuah karya yang bisa disebut sebagai Rembrandt, sekaligus bukan. Algoritma menganalisis karya-karya Rembrandt yang otentik, lalu mengidentifikasi subjek paling umum (misalnya potret pria Kaukasia berusia 30–40 tahun), ciri khas yang sering muncul (seperti rambut wajah, arah pandangan ke kanan, topi, kerah, pakaian gelap, dan seterusnya), serta gaya lukisan yang khas. Berdasarkan analisis ini, sistem AI merekonstruksi sebuah lukisan baru dengan menggabungkan semua elemen khas tersebut.

Hasil karya itu kemudian dicetak menggunakan printer 3D untuk meniru kedalaman, tekstur, dan lapisan warna, sehingga menyerupai teknik melukis Rembrandt sedekat mungkin. Hasil akhirnya adalah sebuah mahakarya yang tampak otentik sebuah Rembrandt yang sejatinya tidak pernah dilukis oleh Rembrandt sendiri. Lukisan ini menantang konsep tradisional tentang “keaslian” dan “orisinalitas,” mengingat adanya hubungan yang begitu kuat dengan sang seniman besar.

Nilai jual lukisan tersebut memang tidak diketahui secara pasti. Namun, jika dilelang dengan klaim keaslian yang sah, besar kemungkinan harganya akan sangat tinggi, mungkin setara dengan lukisan Rembrandt asli bernilai puluhan hingga ratusan miliar rupiah. Persoalannya, kita tidak memiliki istilah yang tepat untuk mendeskripsikan artefak semacam ini. Karena itu, saya mengusulkan istilah *ektipe*.

Kata *ektipe* berasal dari bahasa Yunani dan memiliki nuansa makna yang sangat relevan. Ektipe bukan sekadar salinan biasa, melainkan salinan yang memiliki hubungan istimewa dengan sumber atau *arketipe*. Secara khusus, ektipe bisa dipahami sebagai kesan yang ditinggalkan oleh sebuah segel. Ia bukan benda aslinya, namun tetap berhubungan erat dan autentik dengan benda asli itu sendiri. Filsuf John Locke, misalnya, menggunakan istilah “*ectypes*” untuk menyebut ide atau kesan yang berhubungan dengan realitas eksternal (arketipe), meskipun hubungan itu tidak sepenuhnya memadai.

Dalam kerangka teknologi digital, konsep ini menjadi semakin penting. Menurut Floridi (2017), teknologi mampu memisahkan arketipe misalnya ide murni dalam pikiran seorang

seniman dari proses kreatif berupa gaya, metode, atau prosedur yang menghasilkan sebuah artefak. Jika pemisahan ini terjadi, kita bisa mendapatkan *ek tipe* yang “asli” dalam gaya dan konten, tetapi tidak “asli” dari segi arketipe seperti halnya Rembrandt buatan Microsoft. Sebaliknya, bisa juga ada *ek tipe* yang benar-benar “asli” dari segi arketipe (misalnya memang berasal dari sumber otentik), tetapi tidak “asli” dalam metode dan proses produksinya.

Dengan kata lain, *ek tipe* bisa hadir sebagai artefak yang autentik namun tidak orisinal, atau artefak yang orisinal namun tidak autentik. Dalam dunia seni digital, kategori baru inilah yang semakin relevan untuk menjelaskan karya-karya yang lahir dari persilangan kreativitas manusia dan teknologi.



Gambar 15.2 Rembrandt yang bukan Rembrandt. Microsoft Project dengan Museum Rumah Rembrandt

Sebuah contoh yang bagus dari *ek tipe* asli yang tidak autentik diberikan pada bulan Maret oleh rekaman audio pidato terakhir John F. Kennedy. Meskipun merupakan pidato biasa dari jalur kampanye yang sudah berusia puluhan tahun, pidato itu tiba-tiba menjadi berita utama. Karena itu adalah pidato Dallas Trade Mart pada 22 November 1963, teks yang seharusnya dibacakan JFK, seandainya ia tidak dibunuh beberapa saat sebelumnya, dalam perjalanan untuk menyampaikannya. Teksnya asli: berasal dari sumbernya.

Namun, suara yang dibacakan tidak autentik, karena disintesis oleh perangkat lunak yang menganalisis 831 rekaman pidato dan wawancara Kennedy, untuk "belajar" berbicara seperti dirinya. Perangkat lunak tersebut akhirnya dapat menyuarakan pidato terakhir JFK 55 tahun kemudian. Jadi, inilah Kennedy yang merupakan Kennedy dan bukan Kennedy, serupa namun berbeda dari Rembrandt yang merupakan Rembrandt dan bukan. Keduanya merupakan *ek tipe* (lihat Tabel 15.1).

Tabel 15.1 Arketipe, palsu dan ektipe

Karya / Produksi	Sumber Asli	Produksi Otentik
<i>Mona Lisa karya Leonardo</i>	Ya	Ya
<i>Lukisan palsu Vermeer karya Han van Meegeren</i>	Tidak	Tidak
<i>Rembrandt versi Microsoft (AI Rembrandt)</i>	Tidak	(Berkualifikasi) Ya
<i>Pidato Trade Mart JFK</i>	Ya	Tidak

Kita melihat bahwa produksi ektipe tidak berhenti pada karya seni, tetapi melibatkan artefak apa pun, mulai dari teks hingga foto, dari rekaman audio hingga video. Sudah diketahui umum bahwa sejarah manuskrip, percetakan, fotografi, sinema, dan televisi dipenuhi dengan kepalsuan. Nantikan juga lebih banyak ektipe. Khususnya, para seniman gemar mendobrak batasan dan mudah dibayangkan bahwa, seperti Magritte yang memalsukan lukisannya sendiri, mereka akan mulai memproduksi ektipe mereka sendiri. Bayangkan seorang pelukis menggunakan perangkat lunak yang dikembangkan oleh Microsoft untuk menghasilkan karya-karya barunya sendiri. Itu tetaplah sebuah ektipe, dan ini akan menjelaskan mengapa (dengan kualifikasi) proses tersebut akan menangkap keaslian. Reproduksi karya seni dengan cara mekanis akan memperoleh makna baru.

Dengan ektipe, kita biasanya tahu di mana letak suatu hal. Namun, seseorang bisa saja berbuat curang. Mei lalu, Google memperkenalkan Google Duplex, versi asisten AI-nya yang mensimulasikan kehidupan manusia untuk membantu pengguna dengan tugas-tugas interaktif sederhana, seperti memesan meja restoran. Perusahaan tersebut dengan cepat menyatakan bahwa mereka tidak akan sengaja menyesatkan siapa pun, dan akan selalu memastikan untuk mengklarifikasi ketika pengguna berinteraksi dengan agen buatan. Namun, orang lain dapat menggunakan teknologi ini untuk tujuan kriminal atau jahat. Inilah yang terjadi dengan Deepfake, serangkaian teknik yang digunakan untuk mensintesis produk visual baru, misalnya dengan mengganti wajah pada gambar asli.

Kasus-kasus umum melibatkan film porno di mana wajah aktris terkenal seperti Gal Gadot atau Scarlett Johansson (ini biasanya tentang wajah perempuan) digunakan untuk menggantikan wajah asli. Dalam kasus ini pun, basis data besar diperlukan untuk menginstruksikan perangkat lunak (yang tersedia gratis, dan juga tersedia dalam bentuk aplikasi), jadi jika Anda bukan figur publik, risikonya lebih rendah. Deepfake juga menyangkut politik, seperti Presiden Obama, misalnya.

Bagaimana masa depan kita? Teknologi digital tampaknya meruntuhkan kepercayaan kita pada keaslian, keaslian, dan autentik dari apa yang kita lihat dan dengar. Namun, apa yang dirusak oleh teknologi digital juga dapat diperbaiki, tidak seperti pertarungan tanpa akhir antara virus perangkat lunak dan antivirus. Dalam kasus kami, selain mengedukasi masyarakat, memperoleh kepekaan baru, dan memiliki kerangka hukum yang tepat, setidaknya terdapat beberapa strategi digital yang menarik. Untuk artefak yang sudah tersedia, mudah dibayangkan.

Sistem AI yang membantu kita. Akan menarik untuk menganalisis pidato Rembrandt dan Kennedy karya Microsoft dengan sistem buatan untuk melihat apakah sistem tersebut mendeteksi ektipe. Penelitian tentang metode untuk mengekspos video Deepfake yang dihasilkan dengan jaringan saraf sudah tersedia. Singkatnya, mari kita ingat perangkat lunak yang dikembangkan untuk menganalisis gambar: ada banyak alat canggih untuk mendeteksi pemalsuan gambar.

Dan kemungkinan akan lebih banyak lagi yang dikembangkan seiring meningkatnya permintaan. Selanjutnya, terkait artefak baru, karena orisinalitas dan keaslian juga merupakan masalah kontinuitas historis yang dapat dibuktikan dari sumber hingga produk melalui proses produksi, blockchain yang sangat digembar-gemborkan, atau solusi serupa, dapat membuat perbedaan besar. Blockchain bagaikan sebuah register yang menyimpan transaksi dengan cara yang akrual, aman, transparan, dan terlacak.

Sebagai sebuah register transaksi yang aman dan terdistribusi, blockchain sedang dieksplorasi sebagai sarana untuk mensertifikasi asal-usul dan sejarah produk-produk tertentu secara andal: baik dalam hal mengamankan rantai pasok pangan, maupun dalam mencatat berbagai tindakan penciptaan dan kepemilikan yang saling terkait yang menentukan asal-usul sebuah karya seni.

Di masa depan, kita mungkin akan mengadopsi solusi yang sama di mana pun ada kebutuhan untuk memastikan (atau menetapkan) orisinalitas dan autentisitas suatu artefak, baik berupa dokumen tertulis, foto, video, maupun lukisan. Dan tentu saja, seorang seniman masa depan mungkin ingin memastikan, melalui blockchain, bahwa karya seninya sebagai sebuah ektipe benar-benar sesuai dengan apa yang dikatakannya. Pada titik itu, kita akan kembali ke titik awal, karena kita akan memiliki "ektipe sejati", seperti Rembrandt karya Microsoft, atau pidato Kennedy.

15.1 ETIKA AI DALAM PELAYANAN KESEHATAN

Perkembangan kecerdasan buatan (AI) dalam bidang pelayanan kesehatan membuka peluang besar untuk meningkatkan kualitas layanan, efisiensi sistem, serta menurunkan biaya operasional. Namun, bersamaan dengan potensi manfaat tersebut, muncul pula berbagai persoalan etika yang kompleks. Isu-isu etika ini penting diperhatikan agar penerapan AI dalam layanan kesehatan tidak justru menimbulkan kerugian, seperti hilangnya kepercayaan publik, ketidakadilan dalam distribusi manfaat, atau bahkan kegagalan sistem kesehatan secara luas.

Penelitian-penelitian sebelumnya menunjukkan bahwa tantangan etika dalam AI-kesehatan dapat dikategorikan dalam tiga ranah utama. Pertama, isu epistemik, yang berkaitan dengan penggunaan data dan bukti, termasuk risiko misinformasi, kualitas data yang lemah, serta keterbatasan transparansi algoritme. Kedua, isu normatif, yang mencakup persoalan keadilan, bias, eksklusi sosial, dan efek transformasi sosial akibat penggunaan teknologi AI. Ketiga, isu ketertelusuran, yang menekankan pentingnya akuntabilitas, transparansi, dan kemampuan untuk melacak proses pengambilan keputusan oleh sistem berbasis AI.

Lebih jauh, isu-isu tersebut dapat muncul di berbagai tingkatan, mulai dari level individu, interpersonal, kelompok, institusional, hingga masyarakat atau sektoral. Hal ini

memperlihatkan bahwa dampak etika AI tidak hanya bersifat personal, tetapi juga menyentuh dimensi struktural dan sosial yang lebih luas. Misalnya, bias data bukan hanya memengaruhi pasien secara individual, tetapi juga dapat menghasilkan ketidakadilan sistemik dalam akses layanan kesehatan.

Tinjauan literatur komprehensif terhadap 156 publikasi ilmiah mengenai etika AI-kesehatan mengungkap setidaknya 11 pertimbangan utama yang perlu diperhatikan oleh para pembuat kebijakan, regulator, dan pengembang. Pertimbangan ini berkaitan dengan isu berulang dalam literatur, termasuk bias data, kualitas data, inklusi sosial, perlindungan kesetaraan perawatan, distribusi manfaat yang adil, serta pelestarian nilai-nilai sosial dalam layanan kesehatan. Meskipun beberapa solusi teknis telah diusulkan untuk mengatasi permasalahan tersebut, banyak di antaranya masih bersifat eksperimental dan belum terbukti efektif dalam praktik nyata.

Jika masalah-masalah etika ini tidak segera ditangani secara serius, terdapat risiko terjadinya “musim dingin AI” baru, yaitu fase penurunan kepercayaan dan penggunaan AI akibat ekspektasi yang gagal terpenuhi serta dampak negatif yang ditimbulkannya. Oleh karena itu, penelitian ini berupaya memberikan kerangka konseptual yang dapat memandu diskusi, dengan membedakan isu-isu etika berdasarkan jenisnya (epistemik, normatif, ketertelusuran) dan tingkat abstraksinya (individu hingga sektoral).

Dengan demikian, bab ini bertujuan untuk memberikan dasar pemahaman yang kuat mengenai dilema etika AI dalam pelayanan kesehatan. Diskusi ini diharapkan dapat membantu para akademisi, praktisi, pembuat kebijakan, dan regulator dalam merumuskan strategi pengembangan serta penerapan AI yang lebih etis, adil, dan dapat dipercaya, sehingga manfaat teknologi dapat dimaksimalkan tanpa mengorbankan nilai-nilai fundamental dalam pelayanan kesehatan modern.

Sistem layanan kesehatan global saat ini tengah menghadapi tekanan serius berupa meningkatnya biaya operasional serta menurunnya kualitas hasil kesehatan masyarakat. Kondisi ini melahirkan apa yang disebut sebagai *masalah pelik*, yaitu persoalan yang memiliki banyak penyebab, sulit dipahami maupun didefinisikan, serta membutuhkan pendekatan multidisipliner untuk menanganinya. Dalam konteks ini, semakin banyak pemangku kepentingan mulai dari pembuat kebijakan, politisi, wirausahawan klinis, hingga ilmuwan komputer dan data yang menilai bahwa Kecerdasan Buatan (AI), khususnya Pembelajaran Mesin (*Machine Learning*), merupakan salah satu elemen kunci dalam menemukan solusi.

Namun, argumen ini tidak didasarkan pada asumsi bahwa dalam waktu dekat semua kebutuhan layanan kesehatan akan diambil alih oleh “dokter robot”. Sebaliknya, AI dipahami sebagai kumpulan teknik yang memungkinkan mesin menyelesaikan tugas dengan cara yang dianggap cerdas apabila dilakukan oleh manusia. Berbagai penelitian telah menunjukkan potensi AI dalam dunia medis: pohon keputusan untuk mendiagnosis kanker payudara, *Support Vector Machines* (SVM) untuk klasifikasi gen maupun diagnosis diabetes, metode pembelajaran *ensemble* untuk memprediksi hasil pengobatan kanker, hingga jaringan saraf tiruan untuk diagnosis stroke.

Dengan demikian, AI merepresentasikan sebuah agen teknologi interaktif, otonom, dan adaptif seringkali dengan kemampuan belajar mandiri yang dapat digunakan sesuai kebutuhan. Hal ini membuka peluang kolaborasi transformatif antara manusia dan mesin, khususnya antara klinisi dan sistem AI. Jika dimanfaatkan secara efektif, kerja sama ini dapat meningkatkan mutu layanan kesehatan, mempercepat diagnosis, mendorong penemuan obat, memajukan epidemiologi, mengembangkan pengobatan yang dipersonalisasi, serta meningkatkan efisiensi operasional sistem kesehatan. Kendati demikian, integrasi AI dalam praktik klinis membutuhkan kerangka tata kelola yang jelas untuk mencegah risiko penyalahgunaan atau perilaku tidak etis.

Untuk itu, para pembuat kebijakan perlu memperhatikan risiko etika yang muncul dari penerapan AI dalam layanan kesehatan. Tugas utama kajian ini adalah mengklasifikasikan risiko-risiko etis tersebut, menghubungkannya dengan pertanyaan kebijakan yang relevan, serta memberikan contoh tindakan yang dapat dilakukan lembaga kesehatan guna merumuskan kerangka tata kelola yang tepat. Jika risiko etika ini diabaikan, terdapat potensi kerugian besar berupa penolakan sosial, pembentukan regulasi yang tidak proporsional, bahkan hambatan terhadap kemajuan ilmu data.

Beberapa isu etis yang relatif jelas meliputi privasi data pasien, efektivitas teknologi, aksesibilitas layanan, serta manfaat praktisnya. Namun, ada pula isu yang lebih rumit, seperti keterbatasan transparansi proses AI, problem akuntabilitas, dan potensi bias yang dapat menimbulkan diskriminasi atau kesalahan dalam pengambilan keputusan. Untuk mengidentifikasi dan menanganinya, diperlukan masukan lintas disiplin, melibatkan kedokteran, ilmu komputer, hukum, ekonomi, hingga ilmu sosial. Sayangnya, penelitian saat ini masih sering terjebak dalam silo disiplin, terlalu fokus pada individu, atau gagal mempertimbangkan variasi etis berdasarkan tahap pengembangan algoritma.

Meskipun teknologi AI-Health masih berada pada tahap awal dan belum memberikan dampak besar di garis depan layanan klinis, peluang untuk membangun kerangka etika yang kokoh masih terbuka. Namun, peluang ini semakin menyempit seiring percepatan persetujuan penggunaan solusi AI dalam layanan kesehatan, khususnya di Amerika Serikat. Bahkan, pemerintah Tiongkok dan Inggris telah secara terbuka menyatakan komitmen investasi besar-besaran dalam adopsi AI-Health.

Dengan latar belakang tersebut, tujuan utama kajian ini adalah menyajikan tinjauan lintas disiplin mengenai potensi implikasi etis dari penerapan AI-Health. Fokusnya adalah mendukung diskusi kebijakan untuk mengarahkan desain yang lebih bertanggung jawab, strategi penerapan yang transparan, serta praktik yang akuntabel dalam kerangka etika digital. Artinya, perhatian khusus diberikan pada aspek moral yang berkaitan dengan data, algoritma, serta praktik penggunaannya, sehingga pemerintah maupun sistem layanan kesehatan dapat mengadopsi teknologi AI secara lebih etis dan berkelanjutan.

15.2 METODOLOGI

Metodologi tinjauan pemetaan digunakan untuk menemukan literatur dari lintas disiplin ilmu yang menyoroti isu-isu etika yang unik terkait penggunaan algoritma AI dalam

layanan kesehatan. Jenis tinjauan ini digunakan untuk memetakan dan mengkategorikan literatur yang ada mengenai topik tertentu (dalam hal ini etika AI) dan mengontekstualisasikan temuan dalam literatur yang lebih luas. Metodologi tinjauan pemetaan ini dikembangkan oleh Evidence for Policy and Practice Information and Co-ordinating Centre di London untuk menawarkan kepada para pembuat kebijakan, praktisi, dan peneliti cara yang eksplisit dan transparan untuk mengidentifikasi pertanyaan tinjauan yang lebih sempit terkait kebijakan dan praktik.

Karena tujuan kami adalah untuk mendukung diskusi kebijakan dan dengan isu-isu ini mengarahkan pengembangan praktik desain yang lebih baik, serta strategi penerapan yang transparan dan akuntabel, metodologi ini merupakan yang paling tepat. Pertanyaan tinjauan kami berfokus pada "bagaimana risiko etika utama yang ditimbulkan oleh AI-kesehatan dapat dikategorikan, dan apa yang harus dipertimbangkan oleh para pembuat kebijakan, regulator, dan pengembang agar 'berhati-hati secara etika'?" Kami berfokus pada pengkategorian isu-isu untuk memfasilitasi penelitian dan diskusi di masa mendatang.

Kami memilih lima basis data literatur yang relevan dengan isu-isu ini dan yang berada di persimpangan literatur teknis, medis, etika, dan ilmu sosial: Scopus, Google Scholar, Philpapers, Web of Science, Pub Med. Pencarian tinjauan literatur kami dilakukan pada bulan April 2019, dengan referensi yang ditambahkan atau dihapus selama iterasi penyusunan. Mesin pencari tidak identik, sehingga kami menggunakan variasi dari rangkaian istilah pencarian generik berikut: etika* DAN algoritma* ATAU AI* ATAU "Kecerdasan Buatan"* ATAU "Pembelajaran Mesin"* DAN kesehatan* (lihat Lampiran untuk detail hasil dan kueri pencarian). Hasil awal disaring berdasarkan judul. Hasil yang dianggap relevan diunduh agar abstrak dan kata kunci dapat ditinjau relevansinya.

Pada tahap ini, kami mengecualikan hasil apa pun yang berfokus pada isu-isu terkait terhadap kesehatan digital secara umum (misalnya, berbagi data, akses data, privasi data, pengawasan/dorongan, persetujuan, kepemilikan data kesehatan, bukti efikasi) agar tetap fokus pada pemetaan perdebatan terkini tentang etika AI secara khusus. Catatan yang telah diketahui sebelumnya oleh penulis, dan relevan dengan pertanyaan penelitian tetapi tidak disertakan dalam pencarian basis data awal, juga ditambahkan.

Untuk memastikan fokus tetap pada isu-isu etika yang unik, peta yang dikembangkan oleh, tentang isu-isu etika epistemik, normatif, dan menyeluruh terkait algoritma digunakan sebagai dasar. Tipologi yang ditawarkan oleh Mittelstadt dkk. mengidentifikasi masalah-masalah yang berkaitan dengan pengambilan keputusan algoritmik dan kemungkinan penyebabnya, seperti kesalahan input atau output yang diskriminatif. Ketertelusuran muncul dari kompleksitas sistem ketika semua bagian disatukan. Tipologi ini akan direferensikan silang dengan setiap tingkat abstraksi (LoA) yang kami usulkan di bawah ini.

Pertama, literatur terpilih ditinjau untuk mengidentifikasi contoh-contoh layanan kesehatan dari setiap isu yang disorot. dalam peta asli, seperti yang ditunjukkan pada Tabel 15.2, dan kemudian ditinjau lebih menyeluruh untuk mengidentifikasi bagaimana isu-isu etika dapat bervariasi tergantung pada apakah analisis sedang dilakukan di: (i) individu, (ii) interpersonal, (iii) kelompok (misalnya keluarga atau populasi), (iv) kelembagaan, (v) sektoral,

dan/atau (v) tingkat abstraksi masyarakat (LoA). LoA dapat dibayangkan sebagai antarmuka yang memungkinkan seseorang untuk mengamati beberapa aspek dari sistem yang dianalisis, sementara membuat aspek lain buram atau bahkan tidak terlihat. LoA umum dalam ilmu komputer, di mana sistem dijelaskan pada LoA yang berbeda (komputasi, perangkat keras, berpusat pada pengguna, dll.).

Perhatikan bahwa LoA dapat digabungkan dalam set yang lebih kompleks, dan dapat, tetapi tidak harus hierarkis, dengan 'resolusi' atau granularitas informasi yang lebih tinggi atau lebih rendah. Ini membantu tinjauan menghindari fokus sempit pada dampak tingkat individu yang disorot dalam pendahuluan. Pendekatan ini tidak dimaksudkan untuk menyiratkan bahwa tidak ada tumpang tindih antara tingkatan.

Tabel 15.2 Ringkasan mengenai kekhawatiran epistemik, normatif dan etika menyeluruh terkait penggunaan algoritma dalam layanan kesehatan berdasarkan Mittelstadt et al. (2016) dari (Disunting untuk anonimitas)

	Kekhawatiran Etis	Penjelasan	Contoh medis
Kekhawatiran epistemik	Tidak Meyakinkan Bukti	Hasil algoritmik (misalnya klasifikasi) bersifat probabilistik dan tidak sempurna. Hasil tersebut jarang cukup untuk menyatakan adanya hubungan kausal.	Pembaca EKG di jam tangan pintar dapat 'mendiagnosis' pasien menderita aritmia padahal hal ini mungkin disebabkan oleh kesalahan pada jam tangan yang tidak dapat membaca detak jantung pengguna secara akurat (misalnya karena warna kulit mereka) atau 'norma' yang dikalibrasi tidak tepat untuk individu tersebut.
	Bukti yang Tak Terpenuhi Bukti	Penerima keputusan algoritmik sangat jarang memiliki pengawasan penuh atas data yang digunakan untuk melatih atau menguji algoritma atau titik data yang digunakan untuk mencapai keputusan tertentu.	Sistem pendukung keputusan klinis yang diterapkan di rumah sakit dapat memberikan rekomendasi pengobatan, namun mungkin tidak jelas atas dasar apa sistem tersebut membuat 'keputusan' tersebut sehingga menimbulkan risiko bahwa sistem tersebut telah menggunakan data yang tidak sesuai untuk individu yang dimaksud atau terdapat bug dalam sistem yang menyebabkan masalah dengan resep yang berlebihan atau kurang.

	Bukti yang Salah Arah Bukti	Hasil algoritmik hanya dapat diandalkan (tetapi juga netral) tergantung pada data yang mendasarinya.	Watson for Oncology digunakan secara luas di Tiongkok untuk 'diagnosis' melalui pengenalan gambar; tetapi terutama dilatih pada kumpulan data Barat yang menyebabkan masalah dengan kesesuaian dan hasil yang lebih buruk bagi pasien Tiongkok dibandingkan dengan rekan-rekan mereka di Barat.
Kekhawatiran Normatif	Hasil yang tidak adil	Suatu tindakan dapat ditemukan memiliki dampak yang lebih besar (positif atau negatif) pada satu kelompok orang	Sebuah algoritma 'belajar' untuk memprioritaskan pasien yang diprediksi memiliki hasil yang lebih baik untuk penyakit tertentu. Hal ini ternyata berdampak diskriminatif terhadap orang-orang dalam komunitas kulit hitam dan etnis minoritas.
	Efek transformatif	Aktivitas algoritmik, seperti pembuatan profil, mengkonseptualisasikan ulang realitas dengan cara yang tidak terduga.	Seseorang yang menggunakan aplikasi kesehatan pribadi memiliki pengawasan terbatas atas data pasif apa yang dikumpulkannya dan bagaimana data tersebut diubah menjadi rekomendasi untuk perbaikan, sehingga membatasi kemampuan mereka untuk menentang rekomendasi apa pun yang dibuat dan mengakibatkan hilangnya otonomi pribadi dan privasi data.
Menyeluruh	Ketertelusuran	Kerugian yang disebabkan oleh aktivitas algoritmik sulit untuk di-debug (untuk mendeteksi kerugian dan menemukan penyebabnya), dan sulit untuk mengidentifikasi siapa yang harus bertanggung jawab atas kerugian yang ditimbulkan.	Jika keputusan yang dibuat oleh perangkat lunak pendukung keputusan klinis berujung pada hasil negatif bagi individu, tidak jelas siapa yang harus diberi tanggung jawab dan/atau kewajiban dan oleh karena itu mencegah hal itu terjadi lagi.

15.3 TEMUAN

Berikut ini adalah pembahasan mendetail mengenai isu-isu yang terungkap. Sebanyak 223 judul dipilih, duplikat dihapus, dan, saat pembacaan dimulai, referensi bibliografi yang relevan juga ditambahkan, sehingga menghasilkan sekitar 147 makalah yang dibaca dan dimasukkan dalam tinjauan. Diagram alir di bawah ini mengilustrasikan metodologi kami. Selain itu, peta ringkasan temuan kami (Tabel 15.3) disediakan di akhir bagian ini (Gambar 15.3).

Kekhawatiran Epistemik: Bukti yang Tidak Konklusif, Sulit Dipahami, dan Salah Arah

Banyak faktor yang mendorong percepatan pengembangan AI-Health. Salah satu yang paling menonjol adalah keyakinan bahwa algoritma memiliki kemampuan untuk menghasilkan keputusan klinis yang lebih objektif, konsisten, dan berbasis bukti dibandingkan praktisi kesehatan manusia (HCP) dalam hal diagnosis, prognosis, maupun rekomendasi perawatan. Keyakinan ini tidak sepenuhnya tanpa dasar. Teknik pembelajaran mesin, khususnya metode *ensemble* dan *unsupervised learning*, dapat memproses dan mempertimbangkan jangkauan data yang jauh lebih luas dibandingkan HCP dalam praktik klinis. Data tersebut mencakup lima dari tujuh dimensi utama layanan kesehatan yang diidentifikasi oleh Departemen Kesehatan dan Layanan Kemanusiaan AS, yaitu:

1. Data demografi dan sosial-ekonomi,
2. Data gejala dan diagnosis,
3. Data perawatan,
4. Data hasil klinis, serta
5. Data *omics* lainnya.

Apabila dirancang dengan memperhatikan tantangan epistemik yang ada, algoritma klinis dapat berfungsi sebagai pendamping digital bagi praktisi kesehatan. Peran ini memungkinkan pengurangan asimetri informasi antara HCP dengan pasien, melalui penyediaan data dan wawasan yang dapat diakses kedua belah pihak. Dengan demikian, keputusan medis yang paling tepat tetap dapat diambil oleh pihak yang memiliki kewenangan klinis, dengan dukungan analisis yang lebih menyeluruh.

Bukti empiris menunjukkan bahwa AI-Health mampu meningkatkan, bahkan dalam beberapa kasus melampaui kemampuan manusia pada berbagai tugas penting, seperti:

- Analisis faktor risiko; De Langavant et al., (2018); Deng et al., (2018),
- Prediksi penyakit; Moscoso et al., (2019),
- Prediksi infeksi; Barton et al., (2019); López-Martínez et al., (2019),
- Pemantauan kesehatan populasi; Lu et al., (2019); Zacher & Czogiel, (2019),
- Prediksi efek samping obat atau terapi; Ding et al., (2019); Mortazavi et al., (2017),
- Prediksi hasil pengobatan dan tingkat kelangsungan hidup pasien; Dong et al., (2019); Popkes et al., (2019); Topuz et al., (2018), serta
- Analisis rekam medis elektronik; Shickel et al., (2018).

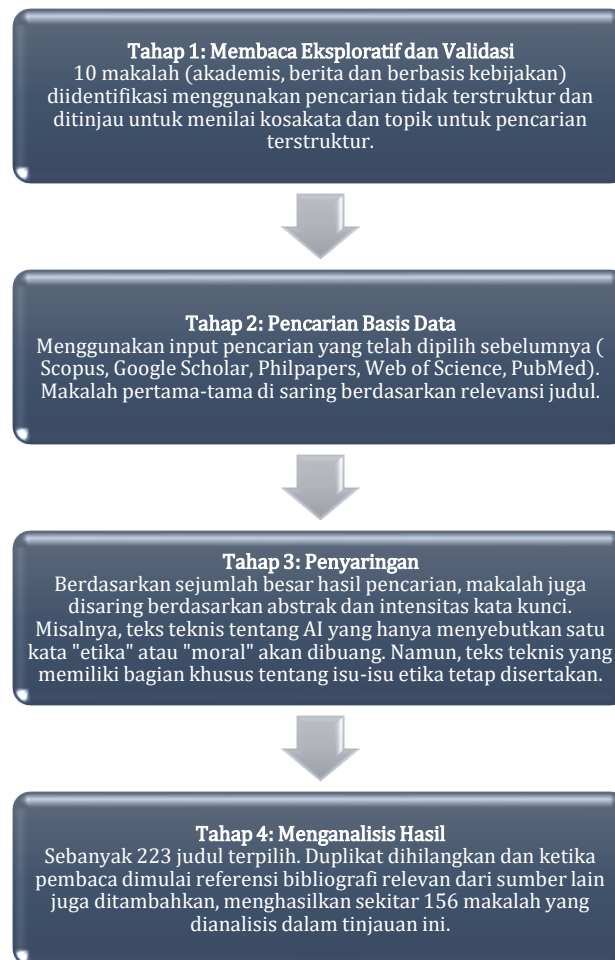
Kemampuan tersebut semakin signifikan karena AI-Health dapat beroperasi dalam skala besar, mendiagnosis atau memprediksi hasil klinis untuk ribuan bahkan jutaan pasien secara simultan sebuah kapasitas yang mustahil dicapai oleh manusia.

Namun, keyakinan yang terlalu besar terhadap “netralitas” dan “kebenaran” algoritma bukan tanpa masalah. Literatur etika AI berulang kali menyoroti bahwa klaim objektivitas algoritma sering kali hanyalah sebuah “mitos yang dikonstruksi”. Fakta bahwa sebuah algoritma mampu mengenali pola tertentu tidak otomatis membuat pola tersebut bermakna secara klinis. Dalam praktik medis, banyak penelitian berbasis AI masih menghadapi tantangan serius berupa overfitting akibat ukuran sampel yang terbatas. Akibatnya, temuan seperti identifikasi faktor risiko atau deteksi keberadaan penyakit sering kali tidak cukup kuat untuk dijadikan dasar pengambilan keputusan klinis.

Tabel 15.3 Ringkasan dari epistemik, normatif dan etika menyeluruh yang terkait dengan AI-Kesehatan di enam LoA berbeda sebagaimana diidentifikasi oleh tinjauan literatur

	A. Individu	B. Interperson al	C. Kelompok	D. Kelembaga an	E. Sektoral	F. Sosial
1. Kekhawatiran epistemik (bukti yang tidak meyakinkan, tidak dapat dipahami, dan menyesatkan)	Salah diagnosis atau diagnosis yang terlewat	Hilangnya kepercayaan dalam hubungan antara Tenaga Kesehatan dan Pasien, depersonalisasi perawatan	Kesalahan diagnosis atau diagnosis yang terlewat dalam skala besar—beberapa kelompok lebih terdampak dibandingkan kelompok lainnya	Pemborosan dana dan sumber daya yang tidak diarahkan ke bidang yang lebih membutuhkan	Pembagian data yang terlalu luas antara entitas publik dan swasta	Penyediaan layanan kesehatan publik yang lebih buruk dan memburuknya hasil kesehatan bagi masyarakat
2. Normatif (dampak transformatif dan hasil yang tidak adil)	Pengawasan & pelemahan otonomi dan integritas diri	Penurunan keterampilan tenaga kesehatan, ketergantungan berlebihan pada perangkat AI, dan melemahnya praktik	Profiling dan diskriminasi terhadap kelompok tertentu yang dianggap kurang sehat atau	Transformasi jalur perawatan & penerapan nilai-nilai spesifik dalam skala besar—mendefinisikan ulang	Siloisasi pengembangan alat AI baru di sektor swasta	Ketimpangan dalam hasil

		persetujuan serta pendefinisian ulang peran dalam sistem layanan kesehatan	berisiko tinggi	'perawatan yang baik'		
3. Menyeluruh (ketertelusran)	'Pengguna Buruk' bisa jadi disalahkan atas kesehatan mereka sendiri yang buruk	Lihat kelembagaan	Kelompok tertentu dianggap lebih tidak bertanggung jawab secara moral terhadap kesehatan mereka dibandingkan kelompok lain	Kurangnya kejelasan mengenai tanggung jawab terkait masalah keselamatan dan efektivitas dapat menghambat adopsi atau menyebabkan kelompok tertentu lebih sering disalahkan dibandingkan kelompok lain.	Lihat kelembagaan	Masyarakat harus memutuskan melalui regulasi jawaban yang lebih baik untuk pertanyaan mengenai tanggung jawab dan alokasi risiko dalam penyediaan layanan kesehatan. Namun, semua kelompok dalam masyarakat mungkin tidak diberi hak bicara yang sama dalam proses ini.



Gambar 15.4 Diagram alir yang ditawarkan dan ikhtisar langkah-langkah yang diambil dalam tinjauan pustaka kami, menyaring dari beberapa ribu judul hingga abstrak yang teridentifikasi dan memilih 156 makalah untuk dibaca

Masalah ini semakin diperparah oleh kurangnya reproduktifitas, dan validitas eksternal, dari hasil. Solusi AI-Health seringkali tidak dapat diterapkan di berbagai pengaturan dan jarang berhasil di pengaturan yang berbeda dari tempat hasil awal diperoleh, sehingga menimbulkan pertanyaan serius tentang ketelitian ilmiah AI-Health dan keamanannya. Lebih lanjut, hasilnya seringkali sarat nilai, berdasarkan definisi 'sehat' oleh orang-orang berpengaruh atau perusahaan yang kuat. Hal ini menimbulkan sejumlah kekhawatiran etika yang signifikan.

Pada LoA individual, terdapat risiko kesalahan diagnosis yang cukup besar. Hal ini dapat terjadi setidaknya dalam dua cara: baik oleh seorang individu yang menggunakan perangkat yang dapat dikenakan yang memiliki bug atau dikalibrasi secara tidak tepat untuk mereka (misalnya mereka dapat 'diberitahu' bahwa mereka menderita suatu kondisi kesehatan padahal tidak, atau sebaliknya), atau, seorang HCP yang mengandalkan perangkat lunak pendukung keputusan klinis (CDSS) dapat diberikan diagnosis atau rekomendasi yang tidak akurat yang tidak mereka pertanyakan karena kecenderungan untuk menerima keputusan sistem otomatis tanpa berpikir kritis.

Selain itu, hal ini dapat berdampak pada praktik medis, menyebabkan ketergantungan yang berlebihan pada diagnostik mesin dan penurunan keterampilan praktisi. Hal ini tidak

hanya menjadi risiko bagi individu, tetapi juga membalikkan keuntungan dari solusi AI-Health yang dapat beroperasi dalam skala besar dengan memperkenalkan kekhawatiran etis LoA kelompok tentang kesalahan diagnosis atau diagnosis yang terlewat yang terjadi berulang kali. Meskipun seorang HCP mungkin memberikan diagnosis yang salah kepada satu orang dan kemudian dikoreksi, algoritma yang salah, berdasarkan bukti yang salah arah, sulit dipahami, atau tidak meyakinkan, dapat memberikan diagnosis yang sama salahnya kepada ratusan atau ribuan orang sekaligus. Skala masalah sama besarnya dengan skala solusinya.

Berdasarkan hal ini, terdapat pula implikasi etis pada LoA interpersonal. Hubungan antara tenaga kesehatan profesional (HCP) dan pasien terutama didasarkan pada kepercayaan dan empati, dan meskipun solusi AI-Kesehatan dapat mengambil alih tugas-tugas yang lebih rutin dan terstandarisasi, solusi tersebut tidak dapat mereproduksi kebajikan emosional yang dimiliki oleh tenaga kesehatan profesional. Akibatnya, ketergantungan yang berlebihan pada bukti 'kuantitatif' dan objektif yang mendorong algoritma klinis dapat mendiskreditkan bentuk diagnosis dan perawatan lain sebuah kekhawatiran utama dalam kasus psikiatri klinis.

Hal ini dapat menyebabkan dehumanisasi atau impersonalisasi penyediaan layanan, dari layanan yang berbasis pada mendengarkan dan teori menjadi layanan yang murni berbasis kategorisasi. Hal ini tidak hanya merupakan 'paternalisme terselubung', tetapi juga dapat menyebabkan hasil kesehatan yang lebih buruk karena kesenjangan antara bukti medis murni dan perubahan perilaku yang nyata. Terakhir, dalam skala ke LoA institusional, sektoral, dan sosial, terdapat kekhawatiran bahwa keputusan kesehatan masyarakat semakin banyak dibuat berdasarkan algoritma AI-Kesehatan prediktif, yang terlalu sering bergantung pada asumsi keliru yang sama seperti yang diuraikan di atas. Mengenai asumsi ini, pertimbangkan contoh Google Flu Trends yang memantau virus influenza. Algoritma awal mendistorsi penyebaran virus di AS sehingga tampak seperti terdapat lebih banyak kasus influenza daripada yang sebenarnya dengan salah mengklasifikasikan penyakit mirip influenza sebagai kasus influenza yang terkonfirmasi.

Studi ini memiliki keterbatasan yang jelas: perilaku pencarian layanan kesehatan dari populasi, misalnya mencari informasi tentang wabah Flu, membuat penelitian ini rentan terhadap distorsi. Distorsi tersebut akan terpengaruh, misalnya, jika terdapat liputan media yang besar tentang suatu epidemi, atau oleh faktor demografis, seperti kesenjangan digital. Jika keputusan kebijakan tentang di mana akan menempatkan sumber daya kesehatan didasarkan pada bukti berkualitas buruk tersebut, hal ini dapat mengakibatkan pemborosan dana publik (misalnya, mempromosikan kampanye vaksinasi di tempat yang tidak diperlukan), merusak perekonomian lokal (misalnya, menakuti wisatawan dari suatu wilayah) yang akan menghasilkan lingkaran umpan balik positif berupa berkurangnya dana yang tersedia untuk belanja publik dan menyebabkan penyediaan layanan kesehatan publik yang lebih buruk kualitasnya, dan dengan demikian hasil kesehatan yang lebih buruk bagi masyarakat luas. Kekhawatiran ini terutama sangat penting ketika dianggap bahwa ambisi akhir AI-Health adalah untuk menciptakan sistem perawatan kesehatan pembelajaran di mana 'sistem' terus-menerus belajar dari data yang diterimanya tentang kinerja intervensinya.

Lebih lanjut, perlu dicatat bahwa, pada titik ini, contoh yang ditawarkan di atas dari Flu Trends tidak mewakili batas minat Google dan anak perusahaannya dan saudara kandungnya di bawah perusahaan induk Alphabet dalam kesehatan masyarakat. Seperti yang kita bahas di bawah ini, keterlibatan antara anak perusahaan AI Alphabet, DeepMind, dan rumah sakit besar di Inggris telah menarik perhatian regulator perlindungan data, pers, dan akademisi. Tantangan untuk memastikan bahwa sistem AI-Health berfungsi secara akurat pada gilirannya telah memicu perdebatan tentang kesesuaian berbagi data antara entitas publik dan swasta.

Menanggapi klaim bahwa data pasien yang ditransfer dari Royal Free Hospital ke DeepMind "jauh melebihi persyaratan kebutuhan yang dinyatakan secara publik", perwakilan DeepMind berpendapat bahwa "data yang diproses dalam aplikasi telah didefinisikan oleh dan saat ini sedang digunakan oleh dokter untuk pemantauan dan perawatan langsung pasien AKI (cedera ginjal akut)". Powles dan Hodson kemudian menanggapi bahwa "merupakan pernyataan fakta bahwa data yang ditransfer lebih luas daripada persyaratan AKI". Sebagaimana ditunjukkan oleh serangkaian klaim dan bantahan ini, kualitas dan kuantitas data yang diperlukan untuk aplikasi AI-Health tertentu kemungkinan akan menjadi bahan perdebatan dalam konteks pengumpulan dan pembagian data pasien dalam pelatihan AI-Health.

Pada akhirnya, data diperlukan untuk praktik medis, demikian pula solusi AI-Health yang dapat menampung data dalam volume yang lebih besar. Namun, data yang dikumpulkan dan digunakan dengan cara ini tidak cukup untuk memberikan informasi bagi praktik medis; data tersebut harus diubah agar bermanfaat. Jika proses transformasi ini cacat, hasilnya bisa sangat merugikan, mengakibatkan pemborosan dana dan penyediaan layanan kesehatan yang lebih buruk, atau pembagian data pasien yang tidak semestinya kepada pelaku sektor swasta dengan kedok AI-Health.

Kekhawatiran Normatif: Hasil yang Tidak Adil dan Efek Transformatif

Sebagaimana dirujuk dalam pendahuluan, sistem layanan kesehatan di seluruh dunia sedang berjuang dengan meningkatnya biaya dan menurunnya hasil dan para administratornya semakin yakin bahwa jawabannya mungkin terletak pada peningkatan kematangan informasi sistem layanan kesehatan dan kemampuan untuk memanfaatkan peluang yang dihadirkan oleh AI-Health secara signifikan untuk meningkatkan hasil bagi pasien, dan mengurangi beban sistem.

Meskipun mengabaikan peluang-peluang ini secara etis merupakan kelalaian, akan sama bermasalahnya secara etis jika mengabaikan fakta bahwa peluang-peluang ini tidak diciptakan oleh teknologi AI-Health semata, melainkan oleh kemampuannya untuk secara fundamental mengubah sifat intrinsik cara layanan kesehatan diberikan dengan menggabungkan, menggabungkan kembali, dan memisahkan berbagai bagian sistem. Hal ini mengubah kemampuan dan kendala yang dimiliki berbagai badan pengatur, regulator, dan agen sistem, sehingga melemahkan mekanisme yang ada untuk meminta pertanggungjawaban dari para pemberi layanan kesehatan dan dengan demikian menimbulkan risiko baru. Misalnya (disunting untuk anonimitas):

- Penggandengan: pasien dan data mereka saling terkait erat dan saling dipertukarkan sehingga pasien tersebut adalah profil genetik mereka, hasil tes darah terbaru, informasi pribadi, alergi, dll. Apa yang disebut undang-undang sebagai "subjek data" menjadi "pasien data";
- Penggandengan Ulang: penelitian dan praktik telah dipisahkan secara tajam sejak penerbitan Komisi Nasional untuk Perlindungan Subjek Manusia pada tahun 1970-an, tetapi dalam skenario digital yang dijelaskan di atas, keduanya kembali disatukan sebagai satu kesatuan;
- De-Coupling: keberadaan Penyedia Layanan Kesehatan (HCP) dan lokasi Pasien menjadi independen, misalnya karena diperkenalkannya konsultasi daring.

Transformasi ini menimbulkan sejumlah kekhawatiran etis.

Dimulai lagi dengan LoA individu: seiring semakin banyaknya intervensi diagnostik dan terapeutik yang berbasis pada solusi AI-Health, individu mungkin terdorong untuk membagikan semakin banyak data pribadi tentang diri mereka sendiri data yang kemudian dapat digunakan secara tidak transparan. Hal ini berarti bahwa kemampuan individu untuk terlibat secara bermakna dalam pengambilan keputusan bersama menjadi sangat terkikis.

Akibatnya, meningkatnya penggunaan pengambilan keputusan algoritmik dalam pengaturan klinis dapat berdampak negatif terhadap otonomi individu. Agar seseorang dapat menggunakan agensi atas keputusan klinis yang berasal dari AI-Health, mereka perlu memiliki pemahaman yang baik tentang data, proses, dan kemungkinan teknis yang mendasarinya yang terlibat dalam pengambilan keputusan tersebut dan mampu memastikan nilai-nilai mereka sendiri dipertimbangkan.

Sebagian besar populasi tidak memiliki tingkat literasi eHealth yang diperlukan untuk ini, dan mereka yang memilikinya (termasuk HCP) dicegah untuk mendapatkan pemahaman ini karena sifat kotak hitam algoritma AI-Health. Dalam kasus ekstrem, ini dapat merusak kepercayaan diri seseorang dalam kemampuan mereka untuk menolak pengobatan. Masalah-masalah tersebut menimbulkan ancaman substansial terhadap integritas diri seseorang (kemampuan seseorang untuk memahami kekuatan yang bekerja pada mereka). Mengingat bahwa kerusakan pada integritas psikologis seseorang dapat dianggap sebagai 'bahaya', tidak memperhitungkan potensi ini menimbulkan risiko menciptakan sistem yang melanggar prinsip pertama etika medis: *primum non nocere* ("*pertama, jangan membahayakan*")

Tidak selalu terjadi bahwa dampak berbahaya akan terutama dirasakan oleh pasien. Pada LoA interpersonal, HCP sendiri mungkin merasa semakin tersisih karena keputusan dibuat oleh pasien dan algoritma 'saran klinis' mereka dalam lingkaran digital tertutup. Akibatnya, HCP mungkin juga merasa tidak mampu menggunakan wewenang mereka sendiri atas kapasitas pengambilan keputusan solusi AI-Health. Meskipun penggunaan pengambilan keputusan algoritmik membuat diagnostik tampak seperti aktivitas langsung untuk mengidentifikasi gejala dan memasukkannya ke dalam kategori buku teks, praktik medis jauh lebih tidak jelas daripada yang terlihat.

Praktik klinis melibatkan serangkaian evaluasi, coba-coba, dan interaksi dinamis dengan pasien dan literatur medis. Akibatnya, protokol perawatan formal harus dilihat lebih sebagai pedoman evaluatif daripada kategori yang terdefinisi dengan baik dan terisolasi. Solusi AI-Health mungkin tidak sesuai dengan praktik terbaik saat ini, yang diperlukan untuk menangani tingkat ketidakpastian yang tinggi dan hanya dapat dievaluasi sepenuhnya oleh dokter. Oleh karena itu, solusi AI-Kesehatan perlu memungkinkan Petugas Kesehatan untuk memberikan pengaruh dalam proses pengambilan keputusan.

Pada LoA kelompok, kekhawatirannya adalah bahwa sistem AI-Kesehatan mungkin dapat mengidentifikasi penyakit dan cedera dengan lebih baik yang memiliki protokol perawatan yang mapan dan cukup baik (dan karenanya dapat diotomatisasi). Protokol ini lebih mungkin ada untuk penyakit yang paling sering diderita oleh pria kulit putih karena terdapat volume data uji medis yang lebih besar untuk kelompok ini dibandingkan dengan hampir semua kelompok lainnya.

Algoritma yang dilatih pada kumpulan data yang bias tersebut dapat menghasilkan prediksi yang jauh lebih buruk, misalnya, untuk perempuan kulit hitam yang lebih muda. Jika HCP tidak dilibatkan sama sekali dan sistem pembelajaran layanan kesehatan terutama bergantung pada keputusan otomatis, terdapat potensi yang besar untuk memperparah kesenjangan yang sudah ada antara kelompok “kaya” dan “miskin” dalam ekosistem layanan kesehatan digital, yaitu mereka yang menghasilkan cukup data tentang diri mereka sendiri untuk memastikan algoritma yang terlatih secara akurat dan mereka yang tidak.

Untuk memitigasi risiko-risiko ini dan risiko terkait, institusi perlu mengajukan pertanyaan krusial: seberapa besar pengambilan keputusan klinis yang harus didelegasikan kepada solusi AI-Health? Jika diketahui bahwa algoritma yang memungkinkan pembuatan profil (misalnya, algoritma yang menentukan profil risiko genetik) dapat mengabaikan outlier dan menjadi dasar diskriminasi, memutuskan apakah layanan kesehatan juga dipandang sebagai sarana untuk mempromosikan keadilan sosial sangatlah penting untuk menetapkan: jenis layanan data apa yang akan diintegrasikan ke dalam sistem; data apa yang harus dikumpulkan; dan nilai-nilai apa yang harus diintegrasikan ke dalam layanan pengambilan keputusan algoritmik.

Keputusan ini juga menentukan jenis perubahan perilaku di tingkat populasi yang harus dapat dicapai oleh sistem kesehatan, bergantung pada manajemen biaya, pengumpulan data, dan keadilan dalam sistem berbasis data (Departemen Kesehatan dan Perawatan Sosial 2018). Jika tidak dipertimbangkan dengan cermat, proses transformasi penyediaan perawatan ini berisiko membuat sistem terlalu pas dengan serangkaian nilai tertentu yang mungkin tidak mewakili nilai-nilai masyarakat luas.

Efek transformatif lain yang lebih halus namun meluas muncul di tingkat sektoral. Powles dan Hodson (2017) berpendapat bahwa salah satu risiko yang mungkin timbul dari kolaborasi antara entitas sektor publik dan swasta seperti antara rumah sakit Royal Free London dan DeepMind adalah bahwa manfaat positif dari “solusi” AI-Health akan terisolasi dalam entitas swasta. Mereka mencatat bahwa dalam kasus Royal Free, “DeepMind [diberi]

keunggulan dalam mengembangkan alat algoritmik baru pada kumpulan data yang sebelumnya dimiliki secara pribadi, tetapi dihasilkan secara publik".

Hal ini, menurut mereka, mungkin berarti bahwa satu-satunya cara yang layak untuk mengembangkan kemajuan di masa depan adalah "melalui DeepMind dengan ketentuan DeepMind". Penafsiran ini dibantah oleh DeepMind, yang menyebutnya "tidak berdasar dan tidak benar" dan mengklaim bahwa Komisioner Informasi setuju dengan pendirian mereka dalam putusannya tahun 2018. Apa pun keadaan kasus khusus ini, risiko yang lebih luas dari solusi AI-Kesehatan yang dimiliki secara pribadi dilatih pada kumpulan data yang telah dihasilkan tentang publik oleh aktor publik tetapi kemudian (secara sah) dibagikan dengan perusahaan swasta layak diwaspadai untuk ke depannya dan menjadi topik diskusi yang berharga dalam etika kesehatan masyarakat.

Seperti yang mungkin sekarang jelas, efek transformatif ini juga memiliki implikasi etis yang signifikan pada LoA masyarakat. Sebelum lembaga dapat menetapkan di mana dan bagaimana (dan, dari perspektif sektoral, apakah) solusi AI-Kesehatan dapat meningkatkan perawatan, masyarakat sendiri harus membuat keputusan sulit tentang apa itu perawatan dan apa yang merupakan perawatan yang baik. Sebagai contoh sederhana, apakah ini berarti hanya menyediakan diagnosis teknis dan resep yang tepat, atau melibatkan perenungan serangkaian kebutuhan manusia yang berpusat pada kesejahteraan? Jika yang pertama, maka relatif mudah untuk mengotomatiskan peran dokter non-bedah melalui AI (meskipun ini tidak berarti dokter harus digantikan oleh sistem AI).

Namun, jika yang kedua, maka menyediakan layanan kesehatan yang baik berarti mencakup kesejahteraan psikologis dan elemen lain yang berkaitan dengan kualitas hidup, yang akan menjadikan interaksi manusia sebagai bagian penting dari penyediaan layanan kesehatan, karena mesin tidak memiliki kemampuan untuk membuat keputusan yang didorong oleh emosi. Akibatnya, keputusan tertentu mungkin sepenuhnya melampaui kemampuan mesin, sehingga mendelegasikan tugas-tugas ini kepada AI-Health akan menimbulkan masalah etika.

Pertimbangkan, misalnya, situasi di mana solusi AI-Health memutuskan pasien mana yang akan dikirim ke Unit Perawatan Intensif (ICU). Perawatan intensif merupakan sumber daya yang terbatas dan hanya orang-orang yang berisiko kehilangan nyawa atau mengalami cedera serius yang dikirim ke sana. Keputusan triase saat ini dibuat oleh manusia dengan tujuan memaksimalkan kesejahteraan sebanyak mungkin orang. Dokter mempertimbangkan berbagai faktor saat membuat keputusan ini, termasuk kemungkinan orang tersebut bertahan hidup jika mereka dikirim ke ICU.

Situasi ini sering kali melibatkan praktisi (secara implisit) yang mengambil sikap moral, dengan memprioritaskan individu berdasarkan usia atau kondisi kesehatan mereka. Kasus-kasus ini pada dasarnya berorientasi pada batasan hukum dan norma medis (misalnya kepatuhan terhadap prinsip-prinsip bioetika atau kode praktik terbaik), namun keahlian, pengalaman, dan nilai-nilai pribadi juga tak terelakkan memainkan peran. Dukungan AI-Health dalam skrining ICU meningkatkan jumlah agen dan mempersulit norma-norma yang terlibat

dalam keputusan ini, karena dokter dapat mengikuti pedoman profesionalnya, sementara algoritma akan berorientasi pada nilai-nilai yang tertanam dalam kodenya.

Kecuali jika ada proses transparan yang melibatkan masyarakat dalam mempertimbangkan nilai-nilai yang tertanam dalam alat pengambilan keputusan ini (misalnya, bagaimana penyediaan perawatan yang 'adil' didefinisikan?), maka penggunaan algoritma dalam skenario semacam itu dapat mengakibatkan sistem kesehatan yang terlalu pas dengan serangkaian nilai tertentu yang tidak mewakili masyarakat luas. Menanggapi risiko ini, beberapa upaya telah dilakukan untuk melibatkan masyarakat luas dalam pengambilan keputusan mengenai desain dan penerapan sistem AI. Pada awal 2019, Kantor Komisioner Informasi Inggris dan Institut Nasional untuk Penelitian Kesehatan menggelar serangkaian "juri warga" untuk mendapatkan pendapat dari perwakilan lintas bagian masyarakat Inggris mengenai penggunaan AI dalam kesehatan. "Juri" disajikan dengan empat skenario, dua berkaitan dengan kesehatan menggunakan AI untuk mendiagnosis stroke, dan menggunakannya untuk menemukan kecocokan potensial untuk transplantasi ginjal dan dua lainnya berkaitan dengan peradilan pidana.

Khususnya, juri "sangat menyukai akurasi daripada penjelasan" dalam dua skenario yang melibatkan AI dalam kesehatan (Institut Nasional untuk Penelitian Kesehatan 2019). Ini hanyalah salah satu contoh penelitian yang mencoba untuk mendapatkan data opini publik mengenai AI dalam kesehatan, dan ada alasan untuk menduga bahwa preferensi yang tampak di antara peserta untuk sistem AI yang akurat daripada yang dapat dijelaskan mencerminkan skenario berisiko tinggi dan bergerak cepat yang disajikan (sebagai lawan dari, katakanlah, penyakit dan cedera yang lebih rutin yang menjadi fokus kita di sini). Meskipun demikian, hal ini menunjukkan masuk akal dan lebih disukainya melibatkan publik dalam merancang sistem AI-Kesehatan.

Sebagai penutup sub-bagian ini, gagasan bahwa teknologi AI-Kesehatan bersifat netral secara etis tidaklah realistis, dan membiarkan teknologi tersebut melakukan pengambilan keputusan dan penegakan moral dapat memicu hasil yang tidak bermoral dan tidak adil. Keterlibatan langsung publik dalam perancangan AI-Kesehatan dapat membantu memitigasi risiko-risiko ini. Hal ini perlu diingat oleh semua pihak yang terlibat dalam transformasi sistem layanan kesehatan berbasis AI.

Kekhawatiran Utama: Ketertelusuran

Sub-bagian sebelumnya menguraikan bagaimana meningkatnya penggunaan AI-Kesehatan secara fundamental mentransformasi pemberian layanan kesehatan dan implikasi etis dari proses ini, terutama dalam hal potensi hasil yang tidak adil. Proses transformasi ini berarti bahwa sistem layanan kesehatan kini bergantung pada serangkaian interaksi yang dinamis, siklis, dan saling terkait antara agen manusia, buatan, dan hibrida. Hal ini membuat identifikasi risiko yang muncul akibat interaksi dan alokasi tanggung jawab menjadi semakin sulit, sehingga menimbulkan kekhawatiran etis terkait tanggung jawab moral.

Tanggung jawab moral melibatkan pandangan ke depan, di mana individu, kelompok, atau organisasi dianggap bertanggung jawab untuk menjamin hasil yang diinginkan, dan pandangan ke belakang untuk mencari kesalahan dan kemungkinan perbaikan yang tepat,

ketika kegagalan terjadi. Dalam sistem layanan kesehatan yang berfungsi dengan baik, tanggung jawab ini didistribusikan secara merata dan transparan di semua simpul sehingga rantai kausal dari suatu hasil dapat dengan mudah direplikasi jika hasilnya positif, atau dicegah agar tidak terulang jika hasilnya negatif.

Dalam sistem layanan kesehatan yang digerakkan oleh algoritma, satu alat diagnostik AI mungkin melibatkan banyak orang yang mengorganisir, mengumpulkan, dan menjadi perantara data, serta melakukan analisis terhadapnya, sehingga alokasi tanggung jawab yang transparan ini hampir mustahil. Intinya, proses pengambilan keputusan dari satu algoritma saja tidak hanya merupakan kotak hitam, tetapi seluruh rantai aktor yang terlibat dalam produk akhir solusi AI-Kesehatan juga sangat kompleks. Hal ini membuat seluruh ekosistem AI-Kesehatan tidak dapat diakses dan tidak transparan, sehingga menyulitkan tanggung jawab dan akuntabilitas.

Untuk menguraikan implikasi etis dari kurangnya ketertelusuran dalam skala besar ini, mari kita ambil contoh monitor detak jantung digital yang 'secara cerdas' memproses data biologis dan lingkungan untuk memberi sinyal kepada penggunanya tentang risiko mereka terkena penyakit jantung. Pada LoA individu, proses ini bergantung pada apa yang disebut sebagai 'pandangan medis digital' (Disunting untuk anonimitas) dan didasarkan pada siklus mikro refleksi diri yang diadaptasi dari:

1. **Memperoleh Pengetahuan:** Algoritma membaca kumpulan data multi-omik untuk menentukan risiko serangan jantung dan memberikan individu 'skor kesehatan jantung'.
2. **Memperoleh Kesadaran:** atas saran algoritma, individu mulai memantau tingkat aktivitas mereka dan menyadari seberapa aktif mereka.
3. **Refleksi diri:** sesuai arahan algoritma, individu merefleksikan berapa banyak makanan berlemak tinggi yang mereka konsumsi dalam sehari dan membandingkannya dengan diet optimal mereka berdasarkan profil genomik dan tingkat aktivitas mereka.
4. **Tindakan:** individu mengikuti saran algoritma dan mengambil tindakan spesifik untuk meningkatkan skor kesehatan jantung mereka, misalnya, memulai olahraga teratur.

Jika proses refleksi diri ini tidak 'berhasil' dalam arti tidak menghasilkan tindakan yang tepat untuk meningkatkan kesehatan jantungnya, karena berbagai alasan, termasuk ketidakakuratan data, dan individu tersebut tetap mengalami gagal jantung, proses pengawasan algoritmik ini berisiko menciptakan mekanisme rumit untuk menyalahkan korban.

Individu tersebut dapat dianggap sebagai 'pengguna yang buruk' karena gagal menindaklanjuti saran algoritma yang konon objektif dan berbasis bukti, dan oleh karena itu dapat dianggap bertanggung jawab secara moral atas kesehatannya yang buruk dan tidak layak mendapatkan layanan kesehatan yang disediakan negara. Namun, karena kurangnya ketertelusuran, tidak ada kepastian bahwa hasil yang buruk tersebut disebabkan oleh kurangnya tindakan dari individu: bisa jadi perangkat yang rusak, kode yang bermasalah, atau hasil dari kumpulan data yang bias.

Lebih lanjut, bahkan jika hasil negatif semata-mata disebabkan oleh individu yang mengabaikan panduan, adopsi infrastruktur digital yang memungkinkan kegagalan dikaitkan

dengan individu yang secara moral 'bersalah' itu sendiri merupakan masalah etika. Wawasan baru ini dapat memungkinkan nyawa diselamatkan dan kualitas hidup ditingkatkan secara drastis, namun juga mengalihkan beban etika 'hidup dengan baik' langsung ke individu yang baru bertanggung jawab. Pergeseran ontologis yang diizinkan oleh infrastruktur baru ini dari individu-sebagai-pasien yang berhak mendapatkan layanan kesehatan berkualitas, terlepas dari pilihan mereka sebelumnya sebagai manusia yang dapat berbuat salah, menjadi individu-sebagai-agen yang diharapkan mengambil langkah aktif untuk mencegah hasil negatif menimbulkan pertanyaan tajam bagi bioetika, yang secara tradisional dipandang sebagai "etika penerima".

Lebih lanjut, perubahan teknologi ini mungkin mendorong pergeseran dalam kerangka etika, membebani individu, sementara tidak menyediakan sarana perubahan perilaku secara *de facto*. Banyak kekhawatiran bermula dari isu sosio-demografis yang melibatkan kebiasaan buruk, dan tidak dapat disederhanakan menjadi masalah penyampaian informasi yang memadai kepada pasien. Akibat isu bias, terdapat risiko etika LoA kelompok yang memungkinkan beberapa kelompok dianggap lebih tidak bertanggung jawab secara moral terkait perawatan kesehatan mereka dibandingkan kelompok lain.

Monitor detak jantung, misalnya, terkenal kurang akurat bagi mereka yang berkulit gelap, yang berarti monitor tersebut dapat memberikan saran yang jauh kurang akurat kepada orang kulit berwarna dibandingkan mereka yang berkulit terang. Jika hal ini mengakibatkan orang kulit berwarna kurang mampu menggunakan saran AI-Health untuk meningkatkan kesehatan jantung mereka, maka kelompok orang ini dapat dianggap tercela secara moral dalam hal kesehatan mereka.

Lebih lanjut, layanan kesehatan kemudian dapat 'belajar' memprediksi bahwa orang kulit berwarna memiliki kesehatan jantung yang lebih buruk, yang berpotensi mengakibatkan kelompok individu ini didiskriminasi oleh, misalnya, perusahaan asuransi. Pada LoA interpersonal, institusional, dan sektoral, tanggung jawab moral ini diterjemahkan menjadi liabilitas. Jika misalnya, alih-alih algoritma kesehatan jantung memberikan saran kembali kepada individu, algoritma tersebut memberikan data kepada HCP individu tersebut dan HCP memberikan saran yang gagal mencegah kejadian buruk atau secara langsung menyebabkan kejadian buruk, hal ini dapat menjadi dasar gugatan malapraktik medis.

Dalam skenario ini, masih belum jelas di mana liabilitas pada akhirnya akan berada. Hukum yang berlaku saat ini menyiratkan bahwa Tenaga Kesehatan Profesional (HCP) akan bersalah, dan oleh karena itu bertanggung jawab, atas suatu kejadian yang tidak diinginkan karena algoritma dalam skenario ini akan dianggap sebagai alat bantu diagnostik seperti tes darah tanpa kapasitas pengambilan keputusan. Oleh karena itu, HCP bertanggung jawab untuk bertindak sesuai informasi yang diberikan. Namun, rantai pasok untuk setiap algoritma klinis jauh lebih kompleks dan kurang transparan dibandingkan alat diagnostik yang lebih tradisional. Hal ini berarti banyak yang mempertanyakan apakah hukum akan ditafsirkan seperti ini di masa mendatang.

Misalnya, apakah tanggung jawab benar-benar terletak pada HCP karena tidak mempertanyakan hasil algoritma, bahkan jika mereka tidak dapat mengevaluasi kualitas

diagnostik terhadap sumber informasi lain, termasuk pengetahuan pribadi mereka sendiri tentang pasien karena sifat kotak hitam dari algoritma itu sendiri? Dan bagaimana dengan peran rumah sakit atau fasilitas perawatan: apakah mereka memiliki tanggung jawab untuk menerapkan kebijakan yang memungkinkan HCP untuk menolak saran algoritmik ketika hal ini tampaknya diindikasikan?

Demikian pula, apa peran komisioner atau pengecer perangkat yang berisi algoritma tersebut? Apakah mereka tidak memikul sebagian tanggung jawab karena tidak memeriksa keakuratannya, atau apakah mereka berasumsi bahwa tanggung jawab ini berada pada regulator (misalnya, MHRA di Inggris, FDA di AS atau CFDA di Tiongkok) yang seharusnya, menanggung beban karena tidak menilai produk dengan tepat sebelum digunakan di pasar? Bagaimana jika masalahnya lebih jauh kembali dalam rantai tersebut, yang berasal dari pengkodean yang tidak akurat atau data pelatihan berkualitas buruk? Terdapat kurangnya tanggung jawab yang terdistribusi secara jelas suatu masalah yang diperburuk oleh kurangnya transparansi yang membuat sulit untuk meminta pertanggungjawaban masing-masing bagian dari rantai tersebut atas hasil yang buruk, yang menimbulkan risiko etika yang signifikan.

Dalam tinjauan mereka tentang isu-isu keselamatan pasien dengan AI dalam perawatan kesehatan, He et al. (2019) menyatakan bahwa mereka yang bekerja di lapangan sedang mencoba untuk membangun pendekatan seluruh sistem yang tidak menyalahkan individu atau perusahaan individu, tetapi menyimpulkan bahwa di mana tanggung jawab pada akhirnya akan berada masih harus dilihat. Ini bermasalah karena, seperti yang ditekankan Hoffman et al. (2019), penerimaan alat pengambilan keputusan algoritmik oleh komunitas klinis sangat tidak mungkin sampai pertanyaan tanggung jawab ini diselesaikan, yang dapat mengakibatkan masalah etika menyeluruh yang diangkat dalam pendahuluan yaitu peluang yang hilang secara signifikan.

Banyak, termasuk, percaya bahwa kemampuan menjelaskan adalah jawaban untuk memecahkan masalah ini dan bahwa, jika HCP dapat memahami bagaimana suatu keputusan dicapai, maka merefleksikan output dari suatu algoritma tidak berbeda dari alat diagnostik lainnya. Memang Schönberger (2019) berpendapat bahwa secara hukum hal ini memang demikian dan selama dapat dibuktikan bahwa kewajiban perawatan telah dipenuhi, maka kerugian yang disebabkan kepada pasien oleh prediksi yang salah dari sistem AI-Health belum dapat dianggap sebagai kelalaian medis, tetapi mungkin dalam waktu dekat dapat dianggap sebagai kelalaian jika tidak mengandalkan keluaran algoritmik, yang membawa kita kembali ke masalah yang diuraikan dalam Pasal 18.3.1.

Secara keseluruhan, ketidakjelasan ini akan terus berlanjut untuk beberapa waktu, sehingga kembali menjadi isu sosial. Masyarakat pada akhirnya akan menentukan jawaban yang dapat diterima dan disukai secara sosial atas pertanyaan-pertanyaan mendesak ini. Isu etisnya adalah apakah semua lapisan masyarakat akan memiliki hak suara yang setara dalam debat ini, seperti dalam contoh juri warga negara di atas, atau apakah individu atau kelompok dengan suara paling lantanglah yang akan menetapkan aturan. Sebagaimana dibuktikan, ketika memikirkan kekuatan sebuah algoritma, kita perlu berpikir melampaui dampak dan

konsekuensi kodenya, menuju pada cara-cara ampuh di mana gagasan dan ide tentang algoritma beredar di dunia sosial.

15.4 PERLUNYA ETIKA DAN PROPORSIONALITAS DALAM PENDEKATAN

Oleh karena itu, kita perlu mempertimbangkan isu-isu etika epistemik, normatif, dan ketertelusuran pada enam LoA yang berbeda untuk menetapkan bidang diskusi yang berbeda. Melindungi masyarakat dari bahaya AI-Kesehatan lebih dari sekadar melindungi pengumpulan data dan memastikan bahwa model algoritmik telah divalidasi. Diskusi perlu memahami bagaimana isu-isu ini muncul secara berbeda pada berbagai tahap siklus pengembangan algoritmik; isu-isu etika yang muncul pada tahap pengumpulan data kemungkinan akan berbeda dengan yang muncul pada tahap penerapan. Contoh masalah yang sering muncul adalah tantangan hukum terkait alokasi tanggung jawab dalam kasus kesalahan medis.

Diskusi legislatif dan regulasi ini secara langsung bergantung pada pemahaman isu-isu etika di setiap tahap implementasi AI (misalnya pengumpulan data, pelatihan, penerapan) dan pengambilan keputusan normatif tentang bagaimana risiko dan beban ini akan didistribusikan ke seluruh masyarakat. Oleh karena itu, diskusi etika perlu direncanakan sebelum terlibat dalam diskusi kebijakan, regulasi, atau legislatif apa pun.

Demikian pula, banyak tantangan tidak akan diatasi melalui aturan. Sebagian besar risiko dalam penanganan data dan algoritma berasal dari para profesional yang tidak mengadopsi langkah-langkah untuk melindungi privasi dan mendukung keamanan siber. Dalam kasus ini, para pembuat kebijakan dapat menggunakan kerangka kerja kami untuk mengidentifikasi di tingkat mana mereka dapat mengatasi masalah dengan sebaik-baiknya. Misalnya, salah satu isunya adalah peningkatan kapasitas individu, yang solusinya adalah meningkatkan pemahaman dan kendali dokter dan pasien atas perangkat AI; edukasi tentang bagaimana AI-Health menghasilkan prediksi atau rekomendasi yang digunakan dalam rencana perawatan, serta akses dan perlindungan data pasien.

Namun, masalah ini dapat terjadi pada skala organisasi (kelompok), sehingga kontrol yang lebih baik atas bagaimana antarmuka dan desain produk AI-Kesehatan memengaruhi interaksi HCP-pasien-agen buatan dapat mengatasi masalah tersebut. Terakhir, beberapa kasus dapat ditangani di tingkat institusional, dengan pengorganisasian kampanye dan sertifikasi bagi para profesional yang ingin menggunakan perangkat AI-Kesehatan juga diperlukan untuk implementasi dan penggunaan AI yang memadai.

Untuk mengatasi tantangan ini, regulator harus mempertimbangkan mekanisme keras dan lunak, yang berarti apa yang seharusnya dilakukan dan apa yang boleh dilakukan berdasarkan kewajiban moral yang ada. Mekanisme ini harus mempertimbangkan berbagai pemangku kepentingan yang terlibat dalam setiap isu dan LoA, untuk menyeimbangkan kebutuhan melindungi individu dari bahaya, sekaligus tetap mendukung inovasi yang dapat memberikan manfaat nyata bagi sistem dan pasien (Disunting untuk anonimitas).

Singkatnya, sistem layanan kesehatan tidak boleh terlalu berhati-hati dalam mengadopsi solusi AI-Health, tetapi harus memperhatikan potensi dampak etisnya sehingga model tata kelola yang proporsional dapat dikembangkan. Model tata kelola ini, pada

gilirannya, dapat membantu memastikan bahwa mereka yang bertanggung jawab atas sistem layanan kesehatan bertanggung jawab atas pemberian layanan kesehatan yang berkualitas tinggi, adil, dan aman. Apa yang seharusnya dicakup oleh peraturan, standar, dan kebijakan ini dan bagaimana cara pengembangannya masih menjadi pertanyaan terbuka, yang kemungkinan akan 'diselesaikan' berkali-kali oleh berbagai sistem layanan kesehatan yang beroperasi di lingkungan yang berbeda.

Namun, untuk memberikan pendekatan yang lebih sistematis dalam menjawab pertanyaan-pertanyaan yang belum terjawab ini, yang memungkinkan koherensi dan kecepatan yang lebih besar dalam mengatasi tantangan-tantangan ini, pada Tabel 15.4 di bawah ini kami telah menyusun daftar pertimbangan lintas sektoral penting yang muncul dari tinjauan pemetaan kami. Tabel ini menunjukkan aspek tinjauan pemetaan kami (perhatian etis $\times$ LoA, sesuai dengan sel pada Tabel 15.3) yang menunjukkan setiap pertimbangan yang ditetapkan berdasarkan Tingkat Abstraksi yang semakin meningkat: Individu (A), Interpersonal (B), Kelompok (C), Kelembagaan (D), Sektoral (E), dan Masyarakat (F).

Ada langkah-langkah yang diambil untuk regulasi dan legislasi, namun diskusi-diskusi ini seringkali gagal menjawab pertanyaan-pertanyaan etika yang lebih luas seperti "apa yang dimaksud dengan layanan kesehatan yang baik?", "layanan apa yang seharusnya dipertimbangkan dalam standar 'perawatan' kita?", dan lain-lain. Tanpa menjawab pertanyaan-pertanyaan yang lebih besar ini, sulit untuk mengarahkan kerangka kerja normatif yang lebih luas dan menghasilkan koherensi di antara para pemangku kepentingan di setiap LoA.

Karena alasan-alasan ini, perkembangannya berjalan lambat (itulah sebabnya literatur yang relevan kemungkinan besar tidak mencerminkan semua perkembangan terkini) dan hampir semuanya hanya berfokus pada intervensi yang memposisikan diri sebagai intervensi yang berkaitan dengan kesehatan dalam arti medis, bukan dalam arti kesejahteraan yang lebih luas (misalnya, olahraga sehat, pola makan, kebiasaan tidur). Kesadaran akan perlunya mempertimbangkan pertanyaan-pertanyaan ini semakin meningkat, dan berbagai upaya sedang dilakukan di tingkat nasional dan internasional untuk mengadaptasi regulasi yang ada agar tetap sesuai dengan tujuannya (*The Lancet Digital Health* 2019). Badan Pengawas Obat dan Makanan Amerika (FDA) kini berencana untuk meregulasi Perangkat Lunak sebagai Alat Kesehatan (SaMD) (*Food and Drug Administration (FDA)* 2019).

Di Uni Eropa dan Inggris, Peraturan 2017/745 tentang alat kesehatan akan mulai berlaku pada April 2020 dan secara signifikan memperluas cakupan perangkat lunak dan produk non-medis yang perlu diklasifikasikan (dan dinilai) sebagai alat kesehatan. Perdebatan praktis dan normatif ini perlu dikaji lebih lanjut mengenai apa yang diharapkan dari sebuah alat kesehatan, dan oleh karena itu, apa yang dianggap sebagai pengobatan.

Tabel 15.4 Sebelas pertimbangan utama bagi para pembuat kebijakan yang muncul dari tinjauan pustaka, ditandai dengan peningkatan Tingkat Abstraksi: Individu (A), Interpersonal (B), Kelompok (C), Kelembagaan (D), Sektoral (E) dan Masyarakat (F)

Pertimbangan	Literatur Pendukung Utama	Aspek Relevan (urut naik LoA*)	Contoh Lembaga yang Bertanggung Jawab (berdasarkan Sistem Kesehatan Nasional Inggris)
Keterampilan apa yang diperlukan tenaga kesehatan profesional agar dapat menggunakan solusi AI-Health secara aman dan efektif di masa depan?	Kluge et al. (2018)	Epistemik (A, B, C, F) Normatif (B, C, D, E) Menyeluruh (A, C)	Health Education England sebaiknya melakukan survei keterampilan yang tersedia saat ini di tenaga kerja dan melakukan analisis kesenjangan keterampilan yang akan dibutuhkan.
Tugas mana yang sebaiknya didelegasikan kepada solusi AI-Health, dan mana yang tidak?	Di Nucci (2019)	Epistemik (A, B, C, D, F) Normatif (B, C, D, F) Menyeluruh (A, C, D)	Department of Health and Social Care harus melakukan proses keterlibatan multi-pemangku kepentingan untuk memahami tugas mana yang secara sosial dapat diterima untuk didelegasikan kepada AI-Health dan menjadikannya kebijakan resmi.
Bukti apa yang diperlukan untuk 'membuktikan' efektivitas klinis dari solusi AI-Health?	Greaves et al. (2018)	Epistemik (A, B, C, E, F) Normatif (E) Menyeluruh (A, C, D, F)	Medicines and Healthcare products Regulatory Agency (MHRA) harus memperbarui regulasi perangkat medis dengan mencakup standar minimum akurasi dan bukti minimum untuk menunjukkan bahwa produk AI-Health benar-benar mampu berfungsi pada tingkat ini.
Mekanisme apa yang harus disiapkan agar masyarakat dapat melaporkan dan meminta ganti rugi atas kerugian terkait AI-Health?	Schönberger (2019)	Epistemik (A, C, E, F) Normatif (A, C, E, F) Menyeluruh (A, C, D)	Care Quality Commission harus memperbarui kerangka inspeksinya untuk secara rutin memeriksa bahwa produk AI-Health yang digunakan tetap beroperasi dengan aman. MHRA juga harus ikut serta dalam pengawasan.
Mekanisme apa yang harus disiapkan agar semua pandangan pemangku kepentingan relevan dapat diikutsertakan dalam pengembangan solusi AI-Health?	Aitken et al. (2019)	Epistemik (C, E, F) Menyeluruh (A, C, D, F)	Health Research Authority harus memperbarui panduan tentang persetujuan etis bagi penelitian AI-Health dan pengembangan produk, dengan menetapkan persyaratan minimum partisipasi dari beragam pemangku kepentingan.
Bagaimana keterjelasan (explainability) solusi AI-Health dapat dijamin?	Watson et al. (2019)	Epistemik (A, C) Normatif (A, C, E) Menyeluruh (A, D)	Medicines and Healthcare products Regulatory Agency (MHRA) harus memperbarui regulasi perangkat medis untuk menetapkan standar minimum keterjelasan (<i>explainability</i>) bagi produk AI-Health.

Mekanisme apa yang dapat disiapkan untuk menjamin keandalan, replikasi, dan keselamatan solusi AI-Health?	Challen et al. (2019)	Epistemik (A, C, F) Normatif (C, E, F) Menyeluruh (A, C, D)	National Institute for Health and Care Excellence (NICE) harus menjadikan ini sebagai persyaratan dalam penilaian teknologi kesehatan formal. Dalam batas kelayakan teknis dan perlindungan hak kekayaan intelektual, pengembang diwajibkan membuka kode agar dapat direplikasi dan diperiksa kesalahannya.
Bagaimana transparansi dapat dijaga terkait cara integrasi alat algoritmik dalam sistem kesehatan?	Vayena et al. (2015)	Epistemik (A, B, C, D, E, F)	NHS England harus menjadikannya persyaratan bagi semua <i>NHS trusts</i> , rumah sakit, dan institusi terkait untuk memastikan transparansi dalam penerapan alat algoritmik.
Bagaimana alur kerja layanan kesehatan, cara ia membentuk keputusan, serta dampaknya terhadap optimasi proses dalam layanan medis dapat dijamin?	–	Normatif (A, B, D, F) Menyeluruh (A, D, F)	Penyedia layanan kesehatan harus menyatakan kapan solusi AI-Health digunakan dalam jalur perawatan tertentu serta menjelaskan bagaimana aspek keselamatan dan kualitasnya dievaluasi secara rutin.
Bagaimana sumber data kesehatan tradisional dan non-tradisional dapat dimasukkan dalam pengambilan keputusan AI-Health? Dan bagaimana data tersebut dapat dilindungi secara tepat serta diharmonisasikan?	Maher et al. (2019), Ploug and Holm (2016), Richardson Milam et al. (2015), Townend (2018)	Epistemik (A, C, D, E, F) Normatif (A, C, D, E, F) Menyeluruh (A, C, D, E)	Health Research Authority dan NHS Digital harus memperbarui panduan serta regulasi terkait penggunaan sekunder data kesehatan agar mempertimbangkan aspek khusus AI-Health.
Bagaimana konsep bioetika (beneficence, non-maleficence, autonomy, dan justice) menurut Beauchamp & Childress (2013) ditantang oleh AI-Health?	Mittelstadt (2019)	Epistemik (B, F) Normatif (A, C, D, F) Menyeluruh (A, F)	Nuffield Council on Bioethics harus memperbarui panduannya terkait prinsip bioetika untuk inisiatif data agar memasukkan pertimbangan khusus terkait AI-Health.
Bagaimana konsep seperti keadilan, akuntabilitas, dan transparansi dapat dipertahankan pada skala besar (dengan anonimitas terjaga)?	Rosenfeld et al. (2019)	Epistemik (C, D, E, F) Normatif (D, E, F) Menyeluruh (F)	Care Quality Commission harus mengembangkan mekanisme pemantauan dampak ini pada skala besar sebagai bagian dari proses tinjauan reguler untuk menjamin keselamatan dan kualitas layanan. Hal ini mungkin memerlukan Department of Health and Social Care memperluas kewenangan regulasinya.

Ditunjukkan dengan meningkatnya Tingkat Analisis: Individu (A), Interpersonal (B), Kelompok (C), Kelembagaan (D), Sektoral (E) dan Masyarakat (F)

Demikian pula, ada upaya untuk mengesahkan kode etik tanpa mempertimbangkan kepentingan dan tantangan yang berlapis-lapis ini. Namun, beberapa perubahan patut dicatat. Inggris telah menerbitkan Kode Etik untuk teknologi kesehatan dan perawatan berbasis data, standar untuk bukti efektivitas klinis untuk teknologi kesehatan digital sebuah standar kuesioner penilaian digital untuk aplikasi dan saat ini sedang mengembangkan model 'regulasi sebagai layanan' untuk memastikan adanya pemeriksaan regulasi yang tepat di semua tahap siklus pengembangan AI.

Organisasi Kesehatan Dunia (WHO) memiliki sejumlah proyek yang sedang berjalan untuk mengembangkan panduan bagi negara-negara anggota (Organisasi Kesehatan Dunia 2019). Di Tiongkok, beberapa norma memberikan instruksi spesifik dan terperinci untuk memastikan keamanan dan kerahasiaan data kesehatan guna memastikan bahwa kumpulan data besar kesehatan dan medis dapat digunakan sebagai sumber daya nasional untuk mengembangkan algoritma demi peningkatan kesehatan masyarakat.

Pertanyaan etis yang terkait dengan penggunaan AI untuk layanan kesehatan merembet ke isu-isu tentang hal-hal apa yang dapat atau harus diatur dalam lingkup layanan kesehatan, dibandingkan dengan apa yang dianggap hanya sebagai layanan kesejahteraan. Oleh karena itu, berpikir dalam kerangka kerja yang diusulkan membantu para pembuat kebijakan memahami dan menggambarkan ruang lingkup regulasi mereka. Misalnya, beberapa perangkat algoritmik berpotensi memungkinkan orang untuk sepenuhnya melewati sistem layanan kesehatan formal dan teregulasi dengan baik dengan mengakses teknologi secara langsung, baik dengan menggunakan perangkat yang dapat dikenakan (*wearable device*) maupun dengan mengakses basis data daring.

Perlu ada diskusi, dengan mempertimbangkan LoA dan kekhawatiran, mengenai apakah layanan-layanan ini secara *de facto* telah melampaui batas dalam layanan kesehatan di salah satu tingkatan tersebut. Demikian pula, meskipun beberapa solusi teknis telah diajukan untuk memitigasi masalah bias data dan kualitas data serta memastikan inklusi sosial dalam pengambilan keputusan, solusi-solusi ini masih relatif belum teruji. Kecuali jika keunggulan kompetitif dari langkah-langkah pro-etika tersebut menjadi jelas tanpa mewajibkan pendekatan-pendekatan ini, kecil kemungkinan langkah-langkah tersebut akan berdampak signifikan terhadap dampak etis AI-Health dalam waktu dekat.

Akibatnya, masih terdapat sedikit kendali atas prosedur yang diikuti dan mekanisme kendali mutu yang terlibat dalam pengembangan, penerapan, dan penggunaan AI-Health. Karena masalah-masalah yang relatif lebih mudah diatasi ini belum memiliki solusi yang memadai, tidak mengherankan bahwa isu-isu yang lebih besar mengenai perlindungan kesetaraan perawatan, distribusi manfaat yang adil dan perlindungan dan promosi nilai-nilai masyarakat hampir tidak dipertimbangkan.

Mengingat bahwa sistem perawatan kesehatan dalam banyak hal bertindak sebagai inti dari masyarakat modern, hal ini mengkhawatirkan. Jika kesalahan dilakukan terlalu dini dalam adopsi dan implementasi AI dalam perawatan kesehatan, dampaknya bisa cukup signifikan untuk merusak kepercayaan publik, yang mengakibatkan biaya peluang yang signifikan, dan berpotensi mendorong individu untuk mencari perawatan kesehatan mereka dari luar sistem

formal di mana mereka mungkin menghadapi risiko yang lebih besar. Pendekatan yang koheren diperlukan dan segera, semoga tinjauan sistematis tentang masalah-masalah yang harus dipertimbangkan ini dapat membantu mempercepat pengembangannya.

15.5 KESIMPULAN

Kesimpulan dari bacaan tersebut adalah bahwa pemanfaatan teknologi AI dalam layanan kesehatan menawarkan potensi besar, namun juga membawa tantangan etika yang perlu ditangani secara serius. Tinjauan pustaka ini berupaya mengelompokkan isu-isu etika ke dalam tema dan tingkat abstraksi tertentu agar pembuat kebijakan dan regulator dapat memahami serta meresponsnya dengan lebih efektif.

Meski begitu, terdapat keterbatasan, seperti keterfokusan pada literatur berbahasa Inggris dan kesenjangan antara perkembangan teknologi dengan regulasi maupun penelitian akademis. Karena itu, penelitian lanjutan yang lebih luas dan studi kasus mendalam diperlukan untuk melengkapi gambaran etika ini, termasuk menilai kebijakan dan regulasi global yang sedang berkembang.

Tujuan akhirnya adalah membangun sistem tanggung jawab yang transparan dan adil, sehingga semua pihak yang terlibat dalam penggunaan AI di bidang kesehatan dapat dimintai pertanggungjawaban dengan tepat. Dengan cara ini, risiko etika dapat diminimalkan dan adopsi teknologi AI dalam layanan kesehatan dapat dilakukan dengan lebih aman dan dapat diterima oleh komunitas medis.

BAB 16

KENDARAAN OTONOM

Mobilitas selalu menjadi bagian mendasar dari kehidupan manusia. Setiap masyarakat bergantung pada kemampuan warganya untuk berpindah, baik untuk bekerja, belajar, berinteraksi sosial, maupun sekadar memenuhi kebutuhan sehari-hari. Oleh karena itu, setiap perubahan besar dalam cara manusia bergerak akan membawa dampak langsung terhadap struktur dan dinamika masyarakat itu sendiri. Dari berbagai bentuk transformasi mobilitas, sulit membayangkan sesuatu yang lebih radikal dibandingkan hadirnya teknologi mengemudi otonom.

Kendaraan otonom atau kendaraan tanpa pengemudi tidak hanya menjanjikan efisiensi dan kenyamanan, tetapi juga berpotensi mengubah cara kita memaknai perjalanan. Perdebatan mengenai manfaat dan kelemahannya pun semakin menguat, karena isu ini menyentuh aspek mendasar: kesejahteraan individu maupun masyarakat secara keseluruhan. Dengan kata lain, memahami sikap dan pandangan terhadap teknologi ini berarti memahami pula arah perkembangan sosial di masa depan.

Dalam bab ini, saya berpendapat bahwa teknologi digital memungkinkan kita untuk memisahkan antara “perjalanan” dan “pengalaman bepergian.” Selama ini, bepergian sering kali dilakukan karena tuntutan kita perlu ke kantor, menghadiri rapat, atau memenuhi kewajiban lain. Namun, dengan hadirnya teknologi otonom, kemungkinan terbuka bagi manusia untuk menikmati pengalaman bepergian itu sendiri, terlepas dari beban mengemudi atau kewajiban praktis. Dengan kata lain, perjalanan dapat berubah dari kebutuhan menjadi pilihan, dari kewajiban menjadi aktivitas yang lebih bebas dan bahkan menyenangkan.

Perubahan mendasar ini bukan sekadar soal teknis transportasi, melainkan menyangkut cara kita menata waktu, ruang, serta pola interaksi dalam kehidupan sehari-hari. Oleh karena itu, membahas kendaraan otonom tidak hanya berarti membicarakan inovasi teknologi, tetapi juga menyelami bagaimana inovasi tersebut akan membentuk kembali fondasi masyarakat modern.

16.1 PENDAHULUAN

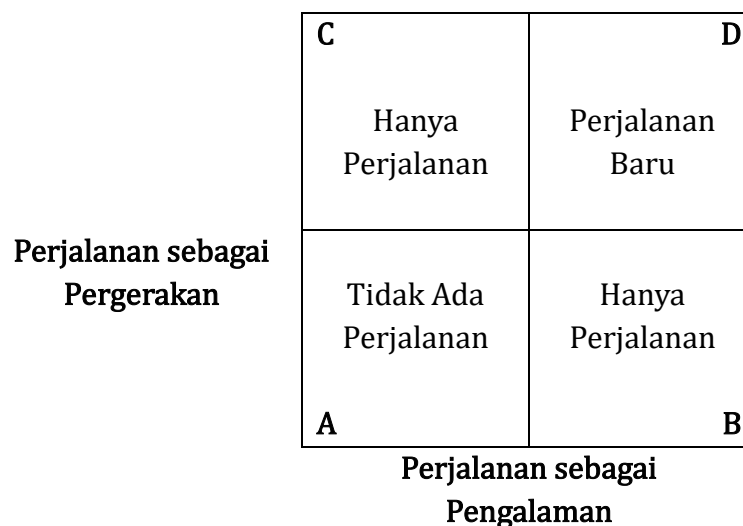
Ketika drama *Death of a Salesman* karya Arthur Miller dipentaskan pada tahun 1949, pekerjaan sebagai penjual keliling yang menawarkan barang dagangan dari satu wilayah ke wilayah lain masih merupakan profesi yang lazim. Zaman itu sangat berbeda dengan masa kini standarisasi kontainer intermodal belum dikenal, konsumerisme sedang tumbuh pesat, dan perjalanan panjang seorang penjual dianggap hal yang wajar. Dalam drama tersebut, tokoh utama Willy Loman merasa lelah dan tidak bahagia dengan rutinitas perjalanannya yang tak ada habisnya, sebuah gambaran yang mencerminkan realitas sosial-ekonomi saat itu.

Hilangnya profesi seperti itu mengingatkan kita bahwa hingga beberapa dekade lalu, banyak orang percaya bahwa perkembangan teknologi digital terutama akan mengurangi kebutuhan manusia untuk melakukan perjalanan fisik. Saat ini, keyakinan tersebut tampak masuk akal: kita berbelanja secara daring, berinteraksi melalui media digital, dan bahkan

menggantikan peran agen penjualan tradisional dengan sistem rekomendasi berbasis algoritma, yang mampu menavigasi ruang informasi atau *infosphere*.

Namun, membatasi pandangan pada gagasan bahwa revolusi digital hanya mengurangi mobilitas jelas terlalu sederhana. Memang benar bahwa kita kini lebih banyak memindahkan byte dibandingkan atom, tetapi hal itu bukan keseluruhan cerita. Digitalisasi tidak semata-mata menghapus kebutuhan akan perjalanan, melainkan juga mengubah bentuk dan maknanya. Sebagian pihak, misalnya para pendukung kendaraan tanpa pengemudi, masih berargumen bahwa teknologi baru akan semakin mengurangi mobilitas manusia. Bahkan, ada yang menggunakan retorika ramah lingkungan untuk mendukung pandangan tersebut. Akan tetapi, argumen ini keliru. Justru, dengan meningkatnya otomatisasi, lebih banyak orang yang sebelumnya tidak dapat mengemudi akan mampu melakukannya di masa depan.

Dengan demikian, lebih tepat dikatakan bahwa teknologi digital tidak sekadar mengurangi mobilitas, tetapi merekonstruksi ulang hakikatnya sebuah proses yang dapat disebut *reontologisasi*. Perubahan ini berlangsung melalui setidaknya empat dimensi utama, sebagaimana ditunjukkan pada Gambar 19.1.



Gambar 16.1 Dua komponen perjalanan: perjalanan fisik dan perjalanan pengalaman

16.2 PERJALANAN, PERJALANAN, DAN TRIP

Setiap perjalanan manusia pada dasarnya terdiri dari dua komponen utama yang dapat dipetakan dalam dua sumbu. Pertama, sumbu y yang merepresentasikan perjalanan fisik, yakni perpindahan nyata seseorang dari satu lokasi geografis ke lokasi lain, seperti dari rumah menuju kantor. Komponen ini bersifat terukur, karena mencakup aspek-aspek seperti biaya, jarak, waktu tempuh, kecepatan, serta moda transportasi yang digunakan. Kedua, sumbu x yang mencerminkan perjalanan pengalaman, yaitu aspek subjektif yang melekat dalam proses perjalanan, misalnya kenyamanan, kenikmatan, kebaruan, maupun persepsi durasi perjalanan. Dengan kata lain, perjalanan tidak hanya dipahami secara fisik, tetapi juga sebagai pengalaman

yang dapat menyenangkan maupun membosankan. Revolusi digital kemudian mengubah lanskap mobilitas secara mendalam, dengan memengaruhi kedua komponen ini sekaligus.

Pada titik A dalam skema tersebut, terdapat opsi *zero-travelling*, yaitu kondisi di mana perjalanan fisik ditiadakan karena semua aktivitas berlangsung di dalam *infosfer*. Telepresensi dan kerja jarak jauh adalah contoh konkret yang mengurangi kebutuhan mobilitas manusia secara langsung. Namun, penurunan ini sering kali bergeser menjadi peningkatan perjalanan dalam bentuk lain, misalnya melonjaknya jasa kurir akibat belanja daring. Artinya, berkurangnya mobilitas manusia justru memicu tumbuhnya mobilitas buatan melalui agen-agen non-manusia seperti drone, robot logistik, hingga kapal tanpa awak. Dengan demikian, semakin besar penggunaan opsi *zero-travelling*, semakin tinggi pula kebutuhan pada bentuk mobilitas buatan.

Berbeda dengan titik D, yang menunjukkan munculnya bentuk mobilitas baru hasil inovasi digital. Pada titik ini, perjalanan fisik dan pengalaman justru menyatu. Sepeda motor atau mobil sport, misalnya, tidak hanya berfungsi untuk berpindah tempat, melainkan juga untuk menghadirkan pengalaman berkendara yang menyenangkan. Karena alasan inilah, sulit membayangkan mobil sport mewah seperti Ferrari dibuat sepenuhnya tanpa pengemudi, sebab daya tarik utamanya justru ada pada pengalaman mengemudi itu sendiri. Revolusi digital, alih-alih menghapus aspek pengalaman, justru memperkaya nilai perjalanan dengan menghadirkan mobilitas yang lebih aman, murah, sekaligus memuaskan secara etis, seperti terlihat pada tren sistem berbagi sepeda di perkotaan.

Sementara itu, pada titik B, teknologi digital bekerja untuk mengubah perjalanan menjadi pengalaman yang semata-mata menyenangkan dengan mengurangi aspek-aspek yang dianggap membosankan. Contoh yang jelas terlihat pada kapal pesiar, di mana perjalanan fisik sepenuhnya terintegrasi dengan pengalaman perjalanan itu sendiri. Fasilitas hiburan di kendaraan modern, sistem navigasi pintar, hingga asisten virtual, semuanya berkontribusi mengubah perjalanan fisik menjadi pengalaman yang lebih nyaman dan menyenangkan. Bahkan, masa depan mungkin menghadirkan mobilitas virtual, di mana liburan tidak lagi memerlukan perjalanan fisik, melainkan digantikan oleh pengalaman digital yang imersif.

Kemudian, pada titik C, fokus bergeser pada mobilitas efisien yang semakin mengecualikan peran manusia. Di sini, perjalanan ditransformasikan menjadi sistem tanpa awak, yang menekankan efisiensi, keamanan, biaya rendah, serta kecepatan. Contoh utamanya adalah otomatisasi transportasi barang, terutama di sektor maritim, yang kini semakin mengandalkan sistem cerdas. Tren ini juga dapat dilihat pada otomatisasi transportasi publik di ruang terbatas, seperti bus bandara atau kendaraan otonom di jalur khusus. Meski secara logis kendaraan otonom penuh (level-5) mungkin terwujud di masa depan, kenyataan teknis menunjukkan bahwa keberhasilan lebih realistis dicapai melalui rekayasa ulang lingkungan agar lebih selaras dengan teknologi. Alih-alih mengandalkan sistem visual yang sempurna untuk mengenali rambu lalu lintas dalam kondisi ekstrem, solusi yang lebih layak adalah membangun rambu pintar yang berkomunikasi langsung dengan kendaraan melalui sinyal digital.

Dengan demikian, peta perjalanan dalam empat kuadran ini menggambarkan transformasi mobilitas manusia di era digital: mulai dari penghapusan perjalanan fisik, penciptaan pengalaman baru, peningkatan kenyamanan, hingga peralihan pada efisiensi berbasis otomatisasi. Semua ini menunjukkan bahwa masa depan mobilitas tidak hanya tentang bergerak dari satu tempat ke tempat lain, tetapi juga bagaimana teknologi mendefinisikan ulang makna perjalanan itu sendiri.

16.3 FOKUS PADA 'BAGAIMANA', BUKAN 'KAPAN' ATAU 'DI MANA'

Mobilitas telah lama menjadi salah satu fondasi kehidupan masyarakat modern. Sejak awal abad ke-21, diperkirakan lebih dari satu miliar mobil penumpang melintasi jalan raya dunia, memainkan peran vital bukan hanya sebagai sarana transportasi, tetapi juga sebagai media rekreasi dan simbol status sosial. Kendaraan, dengan demikian, bukan sekadar alat mekanis, melainkan bagian dari sejarah, budaya, dan identitas manusia. Transformasi dalam cara kita bergerak, terutama melalui kehadiran kendaraan otonom, membawa implikasi yang sangat besar. Sulit membayangkan perubahan dalam mobilitas yang lebih mendalam daripada pergeseran dari kendaraan konvensional ke sistem yang sepenuhnya mengandalkan kecerdasan buatan. Oleh sebab itu, memahami sikap masyarakat terhadap manfaat dan kekurangan mengemudi otonom menjadi kunci untuk menilai dampaknya terhadap kesejahteraan individu maupun sosial secara keseluruhan.

Seperti telah disinggung sebelumnya, teknologi digital memungkinkan kita memisahkan perjalanan dari makna perjalanan itu sendiri. Dalam waktu dekat, manusia berpotensi menikmati pengalaman berkendara tanpa harus terbebani oleh kewajiban untuk mengemudi. Perjalanan dapat berubah menjadi aktivitas yang lebih fleksibel: dilakukan bukan semata karena kebutuhan, tetapi juga karena keinginan. Studi internasional (AUDI, 2019) menunjukkan bahwa masyarakat di sembilan negara mulai dari Tiongkok, Prancis, Jerman, Italia, Jepang, Korea Selatan, Spanyol, Inggris, hingga Amerika Serikat menyimpan kombinasi rasa ingin tahu, minat, sekaligus kekhawatiran terhadap teknologi ini.

Mayoritas responden survei mengekspresikan minat (82%) dan rasa ingin tahu (62%) terhadap kendaraan otonom. Namun, rasa optimisme itu diimbangi dengan kekhawatiran: 70% takut kehilangan kendali, 66% menyoroti adanya risiko teknis yang tidak dapat sepenuhnya dihindari, dan 65% mengkhawatirkan belum adanya kerangka hukum yang jelas. Perbedaan sikap ini dapat dipahami. Menghargai sebuah inovasi hanya menuntut keterbukaan pikiran, tetapi menerima perubahan menuntut komitmen yang melibatkan risiko dan biaya nyata. Hal ini tercermin dari hanya 28% responden yang bersedia membayar lebih untuk kendaraan otonom. Dengan kata lain, mengemudi otonom adalah fenomena ganda: sebuah kebaruan teknologi sekaligus perubahan sosial yang belum pernah terjadi sebelumnya. Untuk mengubah rasa ingin tahu menjadi penerimaan yang lebih luas, diperlukan jaminan keamanan, kemajuan teknologi, serta kerangka hukum dan etika yang kokoh.

Lebih jauh, hanya 8% responden yang merasa mampu menjelaskan konsep mengemudi otonom. Sekilas angka ini tampak rendah, bahkan seolah meragukan validitas survei. Namun, kenyataannya tidak demikian. Pengalaman dengan transmisi mobil dapat menjadi analogi. Di

Amerika Serikat, pada 2018, hanya 3,7% mobil yang terjual melalui CarMax masih menggunakan transmisi manual, sementara transmisi otomatis mendominasi pasar. Meski demikian, sangat sedikit pengemudi yang mampu menjelaskan secara teknis perbedaan antara transmisi variabel konstan, transmisi kopling ganda, dan transmisi otomatis sederhana. Fakta ini menunjukkan bahwa sikap masyarakat biasanya dibangun berdasarkan pengalaman dan keyakinan praktis, bukan pada pengetahuan teknis mendalam.

Dengan demikian, penting untuk tidak meremehkan sikap masyarakat hanya karena sebagian besar dari mereka tidak dapat menjelaskan detail teknisnya. Fakta bahwa 90% responden sudah pernah mendengar tentang kendaraan otonom dan 30% mengaku cukup mengenalnya menunjukkan bahwa kesadaran publik relatif tinggi. Kesimpulan besar dari survei tersebut dapat dirangkum dalam satu kata: variasi. Pertanyaan utama tentang masa depan kendaraan otonom bukanlah “kapan” atau “di mana”, melainkan “bagaimana” teknologi ini akan diwujudkan. Kunci keberhasilan terletak pada penyediaan beragam pilihan tingkat otonomi yang dapat disesuaikan dengan kebutuhan, preferensi, sikap, dan kondisi sosial-ekonomi masyarakat. Dengan pendekatan yang berfokus pada fleksibilitas, kendaraan otonom berpotensi diterima lebih luas dan memberi dampak transformasional yang lebih adil serta berkelanjutan.

16.4 KESIMPULAN

Dalam film fiksi ilmiah yang berkualitas rendah, sering kali digambarkan bahwa semua kendaraan tampak baru, seragam, dan hanya terdiri atas beberapa model saja. Gambaran tersebut jelas menyederhanakan kenyataan, sebab dunia nyata jauh lebih kompleks, berlapis, dan penuh dengan keragaman, layaknya mesin yang berminyak dan lengket ketika digunakan secara nyata. Karena itu, kebijakan publik maupun strategi bisnis terkait pengembangan kendaraan otonom seharusnya tidak memandang variasi sebagai kelemahan, melainkan menjadikannya sebagai kekuatan utama. Upaya tersebut harus diarahkan pada rekayasa ulang lingkungan secara menyeluruh, sehingga tercipta ekosistem yang mendukung keberadaan kendaraan otonom sebagai bagian normal dari kehidupan sehari-hari. Harapannya, seluruh inovasi ini tidak hanya memungkinkan penerapan kendaraan otonom secara luas, tetapi juga memastikan bahwa teknologi tersebut lebih ramah lingkungan dan berkelanjutan dibandingkan kendaraan konvensional yang kita gunakan saat ini.

BAB 17

KEADILAN DAN TATA KELOLA AI DALAM SEKTOR KEUANGAN

Perkembangan kecerdasan buatan (AI) telah membawa dampak besar dalam berbagai sektor, termasuk industri jasa keuangan (JK). Banyak perusahaan mulai mengadopsi solusi berbasis AI untuk meningkatkan efisiensi operasional, memperoleh wawasan strategis, serta memperkuat keterlibatan dengan pelanggan. Potensi AI dalam industri ini sangat luas, mulai dari otomatisasi proses bisnis, deteksi penipuan, analisis risiko, hingga penyediaan layanan personal bagi nasabah.

Namun, tingkat adopsi AI di sektor jasa keuangan masih tergolong rendah. Salah satu penyebab utama adalah adanya kekhawatiran mengenai kompleksitas teknologi dan sifat belajar mandiri yang dimiliki AI. Dalam industri yang sangat teregulasi, hal ini menimbulkan tantangan besar terutama terkait dengan aspek auditabilitas. Bagaimana memastikan bahwa sistem AI berjalan sesuai aturan, adil, transparan, dan tidak menimbulkan risiko baru bagi perusahaan maupun masyarakat menjadi pertanyaan mendasar yang harus dijawab.

Sayangnya, literatur yang membahas secara khusus mengenai tata kelola dan kontrol atas penerapan AI dalam jasa keuangan masih terbatas. Padahal, audit terhadap AI tidak bisa hanya berfokus pada algoritmanya saja. Risiko yang ditimbulkan AI meresap ke dalam keseluruhan organisasi, memengaruhi keputusan bisnis, kepatuhan regulasi, hingga kepercayaan publik.

Bab ini akan membahas strategi tata kelola dan kerangka kerja kontrol yang menyeluruh untuk membantu perusahaan jasa keuangan mengelola risiko AI. Dengan mengambil contoh risiko ketidakadilan, akan ditunjukkan bagaimana AI berpotensi menimbulkan masalah etika dan regulasi jika tidak dikelola dengan baik. Kerangka kerja yang dibahas dalam bab ini diharapkan dapat memberikan panduan praktis bagi para pemimpin industri untuk memitigasi risiko sekaligus tetap berinovasi dengan percaya diri.

17.1 PENDAHULUAN

Adopsi kecerdasan buatan (AI) dalam sektor jasa keuangan hingga kini masih berada pada tahap awal. Hal ini tercermin dari survei terbaru terhadap lebih dari 3.000 eksekutif tingkat atas yang dilakukan oleh Deloitte dan Asosiasi Manajemen Keuangan Eropa (EFMA), yang menunjukkan bahwa 11% responden sama sekali belum memulai aktivitas di bidang AI, sementara 40% lainnya masih sebatas mempelajari bagaimana teknologi ini dapat diterapkan di organisasi mereka. Dalam pembahasan ini, istilah AI digunakan secara luas untuk merujuk pada berbagai teknik yang mengandalkan pembelajaran mesin guna menyelesaikan tugas-tugas yang biasanya membutuhkan kecerdasan manusia, seperti pemrosesan bahasa alami, pengenalan suara, hingga pengambilan keputusan dalam kondisi yang penuh ketidakpastian. Sebelum hadirnya teknologi ini, sebagian besar tugas tersebut ditangani melalui proses berbasis manusia atau aturan sistemik, misalnya pemberian pinjaman yang ditentukan oleh kebijakan manajer bank atau penggunaan kartu skor untuk menghitung tingkat risiko nasabah.

Kemajuan dalam ketersediaan daya komputasi yang lebih murah, ditambah dengan peningkatan volume dan variasi data dalam jumlah besar, telah mendorong lahirnya algoritma yang lebih canggih. Algoritma ini memungkinkan analisis informasi dilakukan secara lebih cepat dan mendalam. Alat AI dapat bersifat statis seperti model prediksi pendapatan yang diperbarui setiap kuartal fiskal atau bersifat dinamis dengan kemampuan belajar berkelanjutan melalui umpan balik waktu nyata, contohnya chatbot yang terus memperbaiki kualitas responsnya berdasarkan masukan pengguna. Meskipun adopsi AI berjalan lambat, perusahaan jasa keuangan kini mulai meneliti cara untuk memanfaatkannya dalam rangka menekan biaya dan mempertahankan daya saing. Survei menunjukkan bahwa 65% eksekutif perbankan melihat potensi terbesar AI pada peningkatan layanan pelanggan, sedangkan 78% eksekutif asuransi menilai bahwa AI paling bermanfaat untuk mendukung operasional dan fungsi back office.

Namun, perlu dicatat bahwa sektor jasa keuangan merupakan industri yang sangat diatur, dengan beragam lini bisnis serta produk yang kompleks. Riwayat panjang sanksi akibat ketidakpatuhan membuat perusahaan berhati-hati dalam menerapkan teknologi baru, terlebih dengan meningkatnya pengawasan regulator terhadap penggunaan AI. Dalam setahun terakhir, Otoritas Perilaku Keuangan (FCA) Inggris bersama Kantor Komisioner Informasi (ICO) telah aktif memberikan pandangan mengenai AI dan pembelajaran mesin. Walaupun belum ada regulasi spesifik yang dirancang untuk AI, regulator menyampaikan ekspektasi mereka melalui sejumlah publikasi, misalnya terkait perdagangan algoritmik, pengawasan model internal, serta Manajer Senior dan Skema Sertifikasi (SM&CS). Mengingat perkembangan teknologi yang jauh lebih cepat dibandingkan dengan perkembangan regulasi, regulator tetap memegang prinsip *netralitas teknologi*. Artinya, prinsip regulasi yang berlaku tetap sama tanpa memandang jenis teknologi yang digunakan perusahaan dalam menjalankan aktivitasnya. Prinsip ini juga dapat dipandang sebagai gambaran awal mengenai bagaimana AI berpotensi diatur di masa mendatang.

Kajian literatur mengenai penggunaan kecerdasan buatan (AI) selama ini umumnya berfokus pada aspek teknis, seperti teknik, alat, dan metodologi yang dirancang untuk menjamin keadilan, akuntabilitas, serta transparansi algoritma. Namun, hanya sedikit penelitian yang berusaha menempatkan temuan-temuan tersebut dalam kerangka regulasi, sementara hubungan antara pendekatan teknis dengan proses tata kelola organisasi sering kali terabaikan. Meskipun telah banyak dikembangkan berbagai formalisasi matematis mengenai konsep keadilan, penerapan praktisnya di industri masih menyisakan ketidakpastian, khususnya dalam hal bagaimana memastikan algoritma benar-benar berjalan secara adil.

Untuk menjawab kesenjangan tersebut, makalah ini akan menggunakan risiko diskriminasi sebagai studi kasus untuk menelaah tantangan nyata dalam pengelolaan risiko yang muncul akibat penggunaan AI. Pembahasan akan diarahkan pada siklus hidup produk AI, dimulai dari tahap identifikasi risiko, penilaian, pengendalian, hingga pemantauan yang dilakukan dalam konteks perusahaan jasa keuangan. Dalam menguraikan hal tersebut, makalah ini akan mengacu pada praktik dan prinsip yang direkomendasikan oleh publikasi

regulator, dengan fokus utama pada regulasi yang berkembang di Uni Eropa dan Inggris, meskipun tidak menutup kemungkinan merujuk pada aturan eksternal lain yang relevan.

17.2 KEADILAN DALAM INDUSTRI JASA KEUANGAN

Pemanfaatan pembelajaran mesin dalam industri asuransi jiwa semakin meluas, terutama dalam pengambilan keputusan penting yang berkaitan dengan nasabah. Teknologi ini kini digunakan untuk mengevaluasi kelayakan kredit, memberikan rekomendasi produk investasi, hingga menentukan besaran premi asuransi. Tidak hanya berimplikasi pada nasabah, pembelajaran mesin juga berdampak pada karyawan melalui penggunaan algoritma untuk menyaring *curriculum vitae* (CV) dan mengukur kinerja kerja.

Secara tradisional, perusahaan asuransi jiwa hanya memanfaatkan jenis data yang terbatas dan secara langsung relevan dengan hasil yang ingin dicapai. Misalnya, dalam asuransi mobil, data yang dipertimbangkan biasanya mencakup riwayat mengemudi dan lama pengalaman mengemudi. Begitu pula dalam menilai risiko kredit, indikator yang digunakan adalah rasio utang terhadap pendapatan serta riwayat pembayaran sebelumnya. Namun, dengan hadirnya analitik big data, perusahaan mulai menggunakan data non-tradisional sebagai proksi risiko. Praktik ini menimbulkan kontroversi, karena penetapan harga asuransi ternyata dapat dipengaruhi oleh faktor-faktor yang tidak relevan, seperti nama domain email atau nama belakang pemohon, sementara keputusan pemberian kredit dapat bergantung pada riwayat penelusuran internet seseorang.

Pada masa sebelum big data dan algoritma pembelajaran mesin, perusahaan dapat menghindari tuduhan diskriminasi dengan menyatakan bahwa perlakuan berbeda terhadap kelompok tertentu tidak disengaja, sebab atribut yang dilindungi secara hukum memang tidak dimasukkan ke dalam proses pengambilan keputusan. Akan tetapi, sistem berbasis AI justru menimbulkan tantangan baru karena sifatnya yang kurang transparan dibandingkan proses berbasis aturan tradisional. Kemampuannya mengekstrak pola dari hubungan antarfitur yang kompleks sering kali menyulitkan pemangku kepentingan untuk menelusuri dasar pengambilan keputusan.

Lebih jauh, perkembangan hukum terbaru telah menggeser fokus dari aspek niat diskriminatif menuju dampak diskriminatif. Di Amerika Serikat, Mahkamah Agung dalam kasus *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.* menegaskan bahwa diskriminasi yang tidak disengaja tetap ilegal apabila terbukti memberikan dampak yang tidak proporsional terhadap kelompok yang dilindungi berdasarkan Undang-Undang Perumahan yang Adil. Sementara itu, di Inggris, perkara *Essop v Home Office* juga menegaskan bahwa diskriminasi tidak langsung dalam praktik perekrutan merupakan pelanggaran hukum. Dengan demikian, perkembangan hukum ini semakin mempertegas pentingnya transparansi dan akuntabilitas dalam penerapan algoritma pembelajaran mesin di sektor jasa keuangan dan asuransi.

Seiring dengan semakin kompleksnya hubungan dalam data sosial yang tertanam dalam algoritma *black box*, persoalan diskriminasi dalam pengambilan keputusan berbasis AI menjadi semakin menonjol. Pergeseran fokus regulasi dari niat diskriminatif menuju dampak

diskriminatif menuntut adanya pendekatan baru untuk mengidentifikasi kerugian yang mungkin timbul dari penggunaan alat pengambilan keputusan otomatis. Berbeda dengan proses berbasis manusia yang cenderung menghasilkan keputusan yang bervariasi akibat adanya subjektivitas dan bias personal, algoritma AI justru berpotensi mereplikasi bias yang sama secara konsisten dalam skala besar. Hal ini tidak hanya memperkuat bias yang sudah ada, tetapi juga menghilangkan ruang ambiguitas yang sebelumnya terdapat dalam pengambilan keputusan manusia, sehingga memperkristal ketidakadilan.

Risiko ini menjadi sangat mengkhawatirkan pada domain yang memiliki sejarah panjang diskriminasi yang terdokumentasi dengan baik, misalnya dalam bidang ketenagakerjaan, perumahan, atau akses kredit. Dalam konteks tersebut, AI dapat secara tidak sengaja memperburuk ketidaksetaraan sosial yang sudah ada. Meskipun tujuan awal penerapan AI adalah untuk mendukung dan meningkatkan kualitas pengambilan keputusan manusia alih-alih menggantikannya sepenuhnya, kenyataannya banyak pengguna bisnis yang kesulitan memahami keluaran teknis algoritma, seperti interval kepercayaan. Kondisi ini membuat mereka cenderung menerima hasil algoritma begitu saja tanpa mempertanyakan validitasnya, bahkan ketika algoritma tersebut mungkin menghasilkan keputusan yang tidak adil.

Dari sisi hukum, terdapat ruang pembenaran dalam kasus diskriminasi tidak langsung, asalkan pihak tertuduh dapat membuktikan adanya kebutuhan bisnis yang sah. Namun, standar bukti yang diperlukan untuk pembenaran ini masih belum jelas dan jarang dikaji secara mendalam. Di Amerika Serikat, misalnya, *business necessity clause* menetapkan bahwa dampak diskriminatif dapat dibenarkan apabila hal itu diperlukan untuk memenuhi batasan kinerja tertentu, dengan catatan bahwa dampak berbeda yang ditimbulkan harus ditekan seminimal mungkin.

Sementara itu, di Inggris Raya, putusan *Essop v Home Office* memperluas pemahaman ini dengan menyatakan bahwa suatu ketentuan, kriteria, atau praktik (*Provision, Criterion, or Practice/PCP*) dapat dianggap sah apabila terbukti merupakan “*cara yang proporsional untuk mencapai tujuan yang sah*.” Dengan demikian, persoalan diskriminasi dalam algoritma AI tidak hanya menuntut solusi teknis berupa perbaikan model, tetapi juga menimbulkan pertanyaan mendasar terkait regulasi, etika, serta tata kelola risiko di organisasi. Tanpa adanya kerangka kerja yang jelas, AI berpotensi memperkuat ketidakadilan sosial alih-alih menjadi alat untuk mengatasinya.

Pada Juli 2018, OJK menulis bahwa meskipun secara tradisional mereka berfokus pada keadilan prosedural dalam menilai perilaku perusahaan, terdapat beberapa kasus yang memerlukan intervensi untuk memastikan keadilan distributif dalam diskriminasi harga. OJK mencantumkan enam pertanyaan berbasis bukti untuk menilai apakah intervensi diperlukan:

- ➡ kerentanan pelanggan;
- ➡ skala dampak buruk;
- ➡ jumlah orang yang terdampak;
- ➡ kurangnya transparansi dalam metodologi penetapan harga;
- ➡ sifat esensial produk atau layanan; dan

➡ pandangan masyarakat tentang ketidakadilan.

Hal ini menunjukkan langkah lebih lanjut dalam fokus regulator pada dampak daripada niat, dan organisasi perlu beralih ke analisis berbasis hasil untuk mengetahui apakah proses mereka adil. Dalam bab ini akan menggunakan risiko diskriminasi yang melanggar hukum sebagai contoh dalam mengeksplorasi bagaimana perusahaan jasa keuangan (FS) akan mengelola risiko ini di sepanjang siklus hidup produk solusi AI.

17.3 MENGELOLA RISIKO AI SEPANJANG SIKLUS HIDUPNYA

Penelitian akademis tentang AI selama ini cenderung berfokus pada risiko model dan algoritmik seperti bias, akurasi, serta kemampuan prediksi secara terpisah. Namun, dalam praktiknya, desain dan kinerja suatu model tidak dapat dilepaskan dari konteks yang lebih luas. Model AI beroperasi di dalam ekosistem organisasi yang kompleks, sehingga efektivitas dan keadilannya juga dipengaruhi oleh berbagai domain risiko non-model. Domain tersebut mencakup risiko regulasi dan kepatuhan, risiko teknologi, risiko sumber daya manusia, risiko pemasok, risiko perilaku, hingga risiko pasar. Dengan demikian, pengelolaan risiko AI harus dipahami secara holistik, tidak semata-mata sebagai persoalan teknis.

Sebagai contoh, menilai kewajaran (*fairness*) suatu model tidak cukup dipandang sebagai isu matematis atau komputasional. Proses penentuan definisi kewajaran, metode penilaian, hingga strategi remediasinya harus mempertimbangkan kerangka regulasi yang berlaku pada kasus penggunaan tertentu. Selain itu, keterbatasan teknologi yang mungkin memengaruhi implementasi juga perlu diperhitungkan. Tidak kalah penting, keselarasan antara penggunaan AI dengan selera risiko organisasi, nilai etika, dan prinsip inti perusahaan harus menjadi landasan utama. Dengan kata lain, kewajaran dalam AI adalah hasil interaksi antara aspek teknis, hukum, sosial, dan etika.

Adopsi AI pada dasarnya tidak mengharuskan perusahaan untuk merombak *Risk Management Framework* (RMF) yang sudah ada. Yang lebih dibutuhkan adalah kesadaran baru tentang bagaimana AI memunculkan tantangan unik dalam mendeteksi risiko, karena manifestasi risikonya sering kali tidak lazim dibandingkan praktik tradisional. Misalnya, volume dan kecepatan data yang diproses oleh algoritma dapat menuntut kemampuan reaksi yang jauh lebih cepat terhadap setiap kesalahan. Selain itu, kompleksitas algoritma pembelajaran mesin sering kali menghambat transparansi, sehingga menyulitkan proses penjelasan (*explainability*) maupun audit yang efektif.

Oleh sebab itu, para pengawas dan regulator berharap agar perusahaan memiliki tata kelola risiko yang kuat dan efektif. RMF yang berlaku harus mampu mengidentifikasi, mengurangi, dan mengendalikan risiko yang muncul dari pengembangan maupun penggunaan berkelanjutan setiap aplikasi AI di seluruh lini bisnis. Lebih jauh, RMF tersebut tidak hanya bersifat administratif, melainkan juga harus memperoleh persetujuan dari dewan direksi, sehingga akuntabilitas terhadap penerapan AI benar-benar berada pada tingkat strategis organisasi.

17.4 DESAIN PERMODALAN AI DALAM KEUANGAN

Laporan terbaru dari Financial Conduct Authority (FCA) menekankan pentingnya perusahaan untuk secara jelas mendefinisikan perdagangan algoritmik, dengan tujuan agar mereka memiliki proses yang tepat dalam mengidentifikasi serta mengelola penggunaannya. FCA bahkan dapat meminta perusahaan untuk menyerahkan deskripsi strategi perdagangan algoritmik mereka dalam jangka waktu 14 hari. Hal serupa berlaku bagi perusahaan jasa keuangan (financial services/FS), yang juga dituntut untuk mendefinisikan secara tegas apa yang dimaksud dengan teknologi atau solusi berbasis kecerdasan buatan (AI). Dengan demikian, perusahaan harus mampu membedakan antara AI dengan sistem berbasis aturan (rule-based system), otomatisasi proses robotik (RPA), maupun model matematika statis. Kejelasan ini menjadi penting agar manajemen dan karyawan dapat memahami batasan serta implikasinya.

Lebih jauh, cakupan definisi tersebut harus mencerminkan konsekuensi regulasi yang melekat pada penggunaan AI dalam perusahaan. Saat ini, semakin banyak solusi AI yang dibangun menggunakan algoritma pembelajaran mesin pihak ketiga sebagai akselerator pengembangan. Walaupun pembangunan dilakukan melalui pihak luar, tanggung jawab akhir atas seluruh risiko yang muncul tetap berada di tangan perusahaan jasa keuangan. Misalnya, chatbot perbankan ritel yang menggunakan Application Programming Interface (API) berbasis Pemrosesan Bahasa Alami (*Natural Language Processing/NLP*) dari penyedia eksternal, tetap harus masuk ke dalam cakupan *AI Risk Management Framework* (AI RMF). Hal ini karena akuntabilitas atas setiap potensi pelanggaran hukum atau peraturan tetap menjadi kewajiban perusahaan, bukan penyedia teknologi. FCA bahkan menegaskan bahwa sekalipun terdapat praktik alih daya teknis, perusahaan tetap sepenuhnya bertanggung jawab terhadap kepatuhan regulasi.

Identifikasi dan Penilaian Risiko

Agar perusahaan dapat mengidentifikasi dan menilai dampak kasus penggunaan AI terhadap selera risiko mereka, mereka harus terlebih dahulu mengembangkan serangkaian kriteria penilaian yang jelas dan konsisten untuk diterapkan pada semua kasus tersebut. Perusahaan harus mengidentifikasi domain risiko yang relevan untuk suatu solusi serta risiko produk tertentu, kemudian menilai apakah tingkat risiko residual dapat diterima dengan mempertimbangkan pengendalian yang ada. Sangat penting bahwa proses penilaian dan manajemen risiko tidak membatasi kreativitas.

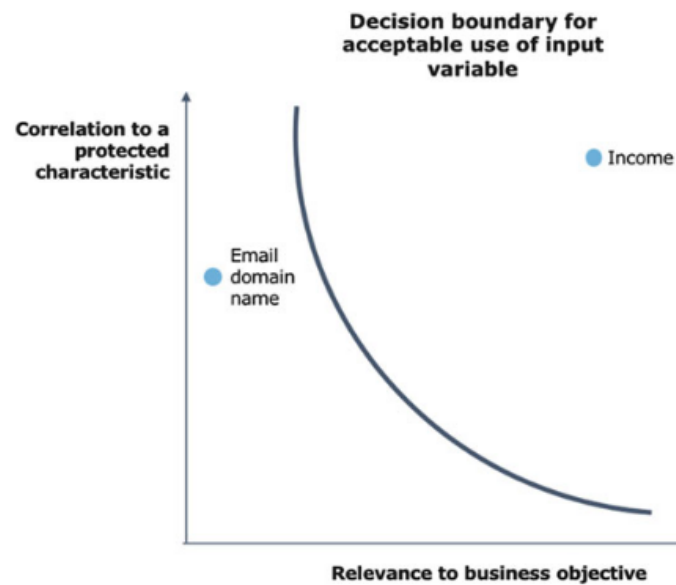
Tujuan utamanya adalah memastikan risiko diidentifikasi sejak dini dan dikelola dengan baik untuk menciptakan lingkungan yang aman bagi inovasi. Beberapa pertimbangan yang relevan meliputi:

- a. Eksternal vs. internal: Target audiens solusi AI akan menentukan implikasi risiko perilaku, serta tingkat kepercayaan dan kinerja ambang batas yang harus dicapai solusi tersebut sebelum penerapan. Misalnya, model penetapan harga asuransi dengan antarmuka pengguna pelanggan memiliki risiko hasil yang tidak adil yang lebih tinggi daripada model validasi pendapatan yang digunakan oleh karyawan.

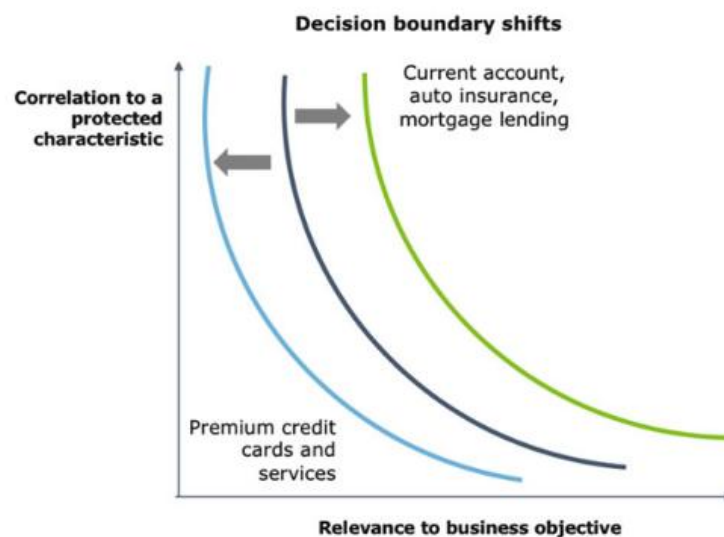
- b. Penggunaan informasi pribadi: Berdasarkan Peraturan Perlindungan Data Umum (GDPR) di Eropa, Penilaian Dampak Privasi harus dilakukan jika organisasi berencana untuk memproses data pribadi. Algoritma yang menggunakan informasi pribadi untuk pengambilan keputusan harus dinilai kewajarannya. Selain itu, GDPR memberi konsumen hak tambahan untuk memahami dan mengendalikan bagaimana perusahaan menggunakan data pribadi mereka. Komisioner Informasi Inggris menyatakan bahwa *"jika suatu keputusan telah dibuat oleh mesin yang berdampak signifikan terhadap seseorang, GDPR mewajibkan mereka memiliki hak untuk menggugat keputusan tersebut dan hak untuk mendapatkan penjelasan."* Meskipun terdapat perbedaan pendapat di antara para akademisi mengenai definisi dan dasar hukum untuk "hak atas penjelasan" ini, perusahaan tetap harus memiliki proses untuk menanggapi pertanyaan pelanggan dengan cara yang bermakna, transparan, dan mudah dipahami, serta mampu menunjukkan bahwa suatu algoritma mematuhi persyaratan perlindungan data.
- c. Akurasi dan kualitas data: Semua data masukan dan pelatihan ke dalam model pembelajaran mesin harus berkualitas tinggi dan sesuai dengan tujuannya. Ini mencakup peninjauan metodologi pengumpulan data untuk potensi bias seleksi dan evaluasi distribusi hasil untuk kemungkinan bias terhadap kelas yang dilindungi. Pandangan masyarakat tentang ketidakadilan, selain menjadi kriteria FCA untuk intervensi, dapat menyebabkan kerusakan reputasi. Ketika sebuah investigasi mengungkapkan bahwa pengendara bernama Mohammed dikenakan biaya asuransi mobil hingga Rp 17.920.500 lebih mahal daripada pria dengan nama khas kulit putih dan Inggris, hal itu memicu protes publik dan seruan untuk boikot. Untuk menghindari pengawasan semacam itu, fitur-fitur yang dimasukkan ke dalam model harus dinilai kesesuaiannya.

Gambar 17.1 memvisualisasikan kemungkinan batasan keputusan untuk menentukan apakah suatu variabel input harus digunakan dalam suatu model atau tidak. Mengingat keputusan dalam kasus *Essop v Home Office* dan *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, meskipun suatu fitur berkorelasi dengan karakteristik yang dilindungi, mungkin ada alasan yang wajar untuk menggunakannya untuk tujuan bisnis.

Misalnya, di antara pemohon pinjaman, tingkat pendapatan mungkin berbeda antara pria dan wanita karena lebih banyak wanita bekerja paruh waktu. Pendapatan tetap dapat digunakan dalam keputusan pemberian pinjaman karena relevansinya yang tinggi terhadap risiko gagal bayar. Sebaliknya, nama domain email mungkin prediktif terhadap risiko, tetapi jika sangat berkorelasi dengan ras, nama tersebut mungkin perlu dihapus dari model karena kurangnya dasar hubungan sebab akibat dengan risiko (Gambar. 17.2).



Gambar 17.1 Batasan keputusan untuk penggunaan variabel input yang dapat diterima



Gambar 17.2 Pergeseran batas keputusan

Batas keputusan ini dapat bergeser tergantung pada kondisi yang digariskan oleh FCA untuk kemungkinan intervensi. Faktor pendorong pengambilan keputusan dalam menyediakan produk-produk penting, seperti rekening giro, asuransi mobil, atau hipotek, mungkin tunduk pada pengawasan yang lebih ketat daripada alasan untuk menawarkan kartu kredit premium. Hal ini juga terkait dengan jumlah orang yang lebih besar dan proporsi nasabah rentan yang lebih tinggi dalam produk-produk keuangan penting. Langkah pra-pemrosesan ini memastikan bahwa keputusan untuk memasukkan fitur-fitur yang berkorelasi dengan karakteristik yang dilindungi dipertimbangkan secara cermat dalam konteks domain yang diatur dan potensi dampaknya terhadap konsumen.

Rencana Manajemen Risiko dan Desain Pengendalian

Rencana manajemen risiko pada dasarnya harus mampu memitigasi berbagai risiko yang telah diidentifikasi dalam proses penilaian, sehingga risiko residual yang tersisa tetap berada dalam batas selera risiko yang telah ditentukan. Upaya ini mencakup penerapan mekanisme pengendalian serta metodologi pengujian yang sesuai, yang dapat berbeda bergantung pada ranah sistem keuangan. Sebagai contoh, dalam kasus risiko diskriminasi yang dilarang oleh hukum, pengendalian yang tepat adalah dengan menguji algoritma guna memastikan tingkat kewajaran (*fairness*). Namun, persoalan muncul karena terdapat banyak definisi matematis mengenai kewajaran, sehingga seringkali memperumit penerapannya dan membuat para pemimpin bisnis kesulitan untuk memastikan implementasi yang konsisten. Oleh karena itu, dalam perumusan rencana manajemen risiko, diperlukan definisi kewajaran yang jelas, tepat, dan dapat ditindaklanjuti sesuai konteks penggunaannya.

Salah satu pendekatan yang dikenal sebagai *kewajaran melalui ketidaksadaran* mencoba menghindari diskriminasi dengan cara mengecualikan atribut-atribut yang dilindungi (seperti gender) dari model algoritma. Akan tetapi, karena algoritma pembelajaran mesin memiliki kemampuan kuat dalam menyimpulkan pola kompleks dari fitur lain, pendekatan ini tidak selalu menjamin hasil yang benar-benar adil. Hal ini dapat dilihat dari sebuah keputusan kontroversial di Uni Eropa yang melarang perusahaan asuransi mobil membedakan premi berdasarkan gender. Tujuan awalnya adalah mengurangi kesenjangan karena laki-laki rata-rata membayar premi lebih tinggi dibandingkan perempuan. Namun, kenyataannya kesenjangan justru semakin melebar, dari sebelumnya sekitar Rp.540.000 menjadi Rp.2.020.000. Hal ini terjadi karena perusahaan asuransi beralih menggunakan variabel proksi yang berkorelasi dengan gender, seperti jenis pekerjaan atau rata-rata lama pengalaman mengemudi, sebagai pengukur risiko.

Meskipun kebijakan tersebut dapat dianggap lebih adil apabila diyakini bahwa harga baru mencerminkan perbedaan risiko yang sebenarnya antara laki-laki dan perempuan, dari perspektif definisi keadilan yang lain, hal ini justru masih menimbulkan ketidakadilan. Model tersebut tetap berpotensi diskriminatif terhadap gender melalui penggunaan proksi. Studi pada tahun 2018 terhadap satu juta penawaran asuransi di Britania Raya menunjukkan bahwa harga median premi tertinggi dialami oleh buruh, seperti pekerja konstruksi, serta tukang cukur pekerjaan yang secara stereotip diasosiasikan dengan laki-laki sedangkan harga terendah justru dialami oleh asisten pribadi dan sekretaris, yang umumnya dikaitkan dengan perempuan.

Dengan mendefinisikan A sebagai atribut yang dilindungi, Y sebagai hasil aktual, dan $\hat{Y}$ sebagai hasil yang diprediksi, metrik keadilan lainnya dalam literatur statistik yang ada meliputi:

Paritas Demografis. Paritas demografis (keadilan kelompok) adalah metrik tingkat populasi di mana hasilnya tidak bergantung pada atribut yang dilindungi. Secara formal:

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$$

Sebagaimana yang dikemukakan Gajane dan Pechenizkiy, metrik ini layak digunakan jika tidak ada data "*ground truth*" yang andal, seperti dalam risiko kredit dan ketenagakerjaan di mana diskriminasi historis terhadap kelompok yang dilindungi terdokumentasi dengan baik. Di sisi lain, metrik ini tidak efektif jika ketidakseimbangan hasil dapat dibenarkan oleh atribut yang tidak dilindungi dan bukan proksi, karena hal ini dapat menyebabkan diskriminasi terbalik dan prediksi yang tidak akurat. Pemilihan hasil yang paling optimal juga tidak ditetapkan. Dalam kasus ini, trade-off antara akurasi dan paritas demografis mungkin terlalu signifikan untuk diterapkan dalam penggunaan penting bisnis, seperti penetapan harga.

Dalam ketenagakerjaan, di mana terdapat kepentingan tambahan untuk meningkatkan keragaman tenaga kerja, paritas demografis dapat menjadi metrik yang berguna untuk memastikan representasi yang adil dari semua kelas yang dilindungi. Keadilan Kontrafaktual³⁴ Model ini menyatakan bahwa dalam suatu model kausal (U, V, F) dengan sekumpulan variabel teramati (V), sekumpulan variabel latar belakang laten (U) yang bukan disebabkan oleh V , dan sekumpulan fungsi (F), kontrafaktual dari kepemilikan kelas yang dilindungi tidak bergantung pada hasilnya. Dengan X merepresentasikan atribut yang tersisa, A merepresentasikan atribut biner yang dilindungi, dan Y adalah hasil aktual, dan $\hat{Y}$ adalah hasil yang diprediksi, secara formal:

$$P(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y | X = x, A = a)$$

Meskipun metodologi inferensi kausal kuat, hubungan kausal seringkali sulit dihipotesiskan dalam domain FS yang kompleks. Sebuah laporan FCA mengakui tantangan dalam memisahkan perbedaan risiko aktuarial (penetapan harga berbasis biaya) dari perbedaan kesediaan membayar (diskriminasi harga). Metrik ini juga rentan terhadap bias retrospektif dan bias hasil.

Keadilan Individu. Keadilan individu menyatakan bahwa individu yang serupa mendapatkan hasil yang serupa. Secara formal, untuk individu yang serupa i dan j :

$$\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$$

Kriteria ini sangat bergantung pada pengukuran "kesamaan" antar individu yang tidak berkorelasi dengan karakteristik yang dilindungi. Kriteria ini juga lebih intensif komputasi daripada metrik tingkat populasi, yang dapat menjadi tantangan bagi solusi waktu nyata apa pun dengan Big Data.

Peluang yang Disetarakan. Peluang yang Disetarakan Peluang yang disetarakan menyiratkan bahwa hasil yang diprediksi berdasarkan hasil aktual tidak bergantung pada atribut terlindungi yang diprediksi berdasarkan hasil aktual. Hal ini menjamin bahwa hasil yang diprediksi memiliki rasio positif sejati yang sama di seluruh karakteristik yang dilindungi. Peluang yang disetarakan berfokus pada positif sejati: berdasarkan hasil positif, prediksi tidak bergantung pada atribut terlindungi. Dengan mendefinisikan A sebagai atribut terlindungi, Y sebagai hasil aktual, dan $\hat{Y}$ sebagai hasil yang diprediksi, formalisasi peluang yang disetarakan adalah:

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0,1\}$$

Demikian pula, kesempatan yang sama memenuhi kondisi berikut:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

Kepentingan relatif metrik akurasi dapat berbeda di setiap kasus penggunaan FS. Misalnya, perusahaan pemberi pinjaman hipotek mungkin paling khawatir tentang tingkat positif palsu algoritma (yaitu menyetujui pinjaman yang menyebabkan gagal bayar). Bank ritel dengan algoritma untuk memprediksi churn yang diharapkan mungkin berfokus pada tingkat negatif palsu (yaitu ditawarkan suku bunga yang lebih baik tetapi tetap pergi). Metrik peluang yang disamakan dan peluang yang disamakan gagal mengatasi diskriminasi yang mungkin sudah tertanam dalam data.

Untuk AI apa pun yang berisiko diskriminasi terhadap kelas yang dilindungi, definisi kewajaran dan justifikasi yang tepat harus diperlukan, dengan mempertimbangkan kekuatan dan kelemahan setiap opsi serta implikasi regulasi dari implementasinya dalam domain FS. Setelah dipilih, metrik tersebut dapat digunakan untuk menguji prediksi kewajaran sebagai bagian dari proses pengendalian. Perusahaan juga harus melibatkan pemangku kepentingan dari tim risiko hukum dan etika untuk memastikan definisi tersebut selaras dengan nilai-nilai etika dan selera risiko perusahaan.

Peran dan Tanggung Jawab yang Ditetapkan

Mengingat implikasi penggunaan AI yang berpotensi luas terhadap bisnis, perusahaan jasa keuangan (FS) mungkin perlu melibatkan lebih banyak pemangku kepentingan dari lini pertahanan pertama, kedua, dan ketiga di seluruh siklus hidup produk. Dalam SM&CS, manajemen senior harus siap untuk membuktikan kerangka kerja tata kelola dan risiko yang efektif untuk solusi AI. Praktik yang baik melibatkan partisipasi manajemen senior di seluruh proses pengujian dan pengembangan serta pemahaman tentang potensi konsekuensi perilaku pasar.

Fungsi risiko dan kepatuhan harus dilibatkan di setiap tahap proses pengembangan, pengujian, dan implementasi. Dalam publikasi tentang perdagangan algoritmik, OJK secara khusus mencatat bahwa staf kepatuhan harus bertujuan untuk memiliki pengetahuan dan keterampilan yang diperlukan untuk memberikan tantangan yang memadai bagi pengembangan algoritma, yang mungkin awalnya melibatkan pelaksanaan analisis kesenjangan atas kemampuan mereka untuk mengawasi aktivitas perdagangan algoritmik dan menetapkan peran serta tanggung jawab baru jika diperlukan.

Kolaborasi yang erat harus terjalin antara pemilik teknis model AI dan pemilik bisnis dari hasil model, dengan pengalihan tanggung jawab secara bertahap melalui siklus hidup, mulai dari desain hingga produksi.

17.5 MEMBANGUN

Proses pengembangan AI dapat menjadi tantangan bagi manajer risiko tradisional karena pendekatan agile yang sering diadopsi oleh tim ilmu data dan AI. Penting untuk menjembatani kesenjangan antara fungsi risiko tradisional dan tim teknis, karena tim teknis tidak selalu menyadari tantangan bisnis dan risiko. Misalnya, sebuah tim mungkin menggunakan perangkat lunak sumber terbuka tanpa meninjau apakah lisensinya memungkinkan penggunaan komersial.

Oleh karena itu, pengendalian risiko harus diintegrasikan ke dalam proses pengembangan. Misalnya, penggunaan perangkat lunak sumber terbuka harus memicu proses yang diperlukan untuk mendapatkan persetujuan dari tim hukum agar dapat melanjutkan setelah meninjau ketentuan lisensi sumber terbuka. Berikut ini adalah prinsip-prinsip regulasi yang serupa untuk teknologi lain yang berlaku sama untuk AI.

Proses Pengembangan dan Pengujian

Dengan mempertahankan kerangka kerja pengembangan dan pengujian yang kuat, konsisten, dan dipahami dengan baik, perusahaan perlu memastikan bahwa pengembangan algoritma mereka konsisten dengan selera risiko dan ekspektasi perilaku perusahaan. Persyaratannya serupa dengan yang diusulkan oleh OJK untuk perdagangan algoritmik. Sebelum penandatanganan, perusahaan perlu menyelesaikan proses peninjauan dan persetujuan yang komprehensif, dan semua pemangku kepentingan perlu memastikan bahwa tugas yang diberikan telah diselesaikan, diverifikasi, dan didokumentasikan.

Tata Kelola dan Pengawasan

Perusahaan harus bertujuan untuk memiliki komite tata kelola multidisiplin yang independen untuk meninjau dokumentasi dan penyelesaian prosedur pengujian serta memverifikasi bahwa algoritma tersebut konsisten dengan spesifikasi awal. Anggotanya harus dilatih untuk memahami risiko yang terkait dengan aplikasi AI. Masalah kewajaran, misalnya, membutuhkan pengetahuan domain dan pemahaman tentang trade-off matematis. Komite-komite ini harus menetapkan proses pengujian dan jaminan serta meninjau kinerja secara berkala untuk mengidentifikasi masalah yang muncul.

Manajemen Perubahan, Pengujian, dan Persetujuan yang Terdokumentasi

Sepanjang proses pengembangan dan pengujian, perusahaan harus memastikan bahwa mereka memiliki dokumentasi yang memadai dan jejak audit yang komprehensif untuk semua aplikasi AI yang diterapkan di seluruh organisasi mereka, termasuk pemilik terkait dan kepatuhan serta kontrol risiko utama yang telah diterapkan. Jika terjadi perubahan dalam definisi kewajaran, misalnya, hal ini termasuk dalam kategori perubahan material dan persetujuan oleh pemangku kepentingan terkait harus dicatat.

Transparansi dan Keterjelasan

Analisis faktor pendorong model harus mengungkapkan fitur apa pun yang seharusnya tidak memengaruhi model. Jika penyedia alamat email pilihan seseorang disorot sebagai faktor pendorong potensial untuk penetapan harga asuransi, hal ini harus dimasukkan ke dalam analisis kewajaran algoritma untuk memastikan fitur ini tidak digunakan sebagai proksi untuk

karakteristik yang dilindungi. Transparansi metodologis secara eksplisit dicantumkan oleh FCA dan GDPR sebagai persyaratan untuk pengambilan keputusan algoritmik.

Sebagaimana dinyatakan oleh Komisioner Informasi Inggris awal tahun ini, “[k]ita mungkin perlu, sebagai regulator, untuk melihat di balik layar atau di balik layar untuk melihat data apa yang digunakan, data pelatihan apa yang digunakan, faktor apa yang diprogram ke dalam sistem, dan pertanyaan apa yang dilatih untuk dijawab oleh sistem AI.” GDPR akan mengharuskan perubahan dalam hubungan dengan regulator, yang mewajibkan tim urusan regulasi yang didanai dengan tepat untuk membahas setiap pemrosesan data otomatis berisiko tinggi yang direncanakan.

Produksi

Tidak seperti otomatisasi proses robotik dan sistem berbasis aturan dan deterministik lainnya, risiko dalam solusi berbasis AI bersifat dinamis dan lebih sulit dideteksi. Hal ini membutuhkan perubahan pola pikir bagi manajer risiko, yang perlu tetap terlibat dalam proses pemantauan risiko. Sebelum produksi, solusi harus aman untuk diterapkan dalam skala besar dengan menanamkan kontrol otomatis.

Memastikan Solusi Aman untuk Diskalakan

Volume dan kecepatan pemrosesan data yang tinggi mungkin memerlukan kecepatan reaksi yang jauh lebih tinggi untuk setiap kesalahan karena peristiwa risiko dapat menyebar jauh lebih cepat. Harus ada kontrol yang memadai sebelum go-live, dengan aturan dan ambang batas yang diprogram untuk saat intervensi manusia diperlukan. FCA, dalam publikasinya tentang kepatuhan perdagangan algoritmik, mengamanatkan penjelasan yang jelas tentang kondisi yang perlu dipenuhi sebelum diimplementasikan ke dalam lingkungan langsung.

Meninjau Mekanisme Umpan Balik

Untuk algoritma pembelajaran mesin dengan data masuk langsung, harus ada kontrol untuk menandai masukan yang tidak sesuai. Chatbot, misalnya, tidak boleh belajar dari komentar pengguna yang provokatif atau tidak senonoh. Algoritma penetapan harga tidak boleh bereaksi secara tidak menentu terhadap guncangan eksternal.

Kesesuaian siklus umpan balik juga harus dipertimbangkan. Dalam algoritma risiko kredit, bank kemungkinan kekurangan data tentang individu yang ditolak pinjamannya, meskipun mereka kemudian mendapatkan pinjaman di tempat lain. Kontrafaktual dari keputusan tersebut, yaitu apakah mereka akan membayar kembali pinjaman seandainya disetujui, tidak diketahui. Ketidakpastian ini perlu dipertimbangkan saat mengevaluasi akurasi model. Dalam batasan keputusan model, eksperimen berkelanjutan untuk memberikan kredit kepada mereka yang berada tepat di luar titik batas dapat memberikan bukti bagi bisnis tentang apakah kebijakan tersebut tepat.

"Kill Switch" dan Kelangsungan Bisnis

Perusahaan harus mendokumentasikan prosedur dan kontrol untuk "kill switch" manual guna menghentikan algoritma beroperasi setelah kesalahan kritis atau perilaku abnormal terdeteksi. Rencana kelangsungan bisnis mungkin perlu didefinisikan ulang untuk

menyediakan rencana kontingensi untuk kembali ke proses manual dengan gangguan minimal pada proses bisnis penting.

Pemantauan

Karena sifat AI yang terus berkembang, pendekatan pemantauan yang lebih dinamis akan diperlukan untuk memastikan model masih berkinerja sebagaimana mestinya untuk kasus penggunaan spesifiknya. Indikator Kinerja Utama (KPI), termasuk persyaratan non-fungsional seperti kewajaran, perlu terus dipantau untuk memastikan kesesuaian, relevansi, dan akurasi. Selain itu, pengukuran risiko waktu nyata (KRI) dapat membantu menginformasikan lini pertahanan kedua dan ketiga. Contohnya adalah jumlah pengaduan dan banding terhadap keputusan kredit AI atas dasar persepsi ketidakadilan.

Pemantauan dan Pengujian Otomatis

Solusi berbasis AI dapat dimanfaatkan untuk pemantauan risiko AI. Misalnya, solusi berbasis pembelajaran mesin dapat memantau percakapan telepon antara agen asuransi dan nasabah untuk memprediksi kemungkinan penjualan yang salah. Dalam alat yang disebut TrueVoice ini, para ahli di bidang asuransi dan risiko perilaku telah mengembangkan dan melatih metrik khusus seperti kerentanan nasabah, dominasi, dan keengganan kerugian, yang semuanya menunjukkan kemungkinan penjualan yang salah yang lebih tinggi.

Nasabah Rentan

Langkah pasca-pemrosesan potensial lainnya mungkin diperlukan untuk memastikan kewajaran. Jika model menghasilkan variabilitas yang tinggi dalam hasil antar kelas yang dilindungi, terutama jika melibatkan nasabah rentan, organisasi dapat menerapkan pendekatan berbasis aturan untuk membatasi variasi tersebut. Jika seorang nasabah dinilai berisiko tinggi karena keadaan yang tidak biasa di sekitar kerentanannya, diperlukan fleksibilitas. FCA mendefinisikan pelanggan yang rentan sebagai “seseorang yang, karena keadaan pribadi mereka, sangat rentan terhadap kerugian, terutama ketika perusahaan tidak bertindak dengan tingkat kehati-hatian yang tepat.”

Secara khusus, FCA mencantumkan “kurangnya produk yang terjangkau dan sesuai untuk orang-orang dalam beberapa situasi non-standar” sebagai risiko perilaku potensial dan merekomendasikan bahwa “[f]leksibilitas dalam penerapan syarat dan ketentuan produk dan layanan memainkan peran penting [untuk] memastikan kebutuhan konsumen dalam keadaan rentan terpenuhi.” Organisasi FS dapat memasang pembatas untuk membatasi tingkat variabilitas jika pelanggan dianggap rentan.

Validasi Ulang Berkala

Peristiwa eksternal dan internal dapat mengakibatkan perubahan profil risiko organisasi. Perkembangan hukum dan peraturan baru mungkin memerlukan perubahan dalam desain model. Pengawasan media terhadap suatu kasus penggunaan dapat membuat solusi menjadi tidak layak. Tim hukum harus mengomunikasikan setiap perubahan dan implikasinya kepada pemilik bisnis.

Perencanaan Audit Internal

Fungsi Audit Internal (IA) harus mendapatkan pelatihan untuk memperoleh keahlian yang memadai guna memahami risiko yang terkait dengan setiap solusi AI dengan tepat.

Komponen AI harus dipertimbangkan secara eksplisit dalam proses perencanaan audit mereka, terlepas dari sistem yang lebih besar di mana mereka berada. Mereka harus memahami dan menangani pelanggaran kepatuhan serta menentukan frekuensi peninjauan yang diperlukan untuk setiap solusi AI.

Kesimpulan

Meskipun tingkat adopsi masih lambat, AI akan semakin menjadi komponen integral dari strategi perusahaan jasa keuangan untuk mencapai efisiensi operasional, meningkatkan layanan pelanggan, dan mendapatkan wawasan untuk keunggulan kompetitif. Sangat penting bagi organisasi untuk memahami implikasi adopsi ini dari perspektif risiko, sehingga tata kelola dan pengendalian yang tepat dapat diterapkan untuk memitigasi risiko baru dan yang semakin memburuk.

Dalam bab ini mengeksplorasi implikasi praktis manajemen risiko di seluruh siklus hidup produk solusi AI. Dengan fokus khusus pada Inggris Raya dan Uni Eropa, pendekatan yang disarankan dipadukan dengan prinsip dan preseden regulasi. Contoh kasus penggunaan utama yang disorot adalah risiko diskriminasi terhadap kelompok yang dilindungi. Meskipun telah terdapat beragam studi tentang definisi teknis dan teoretis tentang kewajaran, diperlukan penelitian lebih lanjut untuk merancang kerangka kerja guna menentukan definisi mana yang paling tepat dalam implementasi praktis metrik kewajaran di industri Jasa Keuangan.

Risiko AI tidak terbatas pada algoritma itu sendiri, melainkan memengaruhi seluruh organisasi. Pertimbangan khusus AI harus diintegrasikan ke dalam RMF yang ada untuk memastikannya tetap sesuai dengan tujuannya. Hanya dengan demikian perusahaan Jasa Keuangan akan merasa berdaya untuk menggunakan AI, karena memiliki keyakinan bahwa risiko terkait AI dapat diidentifikasi dan dikelola secara efektif.

BAB 18

ROBOT, PEKERJAAN, PAJAK, DAN TANGGUNG JAWAB

Perdebatan mengenai kecerdasan buatan (AI) dan robotika semakin sering muncul ke permukaan, tidak hanya dalam ranah teknologi, tetapi juga dalam ranah etika, hukum, dan sosial. Selama bertahun-tahun, imajinasi masyarakat dibentuk oleh gambaran fiksi ilmiah robot yang suatu hari akan bangkit, mendominasi, dan memperbudak manusia. Namun, narasi mutakhir tidak lagi berpusat pada ancaman robot ala *Terminator*, melainkan pada mesin-mesin yang bekerja tanpa lelah, yang dapat menggantikan manusia secara lebih efisien dan murah, sehingga membuat sebagian peran manusia terasa usang. Dengan kata lain, persoalannya bukan lagi “apakah robot akan melawan kita?”, melainkan “apakah robot akan membuat kita tidak relevan?”.

Perkembangan regulasi dan wacana publik menunjukkan bahwa dunia semakin serius memikirkan status robot. Pada 2017, Parlemen Eropa bahkan mengusulkan kemungkinan memberi robot status sebagai “manusia elektronik”, lengkap dengan kerangka etika dan hukum khusus. Ide ini, meski kontroversial, menunjukkan betapa cepatnya AI memasuki ruang-ruang kehidupan sehari-hari, mulai dari aplikasi lunak seperti algoritma hingga bentuk keras seperti perangkat medis, peralatan rumah tangga, dan asisten digital. Di bidang kesehatan, misalnya, AI kini sudah membantu dokter membaca radiografi, mengendalikan pompa insulin, memantau pasien, bahkan menyusun diagnosis awal tugas-tugas yang dulu dianggap tidak mungkin digantikan oleh teknologi.

Namun, kehadiran AI tidak hanya menciptakan efisiensi; ia juga menimbulkan disrupsi. Banyak pekerjaan manusia berisiko tergantikan, sementara bentuk pekerjaan baru membutuhkan keterampilan yang berbeda. Seorang pengemudi bus yang digantikan oleh bus otonom, misalnya, tidak serta-merta bisa beralih menjadi pengembang web, apalagi ketika profesi itu sendiri juga bisa diotomatisasi. Situasi ini menimbulkan kebutuhan untuk membicarakan kebijakan redistribusi ekonomi, pendidikan ulang tenaga kerja, hingga gagasan *basic income*. Tetapi, muncul persoalan baru: robot tidak membayar pajak. Jika perusahaan yang menggunakan robot juga tidak dikenakan pajak tambahan, siapa yang akan menutup biaya sosial akibat disrupsi ini?

Isu tanggung jawab hukum pun tidak kalah rumit. Jika sebuah robot menimbulkan kerusakan, siapa yang harus bertanggung jawab: pabrik, penjual, pemilik, atau robot itu sendiri? Ada usulan untuk memberi robot “kepribadian hukum” layaknya entitas elektronik yang bertanggung jawab. Namun, pendekatan ini dikritik sebagai solusi keliru. Sejarah hukum menunjukkan bahwa tanggung jawab selalu diletakkan pada pihak yang mengendalikan, bukan pada alat itu sendiri. Hukum Romawi bahkan sejak lama mengenal prinsip *respondeat superior*, di mana pemilik bertanggung jawab atas tindakan yang dilakukan oleh budak atau pekerjanya. Memberikan status “orang elektronik” kepada robot justru berisiko mengalihkan tanggung jawab dari aktor manusia atau korporasi.

Dengan demikian, persoalan besar yang kita hadapi bukanlah apakah robot bisa menjadi manusia, melainkan bagaimana kita menyusun ulang kerangka konseptual dan normatif dalam menghadapi bentuk agensi baru yang muncul di infosfer. Dunia digital memaksa kita untuk berhenti terjebak pada pertanyaan spekulatif yang bersifat fiksi ilmiah, dan sebaliknya menekankan kebutuhan untuk merancang solusi etis, hukum, dan sosial yang nyata. Pertanyaan kuncinya adalah: jenis masyarakat seperti apa yang ingin kita ciptakan bersama teknologi ini? Bagaimana kita memastikan bahwa kehadiran AI meningkatkan kualitas hidup, memperkuat solidaritas sosial, dan tidak memperdalam ketidaksetaraan?

Singkatnya, kita membutuhkan lebih sedikit fiksi ilmiah dan lebih banyak filsafat. Masa depan AI bukan hanya soal mesin, melainkan soal manusia yang harus hidup berdampingan dengan mesin tersebut dalam infosfer yang kita desain bersama. Tantangan terbesar ke depan bukanlah inovasi digital itu sendiri, melainkan tata kelola digital yang adil, inklusif, dan berorientasi pada kesejahteraan bersama.

18.1 BAGAIMANA MASA DEPAN KECERDASAN BUATAN (AI)

Perkembangan Kecerdasan Buatan (AI) dalam beberapa tahun terakhir menunjukkan dinamika yang sangat pesat dan penuh peluang. AI tidak lagi hanya menjadi topik penelitian akademik, tetapi juga telah merambah berbagai aspek kehidupan manusia, mulai dari layanan kesehatan, pendidikan, bisnis, hingga hiburan. Hal ini menandakan bahwa AI semakin berperan penting dalam membentuk arah inovasi digital di masa depan.

Bab ini membahas kemungkinan arah perkembangan AI dalam waktu dekat dengan menyoroti dua tren utama yang patut diperhatikan. Pertama, adanya pergeseran dari penggunaan data historis menuju pemanfaatan data sintetis. Data sintetis menjadi alternatif yang semakin relevan karena mampu mengatasi keterbatasan data nyata, baik dari sisi jumlah, kualitas, maupun isu etika dan privasi. Kedua, perubahan fokus dari tugas-tugas yang sulit dalam hal kemampuan manusia menuju pada tugas-tugas yang kompleks dalam perspektif komputasi. Pergeseran ini menunjukkan bahwa tantangan AI ke depan bukan lagi sekadar meniru kecerdasan manusia, melainkan mengolah kompleksitas dalam skala besar dengan dukungan kekuatan komputasi yang semakin maju.

Kedua tren tersebut diperkirakan akan menjadi strategi utama dalam pengembangan AI, sejauh kondisi teknologi memungkinkan dan kebutuhan manusia menuntut. Dengan memahami arah ini, kita dapat melihat bagaimana inovasi AI akan terus bergerak, sekaligus menyiapkan diri terhadap peluang dan tantangan yang akan muncul.

Kecerdasan buatan (AI) dalam beberapa waktu terakhir telah menjadi sorotan utama, dengan berbagai janji, peluang, tantangan, hingga risiko yang menyertainya. Pertanyaan besar yang muncul adalah: bagaimana masa depan AI dapat diperkirakan? Cara paling aman untuk membuat prediksi sebenarnya adalah dengan melihat ke belakang, sebab sejarah sering kali memberi petunjuk mengenai arah perkembangan. Namun, ketika ramalan bersifat spekulatif atau manipulatif, biasanya orang yang bijak akan memilih membuat prediksi yang sifatnya netral, tidak kontroversial, atau bahkan sulit untuk dibuktikan benar maupun salah.

Pada sisi yang relatif aman, dapat diperkirakan bahwa dalam waktu dekat akan ada tekanan regulasi yang semakin kuat dari para pembuat kebijakan agar penggunaan AI selaras dengan nilai dan norma sosial. Uni Eropa, misalnya, hampir pasti akan mengeluarkan regulasi terkait hal ini, cepat atau lambat. Sementara itu, pada sisi yang lebih spekulatif, tidak sedikit pihak yang akan terus menjual narasi bernuansa bencana. Mereka menawarkan gambaran masa depan distopia, tetapi sengaja menempatkannya jauh ke depan sehingga sulit diverifikasi, ibarat ramalan kiamat yang tidak akan ada yang bisa membuktikan salah ketika waktunya tiba. Narasi ketakutan seperti ini laku keras, layaknya film horor bertema vampir atau zombi.

Namun, tentu saja perkembangan AI tidak sesederhana itu. Tantangan terbesarnya justru terletak pada usaha untuk menebak "benih-benih waktu," yakni tanda-tanda awal mengenai potensi mana yang akan tumbuh menjadi kenyataan dan mana yang tidak. Hal ini serupa dengan ungkapan dalam *Macbeth* (Babak I, Adegan III), tentang upaya membaca masa depan dengan penuh ketidakpastian. Inilah yang coba dilakukan penulis dalam uraian berikutnya, dengan berhati-hati memilih jalur prediksi yang logis, tetapi juga tidak terlalu defensif hingga enggan mengambil risiko salah.

Salah satu kesulitan mendasar dalam memperkirakan arah perkembangan AI adalah menentukan tingkat abstraksi yang tepat. Artinya, kita harus mampu mengidentifikasi serangkaian faktor relevan yang benar-benar memengaruhi jalannya perkembangan teknologi ini. Dalam konteks AI, observasi terbaik menurut penulis dapat diperoleh dari dua hal utama: pertama, sifat data yang digunakan AI untuk mencapai performanya; dan kedua, jenis masalah yang mungkin dapat dipecahkan AI secara efektif. Oleh karena itu, pembahasan selanjutnya dibagi menjadi dua bagian. Bagian 18.2 akan menguraikan mengenai sifat data yang menjadi kebutuhan utama AI, sedangkan Bagian 18.3 akan membahas ruang lingkup permasalahan yang paling mungkin ditangani AI secara sukses.

18.2 AI: DEFINISI KERJA

Kecerdasan buatan (AI) memiliki banyak definisi yang berkembang seiring waktu. Saat ini, AI mencakup beragam cabang ilmu pengetahuan dan teknologi, sebagaimana dirangkum dengan baik oleh Corea (2018) dalam Gambar 18.1. Meski demikian, secara garis besar paradigma AI masih sejalan dengan definisi klasik yang diperkenalkan oleh para perintisnya, yakni John McCarthy, Marvin Minsky, Nathaniel Rochester, dan Claude Shannon, melalui dokumen bersejarah *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* pada tahun 1955. Dalam naskah pendiri tersebut, AI dipahami sebagai upaya membuat mesin berperilaku dengan cara yang dianggap cerdas jika perilaku itu dilakukan oleh manusia.

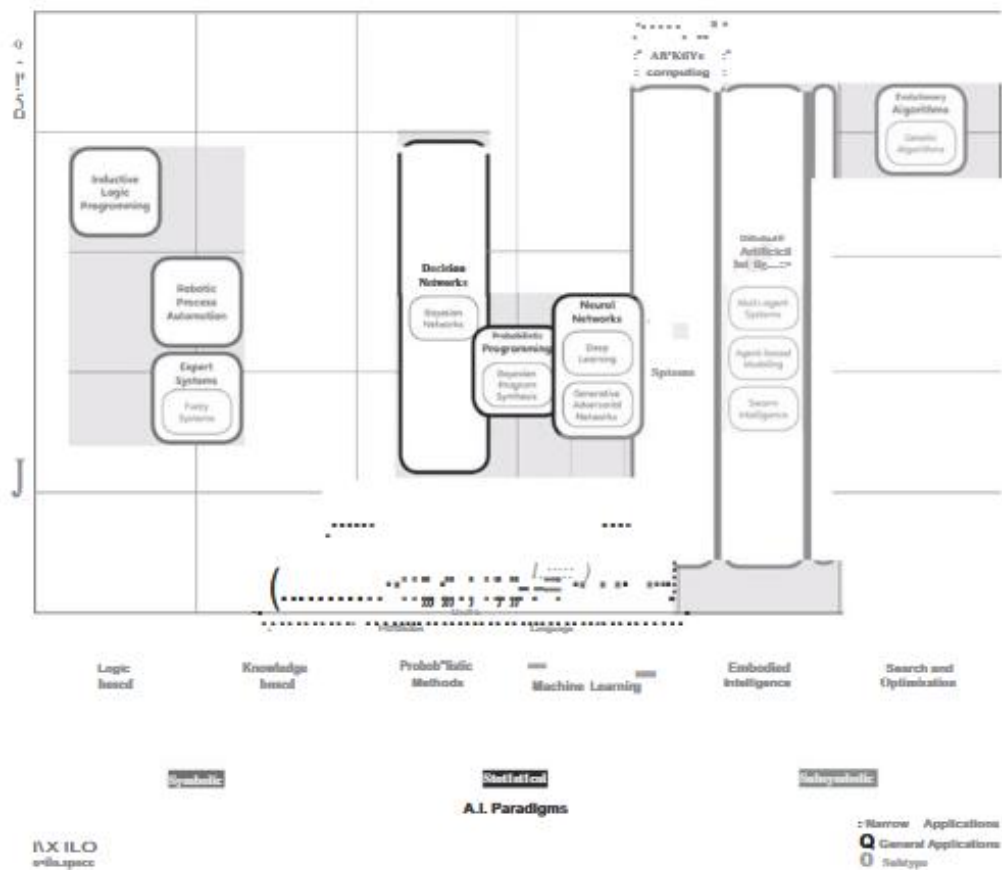
Seperti telah ditegaskan sebelumnya, definisi ini pada dasarnya bersifat kontrafaktual. Artinya, mesin hanya disebut "cerdas" sejauh perilakunya menyerupai perilaku manusia yang dianggap cerdas. Ini tidak berarti bahwa mesin benar-benar memiliki kecerdasan atau kemampuan berpikir. Menganggap mesin benar-benar berpikir hanyalah sebuah kekeliruan, bahkan mendekati sikap takhayul. Sebagai perbandingan sederhana, mesin pencuci piring

mampu membersihkan piring dengan lebih cepat dan konsisten dibandingkan manusia, tetapi itu tidak berarti mesin tersebut menggunakan "kecerdasan" untuk melakukannya, atau melakukannya dengan cara yang sama seperti manusia.

Pemahaman kontrafaktual inilah yang juga menjadi dasar dari *uji Turing*. Uji tersebut tidak menilai apakah mesin benar-benar berpikir, melainkan apakah mesin dapat melakukan suatu tugas sehingga hasilnya tidak dapat dibedakan dari hasil yang dicapai manusia. Dengan demikian, fokusnya tetap pada kinerja yang dihasilkan, bukan pada keberadaan kecerdasan sejati dalam mesin.

Definisi klasik ini membuka jalan untuk memahami AI sebagai sebuah sumber daya yang terus berkembang: interaktif, otonom, dan sering kali mampu belajar sendiri melalui mekanisme *machine learning* (lihat Gambar 18.1). AI kini dapat menjalankan tugas-tugas yang sebelumnya hanya mungkin dilakukan dengan kecerdasan dan intervensi manusia. Namun, perkembangan ini sekaligus menghadirkan tantangan etis. Hal ini karena agen buatan sering kali cukup terinformasi, tampak "cerdas", serta mampu bertindak secara otonom, bahkan dalam konteks tindakan yang memiliki relevansi moral. Walaupun aspek etis ini sangat penting, pembahasan lebih mendalam akan ditunda dan hanya disinggung secara singkat pada bagian kesimpulan.

Dengan kata lain, AI didefinisikan terutama berdasarkan hasil dan tindakannya. Oleh karena itu, dalam pembahasan selanjutnya, AI akan diperlakukan sebagai sebuah bentuk agensi cerdas yang siap digunakan (*ready-made intelligent agency*). Pertanyaan utama yang kemudian muncul adalah: bagaimana teknologi ini akan berkembang, dan dalam konteks apa saja AI dapat digunakan secara efektif? Untuk menjawabnya, diskusi akan dimulai dari salah satu fondasi terpenting AI, yakni data yang menjadi bahan bakarnya.



Gambar 18.1 Peta pengetahuan AI (AIKM).

18.3 MASA DEPAN AI DAN KEBUTUHAN AKAN LUDUFIKASI

Sering kali kita mendengar ungkapan bahwa *data adalah minyak baru*. Perbandingan ini memang menarik, tetapi sejatinya kurang tepat. Minyak bersifat terbatas, hanya dapat digunakan sekali, dan memiliki harga yang relatif pasti di pasar. Sebaliknya, data memiliki karakteristik yang sangat berbeda: ia tahan lama, bisa digunakan berulang kali, mudah dipindahkan, mudah digandakan tanpa kehilangan kualitas, dan bahkan bisa dibagikan secara bersamaan tanpa batas.

Selain itu, nilai ekonomi data tidak ditentukan secara mutlak seperti minyak, melainkan sangat bergantung pada siapa yang menggunakannya dan untuk tujuan apa. Lebih jauh lagi, jika menyangkut data pribadi, muncul persoalan hukum dan etika yang kompleks, termasuk perdebatan mengenai kepemilikan. Klaim “data saya” lebih mirip dengan mengatakan “tangan saya” (sesuatu yang melekat pada identitas dan eksistensi seseorang), ketimbang “minyak saya” (sekadar sumber daya yang bisa diperdagangkan). Dengan demikian, analogi antara data dan minyak sebenarnya bersifat berlebihan.

Namun, bukan berarti analogi tersebut sama sekali tidak bermanfaat. Sebab, data memang memiliki kesamaan dengan minyak dalam hal statusnya sebagai sumber daya bernilai

tinggi yang memerlukan proses pemurnian agar nilai sebenarnya dapat diekstraksi. Tanpa data, algoritma, termasuk kecerdasan buatan (AI), tidak akan dapat menghasilkan apa pun ibarat mesin tanpa bahan bakar. AI membutuhkan data pada dua tahap utama: pertama, untuk melatih modelnya (*training*), dan kedua, untuk mengaplikasikan hasil pelatihan tersebut pada situasi nyata. Dengan kata lain, fleksibilitas AI sangat ditentukan oleh ketersediaan dan kualitas data yang digunakan. Data inilah yang menjadi faktor utama penentu cakupan aplikasi serta tingkat keberhasilan AI.

Sebagai contoh, pada tahun 2016 Google memanfaatkan sistem *machine learning* milik DeepMind untuk mengurangi konsumsi energi di pusat datanya. Sistem tersebut terbukti mampu memahami dinamika kompleks dalam manajemen energi dan kemudian mengoptimalkan penggunaan daya. Lebih jauh lagi, Google menyatakan bahwa kerangka kerja ini berpotensi diterapkan secara lebih luas pada berbagai tantangan lainnya, baik dalam lingkungan pusat data maupun di luar itu. Beberapa kemungkinan penerapannya antara lain peningkatan efisiensi konversi pada pembangkit listrik (sehingga menghasilkan energi lebih besar dari input yang sama), pengurangan penggunaan energi dan air dalam manufaktur semikonduktor, serta peningkatan *throughput* pada fasilitas manufaktur.

Sudah menjadi pengetahuan umum bahwa kecerdasan buatan (AI) belajar dari data yang diberikan kepadanya dan kemudian secara bertahap meningkatkan kualitas hasil yang dihasilkan. Misalnya, jika sebuah jaringan saraf dilatih dengan ribuan hingga jutaan foto anjing, seiring waktu AI akan mampu mengenali anjing dengan semakin baik, termasuk jenis atau gambar anjing yang belum pernah dilihat sebelumnya. Proses pembelajaran ini pada umumnya membutuhkan volume data yang besar, dan sering kali semakin banyak data yang tersedia, semakin baik pula hasil yang diperoleh.

Sebagai ilustrasi, sebuah penelitian terbaru yang dilakukan oleh tim peneliti dari University of California, San Diego, melatih sistem AI menggunakan 101,6 juta titik data rekam medis elektronik (RME). Data tersebut mencakup catatan yang ditulis oleh dokter hingga hasil uji laboratorium dari 1.362.559 kunjungan pasien anak di sebuah pusat medis besar di Guangzhou, Tiongkok. Hasil penelitian menunjukkan bahwa setelah melalui proses pelatihan, sistem AI tersebut mampu mencapai tingkat akurasi diagnostik yang tinggi di berbagai sistem organ, bahkan sebanding dengan kemampuan dokter anak berpengalaman dalam mendiagnosis penyakit umum pada anak-anak. Studi ini memberikan bukti nyata bahwa sistem berbasis AI dapat digunakan untuk membantu dokter menangani volume data yang sangat besar, melengkapi evaluasi diagnostik, serta mendukung pengambilan keputusan klinis dalam kondisi ketidakpastian atau kompleksitas medis. Temuan semacam ini berpotensi memberikan dampak signifikan, terutama di wilayah dengan keterbatasan tenaga medis, meskipun manfaatnya juga bersifat universal.

Akan tetapi, perkembangan AI belakangan ini menunjukkan bahwa fokus tidak selalu lagi pada kuantitas data yang masif, yang kerap disebut dengan istilah *Big Data*. Dalam beberapa kasus, kualitas data ternyata jauh lebih menentukan dibandingkan sekadar jumlah. Hal ini terbukti, misalnya, melalui penelitian yang dilakukan oleh DeepMind pada tahun 2018 bekerja sama dengan Rumah Sakit Mata Moorfields di London, Inggris. Tim tersebut melatih

sistem AI untuk mendeteksi penyakit mata yang berpotensi mengancam penglihatan menggunakan data tomografi koherensi optik (OCT), yaitu teknik pencitraan yang menghasilkan representasi tiga dimensi bagian belakang mata.

Menariknya, meskipun jumlah data pelatihan relatif kecil, yakni hanya 14.884 pemindaian, sistem AI yang dikembangkan justru mampu memberikan rekomendasi rujukan yang akurat, bahkan dalam banyak kasus melampaui rekomendasi para ahli. Temuan ini menggarisbawahi pentingnya kualitas dan relevansi data. Data yang terkurasi dengan baik, dapat diandalkan, serta terus diperbarui terbukti mampu menghasilkan performa AI yang tinggi, bahkan tanpa harus bergantung pada volume data yang sangat besar.

Fakta ini membawa kita pada wacana baru mengenai arah perkembangan AI di masa depan. Jika kualitas data menjadi kunci, maka pertanyaan berikutnya adalah mengenai asal-usul data tersebut. Dalam banyak kasus, data yang digunakan berasal dari institusi atau organisasi tertentu, misalnya rumah sakit. Data semacam ini sering disebut sebagai data *historis*, *autentik*, atau *real-world data* karena benar-benar dihasilkan dari praktik nyata.

Namun, selain data historis, AI juga dapat menghasilkan datanya sendiri. Yang dimaksud di sini bukanlah metadata atau data sekunder tentang proses penggunaannya, melainkan data utama yang benar-benar dibangun oleh AI. Data semacam ini dikenal dengan istilah *data sintetis*. Sayangnya, istilah tersebut memiliki riwayat penggunaan yang agak membingungkan. Pada 1990-an, data sintetis biasanya merujuk pada data historis yang telah dianonimkan untuk melindungi privasi, misalnya melalui proses *masking*. Dalam pengertian kontemporer, data sintetis justru dipahami sebagai data yang sepenuhnya dihasilkan secara artifisial oleh sistem AI, bukan berasal dari catatan historis yang diubah.

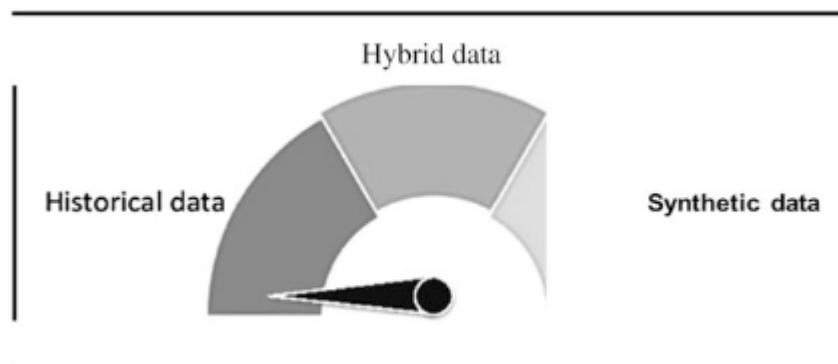
Dalam perkembangan kecerdasan buatan, sumber data yang digunakan dalam proses pelatihan sistem AI menjadi isu yang sangat penting. Secara garis besar, terdapat tiga jenis sumber data yang dapat dibedakan, yakni data historis, data sintetis murni, dan data hibrida. Data historis adalah data yang berasal dari pengalaman nyata manusia maupun sistem dunia nyata, seperti catatan medis pasien, hasil laboratorium, atau basis data permainan catur para Grandmaster. Kelebihan utama data historis terletak pada keotentikannya, sebab data tersebut benar-benar merepresentasikan kenyataan yang telah terjadi. Akan tetapi, penggunaan data historis sering menghadapi sejumlah tantangan, misalnya terkait keterbatasan jumlah data, biaya pengumpulan, hingga masalah privasi yang rawan muncul ketika data melibatkan individu nyata. Bahkan ketika data historis telah dianonimkan atau resolusinya diturunkan demi melindungi kerahasiaan, pada dasarnya asal-usulnya tetap bersumber dari realitas yang telah ada.

Berbeda dengan itu, muncul kategori kedua yang disebut data sintetis murni. Data ini sepenuhnya dihasilkan oleh AI tanpa bergantung sedikit pun pada data historis. Sebuah contoh monumental adalah AlphaZero, sistem AI yang dikembangkan oleh DeepMind. Tidak seperti Deep Blue komputer catur IBM yang mengalahkan Garry Kasparov dengan mengandalkan basis data ribuan permainan manusia AlphaZero hanya diberikan aturan permainan catur, lalu berlatih dengan cara bermain melawan dirinya sendiri. Dengan demikian, seluruh data yang digunakannya adalah hasil ciptaannya sendiri. Menariknya, justru dari data sintetis semacam

ini, AlphaZero mampu mencapai tingkat permainan yang melampaui semua program catur maupun pemain manusia. Dari sudut pandang epistemologi, data sintetis memberikan posisi istimewa bagi AI sebagai “pembuat pengetahuan”, sebab ia memahami sesuatu melalui proses yang diciptakannya sendiri. Keunggulan lain data sintetis adalah bersih, bebas dari masalah privasi, dapat direplikasi tanpa batas, mudah diproduksi ulang, serta telah terformat dengan sempurna sesuai kebutuhan sistem yang menciptakannya. Karena alasan inilah, data sintetis sering digunakan dalam konteks simulasi, uji keamanan digital, dan eksperimen di mana kecepatan serta biaya produksi data menjadi pertimbangan utama.

Di antara kedua kutub tersebut terdapat data hibrida, yaitu data yang lahir dari kombinasi antara historis dan sintetis. Data hibrida memanfaatkan data historis sebagai fondasi, lalu melalui proses generatif menghasilkan data baru yang tidak sepenuhnya identik dengan data asli. Salah satu teknik paling menonjol dalam menghasilkan data hibrida adalah *Generative Adversarial Networks* (GANs). Dalam metode ini, terdapat dua jaringan saraf yang saling berhadapan: Generator yang bertugas menciptakan data tiruan, dan Diskriminator yang berusaha membedakan mana data asli dan mana yang palsu. Melalui interaksi berulang, Generator semakin terampil menghasilkan data yang menyerupai data historis, meskipun ia sendiri tidak pernah melihat data nyata secara langsung. Proses kompetitif ini membuat data yang dihasilkan GAN menjadi representasi sintetis yang tetap memiliki kedekatan dengan realitas.

Perbedaan mendasar dari ketiga jenis data ini terletak pada asal-usulnya. Data historis berakar pada pengalaman nyata, data sintetis sepenuhnya lahir dari proses buatan AI, sedangkan data hibrida berada di tengah-tengah sebagai hasil perkawinan antara realitas dan simulasi. Pemahaman atas perbedaan ontologis ini sangat penting, sebab implikasinya menyentuh ranah epistemologi, etika, serta kebijakan penggunaan data. Dengan demikian, pergeseran dari ketergantungan pada data historis menuju pemanfaatan data sintetis dan hibrida menandai arah baru dalam perkembangan AI modern, di mana kualitas, sumber, dan sifat data menjadi faktor penentu keberhasilan sistem yang dibangun.



Gambar 18.2 Pergeseran dari data historis sepenuhnya ke data sintetis yang sesungguhnya

Dalam proses pembelajaran mesin, Generator pada sistem berbasis GAN (*Generative Adversarial Networks*) mampu menciptakan data sintetis yang menyerupai data masukan yang

sudah dikenal. Di sinilah sifat hibrida mulai terlihat, sebab meskipun data yang dihasilkan Generator bersifat baru, proses pelatihannya tetap bergantung pada data historis yang digunakan Diskriminator. Dengan kata lain, ini bukanlah kasus “partenogenesis” murni seperti AlphaZero, yang menghasilkan seluruh datanya sendiri, melainkan bentuk penciptaan yang berada di wilayah tengah. Meski demikian, karakteristik data yang dihasilkan tetap menarik, karena misalnya wajah sintetis buatan Generator tidak menimbulkan persoalan privasi, persetujuan, ataupun kerahasiaan ketika berada pada tahap pengembangan.

Berbagai metode untuk menghasilkan data hibrida maupun data sintetis kini terus berkembang, bahkan banyak yang dirancang untuk sektor-sektor tertentu. Seiring dengan itu, muncul pula tren altruistik untuk membuka akses publik secara gratis terhadap kumpulan data semacam ini. Fenomena tersebut menunjukkan bahwa masa depan kecerdasan buatan tidak hanya bergantung pada “data kecil” yang berkualitas tinggi, tetapi juga pada meningkatnya kemampuan AI untuk menghasilkan datanya sendiri. Jika arah ini terus berkembang, kita dapat membayangkan terjadinya lompatan besar dalam kemampuan AI, karena sistem tidak lagi sepenuhnya menunggu ketersediaan data historis, melainkan mampu menjadi sumber data itu sendiri.

Perbedaan antara data historis, data sintetis, dan data hibrida pada dasarnya ditentukan oleh jenis aturan yang digunakan dalam proses penciptaannya. Data historis diperoleh melalui pencatatan peristiwa nyata, yang berarti ia merekam perilaku suatu sistem sebagaimana adanya. Sementara itu, data sintetis sering kali lahir melalui pengabstraksian aturan, di mana sebagian aspek data historis ditutupi, dikaburkan, atau dikurangi resolusinya, misalnya melalui anonimisasi. Adapun data hibrida biasanya dibentuk melalui kombinasi antara keduanya, dengan memanfaatkan aturan pembatas ketika aturan konstitutif yang lebih mendasar tidak tersedia.

Konsep aturan ini dapat dipahami lebih jelas dengan membandingkan permainan catur dan sepak bola. Dalam catur, aturan bersifat konstitutif aturan tersebut menciptakan seluruh kemungkinan langkah yang dapat dimainkan dan sekaligus membatasi langkah yang ilegal. Karena itu, sistem AI seperti AlphaZero mampu belajar dengan hanya berpegang pada aturan permainan, lalu menghasilkan jutaan data sintetis melalui latihan mandiri. Fakta bahwa AlphaZero memainkan 44 juta pertandingan dalam sembilan jam menunjukkan skala kemampuan AI yang tidak dapat ditandingi oleh manusia. Sebaliknya, sepak bola adalah permainan dengan aturan pembatas. Menendang bola sudah merupakan aktivitas nyata, lalu aturan hanya menentukan langkah mana yang dianggap legal. Artinya, aturan tidak sepenuhnya membentuk permainan dari awal, melainkan hanya mengarahkan dan mengatur aktivitas yang sudah ada.

Implikasi perbedaan ini sangat penting. Dalam permainan papan seperti catur, shogi, atau Go, aturan konstitutif memungkinkan AI untuk sepenuhnya menciptakan data sintetisnya sendiri. Sebaliknya, dalam permainan yang diatur dengan aturan pembatas, seperti sepak bola, AI hanya dapat mengandalkan data historis berupa rekaman pertandingan, strategi, atau catatan pemain untuk melakukan pembelajaran. Maka, semakin suatu interaksi dapat diformalkan menjadi permainan dengan aturan konstitutif, semakin besar pula peluang AI

untuk menghasilkan data sintetis yang murni. Hal ini membuka jalan bagi konsep “enveloping” (Floridi, 2014), yakni proses di mana fenomena tertentu dapat direkayasa menjadi bentuk permainan yang memungkinkan AI melakukan eksplorasi dan pembelajaran mandiri.

Dengan cara ini, data terbaik bagi AI dapat dipahami sebagai data historis yang benar-benar representatif atau data sintetis yang dihasilkan dari aturan yang sama dengan realitas yang mendasarinya. Dalam permainan papan, kondisi ini tercapai secara otomatis. Namun, dalam dunia nyata, permasalahan menjadi lebih kompleks. Misalnya, dalam meneliti perkembangan tumor pada pasien dengan kondisi tertentu, aturan genetis tidak tersedia secara eksplisit, melainkan harus diekstraksi terlebih dahulu melalui penelitian ilmiah dan kini penelitian semacam itu semakin banyak melibatkan AI.

18.4 MASA DEPAN AI DARI MASALAH SULIT KEMASALAH KOMPLEKS

Kecerdasan buatan (AI) pada dasarnya dapat dipahami sebagai sebuah reservoir agensi, yakni wadah yang mampu digunakan untuk memecahkan berbagai masalah. Cara kerja AI dalam mencapai tujuan ini adalah dengan memisahkan antara kemampuan menyelesaikan suatu tugas secara berhasil dengan kebutuhan untuk memiliki kecerdasan layaknya manusia dalam melakukannya. Sebagai contoh, aplikasi di ponsel saya dapat bermain catur jauh lebih baik dibanding saya, meskipun aplikasi tersebut sama sekali tidak “cerdas” dalam arti manusiawi. Pemisahan inilah yang membuat solusi berbasis AI menjadi mungkin. Oleh karena itu, memahami masa depan AI berarti juga memahami sifat-sifat masalah di mana pemisahan semacam ini secara teknis layak diterapkan dan secara ekonomis menguntungkan untuk dipraktikkan.

Banyak permasalahan yang kita coba pecahkan dengan AI berada di dunia fisik. Contohnya beragam: mulai dari mengemudi mobil, memindai label harga di supermarket, membersihkan jendela atau aliran air, hingga memotong rumput di halaman. Sebagian pembaca mungkin langsung membayangkan AI hanya sebagai robotika, tetapi sebenarnya ruang lingkupnya lebih luas. Aplikasi dan antarmuka cerdas yang hadir dalam *Internet of Things* (IoT) juga termasuk dalam kategori analisis ini. Maka dari itu, untuk memahami perkembangan AI dalam kaitannya dengan dunia fisik, diperlukan sebuah pemetaan masalah berdasarkan sumber daya apa yang dibutuhkan untuk menyelesaikannya, serta seberapa jauh AI mampu memenuhi kebutuhan tersebut.

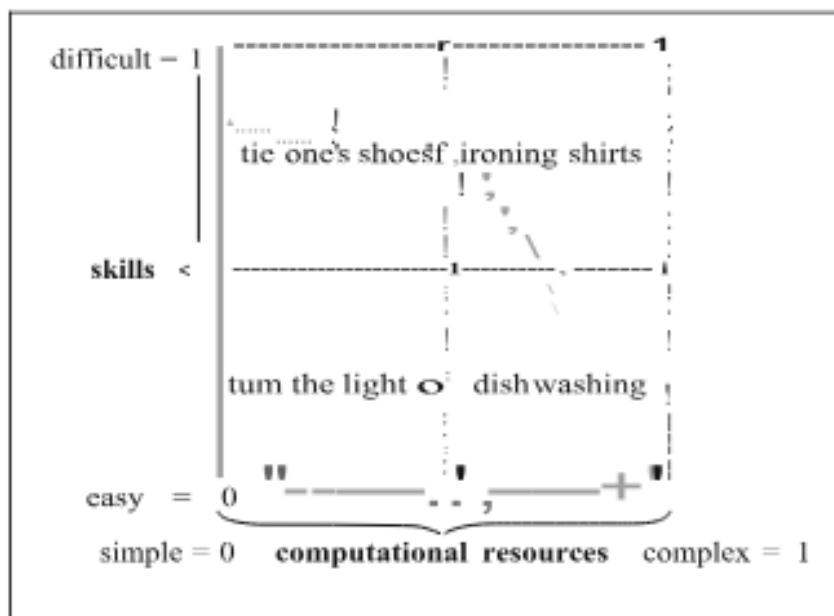
Ada dua jenis sumber daya utama yang dapat digunakan untuk memetakan masalah tersebut: sumber daya komputasi, yang berkaitan dengan tingkat kompleksitas masalah, dan sumber daya keterampilan, yang berkaitan dengan tingkat kesulitan tugas. Kompleksitas suatu masalah telah lama diteliti dalam teori komputasi. Dimensi ini bersifat sangat kuantitatif karena adanya standar pengukuran yang jelas. Sebagai analogi, jika kita memiliki “meter”, maka kita dapat mengukur panjang. Demikian pula, dengan menggunakan Mesin Turing sebagai titik acuan, kita dapat menghitung berapa banyak langkah (waktu) dan seberapa besar ruang memori (pita) yang dibutuhkan untuk menyelesaikan suatu masalah komputasi.

Untuk mempermudah, mari kita sepakati bahwa tingkat kompleksitas suatu masalah dapat dipetakan pada skala antara 0 hingga 1, di mana 0 berarti masalah sederhana dan 1

berarti sangat kompleks. Skala ini memungkinkan kita untuk menggambarkan dengan jelas posisi suatu masalah dalam spektrum kesulitan komputasi, tanpa harus selalu masuk ke detail teknis yang rumit. Selain kompleksitas, ada pula dimensi tingkat kesulitan yang berhubungan dengan keterampilan yang dibutuhkan untuk menyelesaikan suatu tugas. Literatur yang relevan, misalnya dalam kajian perkembangan motorik manusia, umumnya tidak menyusun klasifikasi masalah berdasarkan kebutuhan sumber daya, tetapi lebih pada kinerja agen manusia, yakni kemampuan atau keterampilan yang ditunjukkan dalam menyelesaikan suatu tugas. Dimensi ini lebih kualitatif dibandingkan dengan kompleksitas komputasi.

Salah satu perbedaan standar dalam menilai tingkat kesulitan adalah antara keterampilan motorik kasar dan keterampilan motorik halus. Keterampilan motorik kasar melibatkan penggunaan kelompok otot besar, seperti saat berjalan, melompat, menangkap, atau menendang bola. Sebaliknya, keterampilan motorik halus melibatkan penggunaan otot-otot kecil pada pergelangan tangan, jari tangan, atau jari kaki, misalnya ketika menulis, mengetik, mencuci piring, menggunakan alat, atau memainkan instrumen musik. Dari sini terlihat jelas bahwa ada spektrum tingkat kesulitan yang berbeda-beda.

Untuk keperluan pemetaan, mari kita juga sepakat bahwa tingkat kesulitan suatu masalah berdasarkan keterampilan yang dibutuhkan dapat dinilai pada skala 0 hingga 1, di mana 0 berarti mudah dan 1 berarti sulit. Dengan dua dimensi ini kompleksitas komputasi dan kesulitan keterampilan kita dapat memiliki kerangka kerja yang lebih sistematis untuk menilai tantangan yang dihadapi AI dalam berinteraksi dengan dunia fisik.



Gambar 18.3 Menerjemahkan tugas yang sulit menjadi tugas yang kompleks

Setelah memahami kerangka pemetaan antara kompleksitas dan kesulitan, kita kini dapat menempatkan beberapa contoh nyata ke dalam diagram konseptual tersebut. Misalnya, menyalakan lampu merupakan sebuah masalah dengan tingkat kerumitan yang sangat rendah

hanya membutuhkan sedikit langkah dan kondisi serta tingkat kesulitan yang rendah pula, sebab bahkan anak kecil dapat melakukannya. Namun, contoh berbeda terlihat pada tindakan mengikat tali sepatu.

Secara komputasi, tugas ini sederhana karena tidak membutuhkan banyak langkah atau kondisi, tetapi tingkat kesulitannya tinggi sebab memerlukan koordinasi motorik halus yang kompleks. Tidak mengherankan jika Kasper Rorsted, CEO Adidas, pernah menyatakan pada tahun 2017 bahwa tantangan terbesar industri sepatu adalah menciptakan robot yang mampu memasang tali sepatu. Hingga kini, proses tersebut masih dilakukan secara manual karena belum ada teknologi yang dapat menirukan keterampilan motorik halus manusia dengan sempurna.

Sebaliknya, mencuci piring justru menempati posisi yang berbeda. Aktivitas ini membutuhkan banyak langkah, ruang, dan sumber daya semakin banyak piring, semakin kompleks pula prosesnya. Namun, tingkat kesulitannya relatif rendah, sehingga dapat dilakukan hampir semua orang, bahkan oleh mereka yang tidak terbiasa dengan pekerjaan manual. Lain halnya dengan menyetrika baju, yang dapat ditempatkan pada posisi kanan atas dalam peta. Aktivitas ini menuntut penggunaan sumber daya yang cukup besar sekaligus keterampilan motorik tertentu. Itulah sebabnya menyetrika dianggap sebagai kegiatan yang cukup rumit sekaligus sulit, membuat banyak orang memilih untuk menghindarinya jika memungkinkan.

Jika kita mengaitkan kembali pada contoh bermain sepak bola dan catur, terlihat perbedaan serupa. Sepak bola dapat dianggap sederhana dalam hal aturan, tetapi sulit karena menuntut keterampilan motorik, fisik, dan koordinasi tim. Sementara itu, catur memiliki aturan yang mudah dipelajari hanya dalam hitungan menit, tetapi kompleksitas komputasinya sangat tinggi, sebab jumlah kemungkinan langkah dan strategi tidak terhingga. Itulah sebabnya AI mampu menaklukkan manusia dalam permainan catur, namun keberadaan tim android yang memenangkan piala dunia sepak bola masih sebatas fiksi ilmiah. Menariknya, ada kasus abu-abu seperti robot tenis meja yang memperlihatkan bahwa batas antara kesulitan dan kompleksitas tidak selalu tegas, melainkan bisa berada di antara keduanya.

Berdasarkan pemetaan ini, kita dapat melihat adanya arah perkembangan AI yang lebih realistis. Panah perkembangan tersebut bergerak dari wilayah dengan kompleksitas rendah namun kesulitan tinggi menuju wilayah dengan kompleksitas tinggi namun kesulitan rendah. Hal ini dapat dipahami sebab artefak buatan manusia, betapapun pintarnya, masih belum handal dalam menyelesaikan tugas-tugas yang sangat bergantung pada keterampilan motorik halus. Akan tetapi, AI terbukti unggul dalam menghadapi masalah dengan kompleksitas tinggi, khususnya yang melibatkan pengolahan data dan perhitungan matematis.

Dengan demikian, masa depan AI yang sukses kemungkinan besar tidak hanya ditentukan oleh kemampuan memanfaatkan data sintetis atau hibrida, melainkan juga oleh kemampuannya untuk menerjemahkan tugas-tugas yang sulit menjadi kompleks. Bagaimana penerjemahan ini dilakukan? Caranya adalah dengan mentransformasi lingkungan tempat AI beroperasi agar menjadi lebih “ramah AI.” Transformasi tersebut mungkin akan menambah tingkat kompleksitas tugas yang harus diselesaikan oleh sistem, namun selama dapat

mengurangi kesulitan keterampilan yang diperlukan, hal ini justru membuka peluang bagi AI untuk lebih berhasil. Konsep inilah yang kemudian dikenal dengan istilah *envelopment*, yaitu strategi membungkus atau mengubah lingkungan sehingga AI dapat menjalankan fungsinya dengan lebih efektif.

Dalam dunia robotika industri, terdapat konsep selubung robot, yaitu ruang tiga dimensi yang menentukan batas di mana robot dapat bekerja secara efektif. Prinsip ini menjelaskan mengapa kita tidak membangun robot humanoid seperti C-3PO dalam film *Star Wars* untuk mencuci piring persis sebagaimana manusia melakukannya. Sebaliknya, kita membungkus lingkungan di sekitar robot sederhana agar sesuai dengan kapasitasnya, sehingga ia dapat menyelesaikan tugas dengan baik. Mesin pencuci piring, misalnya, bekerja bukan karena ia mampu meniru cara manusia mencuci piring, melainkan karena lingkungannya telah dirancang khusus: berupa kotak tahan air dengan sistem semprotan dan filter yang menyederhanakan pekerjaannya. Semakin canggih kemampuan robot, semakin sedikit pembungkus yang dibutuhkan; tetapi di sini selalu ada trade-off antara kecanggihan kapasitas mesin dengan kebutuhan untuk membungkus lingkungannya. Contoh serupa dapat ditemukan pada sistem rak robotik Amazon, di mana seluruh gudang didesain secara khusus agar ramah terhadap robot.

Logika ini juga berlaku pada robot memasak atau pembalik hamburger yang sudah dikembangkan. Keberhasilan mobil tanpa pengemudi pun bergantung pada kemampuan kita membungkus lingkungan transportasi agar sesuai dengan kebutuhan teknologi tersebut. Tidak heran jika kendaraan otonom lebih mudah dioperasikan di bandara, karena lingkungannya terkendali dan terstruktur, dibandingkan bus sekolah di pedesaan yang harus beroperasi dalam kondisi cuaca ekstrem, area tanpa sinyal, atau jalan tanpa satelit navigasi. Dengan kata lain, semakin “dibungkus” lingkungannya, semakin besar peluang teknologi otonom berhasil.

Contoh menarik datang dari Nike dengan peluncuran HyperAdapt 1.0 pada tahun 2016. Alih-alih menciptakan AI yang mampu mengikat tali sepatu seperti manusia, Nike justru merekayasa ulang konsep sepatu itu sendiri: dilengkapi sensor, baterai, motor, dan sistem kabel yang dapat menyesuaikan diri secara otomatis dengan kaki pemakainya. Dalam hal ini, bukan AI yang meniru manusia, melainkan desain sepatu yang membungkus kebutuhan agar sesuai dengan kemampuan mesin.

Jika sebelumnya pembungkusan hanya terjadi dalam bentuk perangkat khusus seperti mesin cuci atau mesin pencuci piring atau terbatas pada lingkungan industri, kini strategi ini merambah ke hampir seluruh aspek kehidupan. Kehadiran *Internet of Things* (IoT), jaringan 5G, dan realitas digital yang semakin “*onlife*” membuat proses pembungkusan lingkungan agar ramah AI berlangsung secara masif dan hampir tidak terlihat. Namun, perlu dicatat bahwa tidak semua masalah dapat dibungkus dengan mudah. Hambatan bukanlah soal pembuktian matematis, melainkan lebih pada kecerdikan desain, biaya ekonomi, serta preferensi pengguna.

Robot yang mampu menyetrika baju, misalnya, secara teknis bisa direkayasa. Pada tahun 2012, tim dari Universitas Carlos III Madrid memperkenalkan TEO, robot humanoid seberat 80 kg dengan tinggi 1,8 meter yang mampu menaiki tangga, membuka pintu, hingga

menyetrika kemeja. Tim peneliti membayangkan TEO akan terus dikembangkan untuk melakukan berbagai tugas rumah tangga, bahkan belajar dari manusia hanya dengan observasi. Namun, masa depan semacam ini patut diragukan. Mengapa? Karena pendekatan tersebut gagal membedakan antara tugas yang sulit yaitu yang membutuhkan keterampilan motorik tinggi dengan tugas yang kompleks yang dapat diselesaikan melalui perhitungan dan otomatisasi.

Sejarah teknologi menunjukkan bahwa AI dan robot lebih sukses ketika bekerja pada persoalan yang kompleks tetapi dapat dibungkus agar menjadi mudah, bukan ketika dipaksa meniru manusia dalam tugas-tugas sulit. Mobil otonom, misalnya, lahir bukan dari upaya menempatkan robot di kursi pengemudi, melainkan dari rekayasa ulang ekosistem kendaraan: kursi pengemudi dihilangkan sama sekali, dan seluruh sistem dikemas ulang agar ramah AI. Karena itu, masa depan AI tidak mungkin dipenuhi dengan android seperti TEO yang menirukan perilaku manusia. Justru lebih masuk akal jika kita membayangkan mesin-mesin khusus seperti Effie atau Foldimate, yang didesain hanya untuk satu fungsi: mengeringkan dan menyetrika pakaian. Mesin-mesin ini bekerja dengan logika “kotak” layaknya mesin cuci piring atau mesin cuci pakaian, hanya saja inputnya adalah baju kusut, sementara outputnya adalah baju rapi yang sudah disetrika.

Mungkin harga mesin semacam ini mahal, mungkin pula hasilnya belum sempurna, atau teknologinya akan berkembang ke arah yang saat ini sulit kita bayangkan. Namun, pola dasarnya jelas: jangan mencoba meniru manusia dengan AI, melainkan memanfaatkan kekuatan mesin dalam menghadapi kompleksitas. Kesulitan adalah musuh mesin, sementara kompleksitas adalah sahabatnya. Maka, masa depan AI terletak pada bagaimana kita terus membungkus dunia di sekelilingnya, merancang lingkungan yang mendukung, dan menemukan bentuk-bentuk penerapan baru yang memungkinkan teknologi berkembang progresif, efisien, dan meluas di berbagai sektor kehidupan.

18.5 KESIMPULAN

Masa depan AI tidak hanya ditentukan oleh inovasi teknis, melainkan juga oleh cara kita membentuk lingkungan, data, serta tata kelola digital yang mengaturnya. Ada dua poin penting: pertama, keberhasilan AI akan bergantung pada kemampuan kita menciptakan data yang lebih hibrida dan sintetis, bukan hanya bergantung pada data historis yang terbatas. Kedua, kita perlu merancang ulang realitas agar AI dapat berfungsi lebih baik melalui proses *meludifikasi* (membuat interaksi lebih seperti permainan yang terstruktur) dan *penyelubungan* (mendesain ulang lingkungan agar sesuai dengan kemampuan mesin).

Namun, teks juga menyoroti bahwa masa depan penuh dengan ketidakpastian (*unknown unknowns*), yakni hal-hal yang kita tahu ada tetapi tidak dapat kita ketahui bentuknya. Inilah yang membuat analisis tentang AI harus lebih diarahkan pada pengelolaan, bukan sekadar prediksi. Analisis ke depan menjadi bernilai ketika membantu kita mengantisipasi isu-isu etika, hukum, dan sosial (ELSI), seperti privasi prediktif, otonomi individu, hingga tanggung jawab penggunaan AI.

Dengan demikian, tantangan terbesar bukan hanya bagaimana menciptakan teknologi AI yang lebih pintar, tetapi bagaimana memastikan tata kelola digital berjalan dengan baik. Masa depan AI akan sangat ditentukan oleh sejauh mana kita mampu menyeimbangkan antara inovasi dengan regulasi, antara kemungkinan teknis dengan kepentingan etis, serta antara kompleksitas teknologi dengan kebutuhan manusia.

DAFTAR PUSTAKA

- Abdullah, Y. I., Schuman, J. S., Shabsigh, R., Caplan, A., & Al-Aswad, L. A. (2021). Ethics of artificial intelligence in medicine and ophthalmology. *The Asia-Pacific Journal of Ophthalmology*, 10(3), 289-298.
- Adams, C., Pente, P., Lernermeier, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Comput. Educ. Artif. Intell.*, 4, 100131.
- Ade-Ibijola, A., & Okonkwo, C. (2023). Artificial intelligence in Africa: Emerging challenges. In *Responsible AI in Africa: Challenges and opportunities* (pp. 101-117). Cham: Springer International Publishing.
- Agba, M. S., Agba, G. E. M., & Obeten, A. W. (2023). Artificial intelligence and public management and governance in developed and developing market economies. *Journal of Public Administration, Policy and Governance Research*, 1(2), 1-14.
- Arora, A., Alderman, J. E., Palmer, J., Ganapathi, S., Laws, E., Mccradden, M. D., ... & Liu, X. (2023). The value of standards for health datasets in artificial intelligence-based applications. *Nature medicine*, 29(11), 2929-2938.
- Bélisle-Pipon, J. C., Monteferrante, E., Roy, M. C., & Couture, V. (2023). Artificial intelligence ethics has a black box problem. *AI & society*, 38(4), 1507-1522.
- Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133-167.
- Boppiniti, S. T. (2023). Data ethics in ai: Addressing challenges in machine learning and data governance for responsible data science. *International Scientific Journal for Research*, 5(5), 1-29.
- Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, 64, 101475.
- Buhmann, A., & Fieseler, C. (2023). Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. *Business Ethics Quarterly*, 33(1), 146-179.
- Büthe, T., Djefal, C., Lütge, C., Maasen, S., & Ingersleben-Seip, N. V. (2022). Governing AI—attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence. *Journal of European Public Policy*, 29(11), 1721-1752.
- Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, equity, and inclusion in artificial intelligence: an evaluation of guidelines. *Applied Artificial Intelligence*, 37(1), 2176618.

- Camilleri, M. A. (2024). Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert systems*, 41(7), e13406.
- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and social sciences communications*, 10(1), 1-12.
- Chhillar, D., & Aguilera, R. V. (2022). An eye for artificial intelligence: Insights into the governance of artificial intelligence and vision for future research. *Business & Society*, 61(5), 1197-1241.
- Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10).
- Dara, R., Hazrati Fard, S. M., & Kaur, J. (2022). Recommendations for ethical and responsible use of artificial intelligence in digital agriculture. *Frontiers in Artificial Intelligence*, 5, 884192.
- Davitaia, A. (2025). Artificial Intelligence and machine learning in fraud detection for digital payments. *International Journal of Science and Research Archive*, 15(3), 714-719.
- De Almeida, P. G. R., Dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3), 505-525.
- De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, 3(6).
- Demaidi, M. N. (2025). Artificial intelligence national strategy in a developing country. *Ai & Society*, 40(2), 423-435.
- Elmahjub, E. (2023). Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. *Philosophy & Technology*, 36(4), 73.
- Filgueiras, F. (2024). Artificial intelligence and education governance. *Education, Citizenship and Social Justice*, 19(3), 349-361.
- Floridi, L. (2023). The ethics of artificial intelligence: Principles, challenges, and opportunities.
- Floridi, L. (Ed.). (2021). *Ethics, governance, and policies in artificial intelligence* (Vol. 144). Cham: Springer.
- Guenduez, A. A., & Mettler, T. (2023). Strategically constructed narratives on artificial intelligence: What stories are told in governmental artificial intelligence policies?. *Government Information Quarterly*, 40(1), 101719.
- Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization*.

- Hickman, E., & Petrin, M. (2021). Trustworthy AI and corporate governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review*, 22(4), 593-625.
- Hine, E., & Floridi, L. (2024). Artificial intelligence with American values and Chinese characteristics: a comparative analysis of American and Chinese governmental AI policies. *Ai & society*, 39(1), 257-278.
- Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4), 799-819.
- Huang, L. (2023). Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers*, 16(2), 2577-2587.
- Iphofen, R., & Kritikos, M. (2021). Regulating artificial intelligence and robotics: ethics by design in a digital society. *Contemporary Social Science*, 16(2), 170-184.
- Justo-Hanani, R. (2022). The politics of Artificial Intelligence regulation and governance reform in the European Union. *Policy Sciences*, 55(1), 137-159.
- Kamila, M. K., & Jasrotia, S. S. (2025). Ethical issues in the development of artificial intelligence: recognizing the risks. *International Journal of Ethics and Systems*, 41(1), 45-63.
- Khan, B., Fatima, H., Qureshi, A., Kumar, S., Hanan, A., Hussain, J., & Abdullah, S. (2023). Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials & Devices*, 1(2), 731-738.
- Kochupillai, M., Kahl, M., Schmitt, M., Taubenböck, H., & Zhu, X. X. (2022). Earth observation and artificial intelligence: Understanding emerging ethical issues and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(4), 90-124.
- McStay, A., & Rosner, G. (2021). Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society*, 8(1), 2053951721994877.
- Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2023). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, 41(1), 123-140.
- Mohammad Amini, M., Jesus, M., Fanaei Sheikholeslami, D., Alves, P., Hassanzadeh Benam, A., & Hariri, F. (2023). Artificial intelligence ethics and challenges in healthcare applications: a comprehensive review in the context of the European GDPR mandate. *Machine Learning and Knowledge Extraction*, 5(3), 1023-1035.
- Moon, M. J. (2023). Searching for inclusive artificial intelligence for social good: Participatory governance and policy recommendations for making AI more inclusive and benign for society. *Public Administration Review*, 83(6), 1496-1505.

- Morley, J., Murphy, L., Mishra, A., Joshi, I., & Karpathakis, K. (2022). Governing data and artificial intelligence for health care: developing an international understanding. *JMIR formative research*, 6(1), e31623.
- Nemorin, S., Vlachidis, A., Ayerakwa, H. M., & Andriotis, P. (2023). AI hyped? A horizon scan of discourse on artificial intelligence in education (AIED) and development. *Learning, Media and Technology*, 48(1), 38-51.
- Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B. P. T. (2023). Ethical principles for artificial intelligence in education. *Education and information technologies*, 28(4), 4221-4241.
- Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *Ai & Society*, 39(4), 1871-1882.
- O'Shaughnessy, M. R., Schiff, D. S., Varshney, L. R., Rozell, C. J., & Davenport, M. A. (2023). What governs attitudes toward artificial intelligence adoption and governance?. *Science and Public Policy*, 50(2), 161-176.
- Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., Adebiyi, O. O., Okunleye, O. J., & Ismaila Alao, A. (2024). Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics. *Asian Journal of Advanced Research and Reports*, 18(2), 1-23.
- Panda, S. (2025). *Scalable Artificial Intelligence Systems: Cloud-Native, Edge-AI, MLOps, and Governance for Real-World Deployment*. Deep Science Publishing.
- Porlezza, C. (2023). Promoting responsible AI: A European perspective on the governance of artificial intelligence in media and journalism. *Communications*, 48(3), 370-394.
- Richey Jr, R. G., Chowdhury, S., Davis-Sramek, B., Giannakis, M., & Dwivedi, Y. K. (2023). Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. *Journal of Business Logistics*, 44(4), 532-549.
- Roberts, H., Cows, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2023). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, 39(2), 79-97.
- Roberts, H., Cows, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. In *Ethics, governance, and policies in artificial intelligence* (pp. 47-79). Cham: Springer International Publishing.
- Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*.

- Schiff, D. (2022). Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education*, 32(3), 527-563.
- Shiohira, K. (2021). Understanding the Impact of Artificial Intelligence on Skills Development. *Education 2030*. UNESCO-UNEVOC International Centre for Technical and Vocational Education and Training.
- Slimi, Z., & Carballido, B. V. (2023). Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies. *TEM journal*, 12(2).
- Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). Ethics of artificial intelligence: Case studies and options for addressing ethical challenges (p. 116). Springer Nature.
- Stix, C. (2021). Actionable principles for artificial intelligence policy: three pathways. *Science and engineering ethics*, 27(1), 15.
- Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2022). Ethical principles for artificial intelligence in national defence. In *The 2021 Yearbook of the Digital Ethics Lab* (pp. 261-283). Cham: Springer International Publishing.
- Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and society*, 40(2), 137-157.
- Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIED for personalised learning pathways. *Electronic Journal of e-Learning*, 20(5), 639-653.
- Ulnicane, I. (2022). Artificial Intelligence in the European Union: Policy, ethics and regulation. In *The Routledge handbook of European integrations*. Taylor & Francis.
- Ulnicane, I., Eke, D. O., Knight, W., Ogoh, G., & Stahl, B. C. (2021). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, 46(1-2), 71-93.
- Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G. (2021). Framing governance for a contested emerging technology: insights from AI policy. *Policy and Society*, 40(2), 158-177.
- Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government information quarterly*, 39(4), 101685.
- World Health Organization. (2024). Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance. World Health Organization.
- Yanamala, A. K. Y., & Suryadevara, S. (2023). Advances in data protection and artificial intelligence: Trends and challenges. *International Journal of Advanced Engineering Technologies and Innovations*, 1(01), 294-319.

- Yanamala, A. K. Y., & Suryadevara, S. (2024). Navigating data protection challenges in the era of artificial intelligence: A comprehensive review. *Revista de Inteligencia Artificial en Medicina*, 15(1), 113-146.
- Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3), e12962.
- Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research*, 71, 591-666.
- Zhao, J., & Gómez Fariñas, B. (2023). Artificial intelligence and sustainable decisions. *European Business Organization Law Review*, 24(1), 1-39.

Peran AI dalam Etika, Tatakelola dan Kebijakan

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 dan ilmu sosiologi. 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).

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 978-634-7227-43-0 (PDF)



9

786347

227430